ΤΕΧΝΟΛΟΓΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΣΧΟΛΗ ΕΠΙΚΟΙΝΩΝΙΑΣ & ΜΕΣΩΝ ΕΝΗΜΕΡΩΣΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΙΧΝΕΥΣΗ ΙΣΤΟΣΕΛΙΔΩΝ ΠΟΡΝΟΓΡΑΦΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ ΜΕ ΒΑΣΗ ΤΟ ΚΕΙΜΕΝΟ ΚΑΙ ΤΗ ΔΟΜΗ

ΘΕΟΔΩΡΟΣ ΔΑΝΟΣ

ΤΕΧΝΟΛΟΓΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΣΧΟΛΗ ΕΠΙΚΟΙΝΩΝΙΑΣ & ΜΕΣΩΝ ΕΝΗΜΕΡΩΣΗΣ

ΤΜΗΜΑ **Ε**ΠΙΚΟΙΝΩΝΙΑΣ & **Σ**ΠΟΥΔΩΝ **Δ**ΙΑΔΙΚΤΥΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΙΧΝΕΥΣΗ ΙΣΤΟΣΕΛΙΔΩΝ ΠΟΡΝΟΓΡΑΦΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ ΜΕ ΒΑΣΗ ΤΟ ΚΕΙΜΕΝΟ ΚΑΙ ΤΗ ΔΟΜΗ

ΘΕΟΔΩΡΟΣ ΔΑΝΟΣ

Επιβλέπων Καθηγητής
Δρ. Νικόλας Τσαπατσούλης

Λεμεσός 2015

Πνευματικά δικαιώματα

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Δρ. Νικόλα Τσαπατσούλη για την καθοδήγηση αλλά και τις γνώσεις που μου παρείχε όσον αφορά τις στατιστικές αναλύσεις και τους αλγόριθμους που χρησιμοποιήθηκαν αλλά και συνολικά για την πολύτιμη βοήθεια που μου πρόσφερε σε όλη την διάρκεια εκπόνησης της παρούσας μελέτης. Θα ήθελα ακόμη, να ευχαριστήσω ιδιαίτερα τις καθηγήτριες μου Δρ. Βασιλική Τρίγκα και Δρ. Δήμητρα Μηλιώνη για όλη την βοήθεια που μου πρόσφεραν στην ερευνητική προετοιμασία μου. Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια μου για την βοήθεια σε όλη τη διάρκεια των σπουδών μου. Τέλος, δεν μπορώ να μην αναφέρω τις ευχαριστίες μου στην κοπέλα μου, για την σπουδαία υποστήριξη που μου πρόσφερε στις μεταμεσονύχτιες και ατέλειωτες ώρες μελέτης της παρούσας έρευνας.

Πίνακας Περιεχομένων

Συντομογραφίες

URL:	Uniform Resource Locator

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Απόδοση όρων

Association analysis	Ανάλυση συσχέτισης
Basket analysis	Ανάλυση Καλαθιού
Binary Classification	Δυαδική Κατηγοριοποίηση
Data mining	Εξόρυξη δεδομένων
Dataset	Σύνολο δεδομένων
Decision tree	Δέντρο απόφασης
Document summarization	Περίληψη εγγράφου
Error rate	Ρυθμός σφάλματος
Evaluation Dataset	Σύνολο δεδομένων αξιολόγησης
Flat Classification	Επίπεδη Κατηγοριοποίηση
Functional Classification	Λειτουργική Κατηγοριοποίηση
Hard Classification	Αυστηρή Κατηγοριοποίηση
Heading	Κεφαλίδα
Hierarchical Classification	Ιεραρχική Κατηγοριοποίηση
Informational Page	Ιστοσελίδα πληροφοριών
Instance	Παράδειγμα
Layout	Διάταξη
Link	Σύνδεσμος
Matrix	Πίνακας
Metadata	Μεταδεδομένα
Multiclass Classification	Κατάταξη σε πολλές ετικέτες
Personal Page	Προσωπική σελίδα
Precision	Ακρίβεια
Proxy	Πληρεξούσιος
Recall	Ανάκληση
Research Page	Ιστοσελίδα διερεύνησης
Sentiment Classification	Γνωστική Κατηγοριοποίηση
Single-label Classification	Κατηγοριοποίησης μονής ετικέτας

Soft Classification	Χαλαρή Κατηγοριοποίηση
Standard Dataset	Πρότυπο σύνολο δεδομένων
Stemming	Αποκοπή καταλήξεων
Subject Classification	Θεματική Κατηγοριοποίηση
Tag	Ετικέτα
Title	Τίτλος
Uniform Resource Locator	Ενιαίος Εντοπιστής Πόρου
Web Page Classification	Κατηγοριοποίηση ιστοσελίδας

Περίληψη

Στο διαδίκτυο μέρα με την μέρα δημιουργείται μεγάλος όγκος πληροφορίας. Πληροφορία που όμως είναι αδόμητη. Οι μηχανές αναζήτησης επιτρέπουν ανάκτηση πληροφοριών από αδόμητα δεδομένα βασιζόμενες κατά κύριο λόγο στο κείμενο που υπάρχει στις οικείες ιστοσελίδες. Η ανάγκη για αλγόριθμους βελτιστοποίησης των αποτελεσμάτων (αποδοτικότητα) και απόδοσης των μηχανών αναζήτησης (χρονική αποτελεσματικότητα) είναι μεγάλη. Η ταξινόμηση ιστοσελίδων σε κατηγορίες θεωρείται ότι επιταχύνει την αναζήτηση και επιτρέπει ερωτήματα προσαρμοσμένα στο προφίλ του εκάστοτε χρήστη αυξάνοντας την αποδοτικότητα.

Στην παρούσα εργασία εστιάζουμε στην ταξινόμηση ιστοσελίδων σε πορνογραφικές και μη πορνογραφικές δημιουργώντας έτσι έναν αλγόριθμό εντοπισμού πορνογραφικών ιστοσελίδων. Σε αντίθεση με τις υφιστάμενες τεχνικές που βασίζονται στην ανάλυση των εικόνων των ιστοσελίδων για τον εντοπισμό γυμνού η δική μας τεχνική βασίζεται αποκλειστικά σε χαρακτηριστικά κειμένου και στη δομή της ιστοσελίδας. Τα χαρακτηριστικά κειμένου εξάγονται με τη βοήθεια τεχνικών από την περιοχή της ανάκτησης πληροφορίας και συγκεκριμένα με τη μέθοδο tf-df που αποτελεί μια παραλλαγή της πολύ γνωστής μεθόδου tf-idf. Τα χαρακτηριστικά δομής επιλέχθηκαν με ευρυστικό τρόπο και περιλαμβάνουν τον αριθμό των εικόνων της ιστοσελίδας (κανονικοποιημένο και απόλυτο) και τον αριθμό των υπερσυνδέσμων της ιστοσελίδας (κανονικοποιημένο και απόλυτο). Ο ταξινομητής μας εκπαιδεύτηκε με την τεχνική εκμάθησης μέσω παραδειγμάτων με χρήση των χαρακτηριστικών που αναφέρθηκαν νωρίτερα (κειμένου και δομής) με τη βοήθεια διάφορων αλγορίθμων από την περιοχή της μηχανικής μάθησης. Καταλήξαμε ότι την βέλτιστη επίδοση έχει ο ταξινομητής Bayesian Net τον οποίο και υιοθετήσαμε για τα τελικά μας πειράματα.

Χρησιμοποιώντας ελέγχους σημαντικότητας, και συγκεκριμένα το t-test, δείξαμε ότι η επίδοση του αυτόματου εντοπισμού πορνογραφικών ιστοσελίδων είναι συγκρίσιμη με την χειρωνακτική ταξινόμηση (δηλαδή την ταξινόμηση από ανθρώπους).

Πιστεύουμε ότι πέρα από τα επιστημονικά αποτελέσματα η παρούσα ερεύνα είναι σημαντική και σε πρακτικό επίπεδο (για μηχανές αναζήτησης αλλά και για οργανισμούς ή ιδιώτες όπου επιθυμούν την απαγόρευση των πορνογραφικών ιστοσελίδων). Ακόμη, είναι σημαντική για τον αυτόματο εντοπισμό πορνογραφικών ιστοσελίδων στο διαδίκτυο ή σε ενδοδίκτυα.

Εισαγωγή

Για να μπορέσουμε να επεκτείνουμε τις γνώσεις μας πρέπει να βασιστούμε σε προηγούμενες γνώσεις (“If I have seen further it is by standing on the shoulders of giants”, Isaac Newton 1676). Στην κοινωνία της πληροφορίας οι υπολογιστές παίζουν κύριο ρόλο στην αναπαράσταση της γνώσης και στην αναζήτηση της γνώσης αυτής. Η γνώση κτίζεται με σύνθεση πληροφοριών, πληροφοριών που όμως βρίσκονται, κρυμμένες σε διακομιστές σε ολόκληρο το δίκτυο περιμένοντας κάποια μηχανή αναζήτησης να τις φανερώσει στο ορατό μέρος του διαδικτύου (Bergman, 2001).

Για μεγαλύτερη αποτελεσματικότητα, η μηχανή αναζήτησης πρέπει να κατανοήσει τα δεδομένα τα οποία εντοπίζει έτσι ώστε να τα εμφανίζει με την σωστή σειρά στους χρήστες της χρήστες της. Η διαδικασία αυτή δεν είναι καθόλου εύκολη γιατί υπάρχουν πολλοί τύποι δεδομένων και οι ιστοσελίδες δεν χρησιμοποιούν συγκεκριμένες δομές. Επιπλέον, στα δεδομένα που εμφανίζονται στις ιστοσελίδες δεν υπάρχει δομή, όπως υπάρχει για παράδειγμα σε μια βάση δεδομένων. Ακόμη, η μηχανή αναζήτησης θα ήταν χρήσιμο να αναγνωρίσει τον τύπο της ιστοσελίδας έτσι ώστε να επεξεργαστεί καλύτερα τα δεδομένα και να παρουσιάσει και να εκτιμήσει καλύτερα τη συνάφεια της με το ερώτημα αναζήτησης που έχει υποβάλει ο χρήστης. Στο διαδίκτυο υπάρχουν πολλοί τύποι ιστοσελίδων: “Ειδησεογραφικές ιστοσελίδες”, “Ηλεκτρονικοί κατάλογοι” και “Άλμπουμ φωτογραφίας” κ.ο.κ. Το γεγονός ότι το διαδίκτυο είναι μια μορφή τεράστιας ποσότητας αδόμητων δεδομένων είναι κάτι που απασχολεί την ακαδημαϊκή κοινότητα (Croft *et al*., 2010). Η παρούσα έρευνα προσπαθεί να διερευνήσει ένα μικρό μέρος αυτού του προβλήματος μέσα από την αυτόματη κατηγοριοποίηση ιστοσελίδων.

Διατύπωση ερευνητικού προβλήματος

Η παρούσα μελέτη επικεντρώνεται στην ανάπτυξη ενός υπολογιστικού συστήματος που θα κατηγοριοποιεί πορνογραφικές ιστοσελίδες αυτόματα αφού πρώτα εκπαιδευτεί μέσω ανθρώπινης επίβλεψης.

Ένα τέτοιο υπολογιστικό σύστημα εμπλέκει διάφορους αλγόριθμους και τεχνικές τόσο από το πεδίο της Τεχνητής Νοημοσύνης όσο και από τα πεδία της Ανάκτησης Πληροφορίας και της Εξόρυξης Δεδομένων. Επιπλέον χρειάζεται να αντιμετωπιστεί ένα ακόμη ζήτημα: Ποια είναι εκείνα τα χαρακτηριστικά που μπορούν να εξαχθούν από τις πορνογραφικές ιστοσελίδες και μας επιτρέπουν την αυτόματη ταξινόμηση τους;

Στο διαδίκτυο υπάρχουν πάρα πολλές ιστοσελίδες οι οποίες παρέχουν πληροφορίες. Όμως η ανάγκη πληροφόρησης ενός χρήστη δημιουργεί την επιπρόσθετη ανάγκη για πιο αποτελεσματικούς αλγόριθμους αναζήτησης ως προς τις πληροφορίες αυτές. Όμως, αυξάνοντας την αποτελεσματικότητα ενός αλγορίθμου αυξάνεται συνήθως και ο χρόνος εκτέλεσης του λόγω αυξημένης πολυπλοκότητας. Η ανάγκη ανάπτυξης ενδιάμεσων αλγόριθμων για βελτιστοποίηση της αναζήτησης είναι αναγκαία. Στην παρούσα έρευνα θα αναπτυχθεί και υλοποιηθεί ένας αλγόριθμος που θα παίρνει ως είσοδο μια σειρά από ιστοσελίδες αγνώστου τύπου και θα έχει ως έξοδο τον τύπο τους (εάν μπορούν να κατηγοριοποιηθούν) αφού εκπαιδευτεί με ένα γνωστό συγκεκριμένο σύνολο κατηγοριοποιημένων ιστοσελίδων.

Σημασία Μελέτης

Η παρούσα έρευνα υπάγεται σε ένα ευρύ επιστημονικό πεδίο και έχει στοιχεία που αφορούν πολλές προηγούμενες έρευνες. Εντούτοις, οι παράμετροι και οι τεχνικές που έχουν χρησιμοποιηθεί δεν εφαρμόστηκαν σε καμία άλλη έρευνα έως τώρα. Όταν η έρευνα ολοκληρωθεί θα μπορεί να

αξιοποιηθεί από συστήματα μηχανών αναζήτησης για διάφορες χρήσεις όπως είναι η βελτιστοποίηση της κατάταξης των αποτελεσμάτων και η δημιουργία εστιασμένων μηχανών αναζήτησης – π.χ. μηχανή αναζήτησης μόνο για αναζήτηση σε θέματα πληροφορικής – ή για σκοπούς στοχευμένης διαφήμισης, αλλά και ο εντοπισμός πορνογραφικών ιστοσελίδων. Σε προσωπικό επίπεδο επέλεξα την συγκεκριμένη μελέτη γιατί είχα προηγούμενες εμπειρίες με τα πεδία της εξόρυξης δεδομένων και της τεχνητής νοημοσύνης. Η έρευνα δεν απαιτεί οικονομικούς πόρους που να την περιορίζουν, και μπορεί να εφαρμοστεί στα κατάλληλα χρονικά πλαίσια που έχουν καθοριστεί. Επίσης, δεν υπάρχουν πρακτικές αδυναμίες ως προς την υλοποίηση των αλγορίθμων έτσι οι συνθήκες ευνοούν την επίτευξη του στόχου.

Αναγκαιότητα μελέτης

Όπως έχει προαναφερθεί η αναγκαιότητα της μελέτης πηγάζει από την ανάγκη γρήγορης και αποτελεσματικής πληροφόρησης του χρήστη ως αποτέλεσμα των ερωτημάτων που υποβάλλει. Το θέμα της αυτόματης κατηγοριοποίησης ιστοσελίδων έχει απασχολήσει τόσο την ακαδημία όσο και την βιομηχανία της ανάκτησης πληροφοριών. Υπάρχουν αρκετές μελέτες που έχουν εστιάσει στο συγκεκριμένο πρόβλημα (βλέπε ενδεικτικά Qi & Davison, 2009, Asirvatham & Ravi, 2001). Η βασική διαφορά με τις μελέτες αυτές είναι ο αλγόριθμος εξαγωγής χαρακτηριστικών από τις ιστοσελίδες που θα χρησιμοποιήσουμε στην παρούσα μελέτη. Η εξαγωγή των κατάλληλων χαρακτηριστικών είναι το “κλειδί” ενός συστήματος κατηγοριοποίησης ιστοσελίδων αφού από αυτό εξαρτάται το πώς θα κατηγοριοποιούνται οι ιστοσελίδες και πόσο επιτυχώς αυτό θα γίνεται. Επιτυγχάνοντας έναν αποδοτικό και αποτελεσματικό αλγόριθμο κατηγοριοποίησης ιστοσελίδων, τόσο πιο αποδοτικές θα είναι οι μηχανές αναζήτησης αν τον υλοποιήσουν και τον προσθέσουν στους μηχανισμούς ανάκτησης ιστοσελίδων που διαθέτουν (Tsukada et al., 2001).

Είναι πλέον καθημερινή δραστηριότητα του χρήστη στο διαδίκτυο, να

αναζητά στον ιστό πληροφορίες. Η αναζήτηση και η επικοινωνία μέσω διαδικτύου είναι η πιο συχνή χρήση του υπολογιστή (Croft et. al., 2010). Αυτόματη κατηγοριοποίηση ιστοσελίδων είναι χρήσιμη σε πολλούς τομείς όπως π.χ. στο μάρκετινγκ και στον εντοπισμό πορνογραφικών ιστοσελίδων.

Ένα άλλο πεδίο έρευνας στο οποίο η παρούσα μελέτη είναι χρήσιμη είναι η στοχευμένη αραχνοποίηση (Qi & Davison, 2009). Η αραχνοποίηση είναι η διαδικασία όπου η μηχανή αναζήτησης εντοπίζει τις ιστοσελίδες και τις αποθηκεύει για να την αναλύσει μετέπειτα. Πολλές φορές η πλήρης αραχνοποίηση δεν είναι αποδοτική ή αποτελεσματική. Η στοχευμένη αραχνοποίηση αποθηκεύει μόνο τα έγγραφα που είναι σχετικά με την θεματική όπως είναι για παράδειγμα: “θεματολογία πορνογραφίας”, “θεματολογία υπολογιστών” κ.ο.κ. Επίσης υπάρχει ελλιπής διερεύνηση ως προς το πεδίο των χαρακτηριστικών που εξάγονται από μία ιστοσελίδα. Δηλαδή ποια χαρακτηριστικά πρέπει να χρησιμοποιηθούν από τους ταξινομητές για την κατηγοριοποίηση ιστοσελίδων(Qi & Davison, 2009).

Δεν μπορούμε να παραλείψουμε το γεγονός ότι η έρευνα μπορεί να βοηθήσει στον έλεγχο των ιστοσελίδων που επισκέπτονται τα ανήλικα άτομα. Είναι πραγματικότητα ότι σήμερα τα ανήλικα άτομα έχουν πρόσβαση στο διαδίκτυο το οποίο είναι ανοιχτό προς την πορνογραφία και ανεξέλεγκτο. Υπάρχουν αρκετά λογισμικά τα οποία με την εγκατάσταση τους προστατεύουν τα παιδιά και αυτός ο αλγόριθμος θα βοηθούσε στην υλοποίηση λογισμικών τα οποία θα μπορούν να λειτουργούν χωρίς βάση δεδομένων για έλεγχο της περιήγησης στο διαδίκτυο.

Ένας άλλος τρόπος ελέγχου των ιστοσελίδων που επισκέπτονται οι χρήστες είναι να υλοποιηθεί ένας proxy ο οποίος θα περνά όλο το ρεύμα μεταξύ του εσωτερικού δικτύου και του διαδικτύου και να υλοποιηθεί προσθέτοντας τον αλγόριθμο της παρούσας έρευνας για έλεγχο πορνογραφικών ιστοσελίδων. Φυσικά αυτή η διαδικασία είναι χρονοβόρα και θα έχει αρνητική επίπτωση

στον χρόνο περιήγησης των χρηστών. Ένας proxy με τους αλγόριθμους αυτής της έρευνας μπορεί να εφαρμοσθεί σε πανεπιστήμια, σπίτια, ακόμη και επιχειρήσεις.

Τέλος η βελτίωση των μηχανών αναζήτησης ως προς την αποδοτικότητα, δηλαδή τον χρόνο ανταπόκρισης των αποτελεσμάτων είναι σημαντική. Αυτό όμως εξαρτάται από τους αλγόριθμους της μηχανής αναζήτησης και πώς αυτοί ανταποκρίνονται στα ερωτήματα των χρηστών. Μία έρευνα για αυτόματη κατηγοριοποίηση ιστοσελίδων μπορεί να μειώσει τον χρόνο αυτό από την αλγοριθμική σκοπιά (Tsukada et al., 2001).

Θεωρητική Τεκμηρίωση

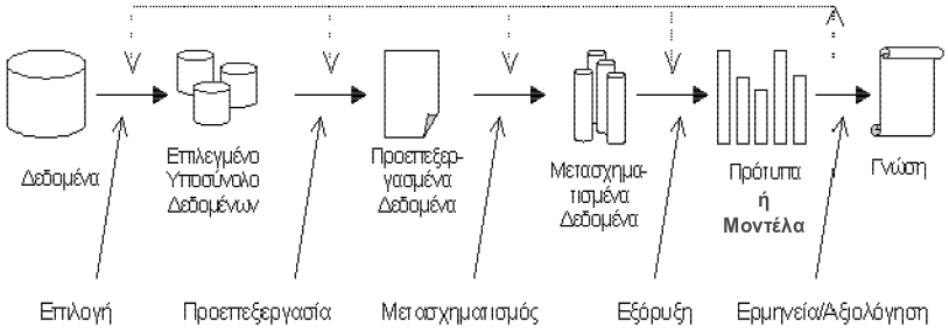
Η παρούσα μελέτη βασίζεται σε θεωρίες και μοντέλα από τις περιοχές της Εξόρυξης Δεδομένων, της Ανάκτησης Πληροφορίας και της Μηχανικής Μάθησης. Για πληρέστερη κατανόηση παρουσιάζουμε συνοπτικά τις τρεις αυτές περιοχές.

Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων (ή ανακάλυψη γνώσης) είναι η ανάλυση δεδομένων από διάφορες οπτικές και μετατροπή τους σε χρήσιμες πληροφορίες. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για διάφορους στόχους όπως για παράδειγμα τη μείωση του κόστους λειτουργίας ή την αύξηση εσόδων μιας επιχείρησης. Η εξόρυξη δεδομένων εστιάζει στην αναζήτηση νέων προτύπων και χρήσιμων αλλά και κατανοητών σχέσεων σε δεδομένα. Ως ερευνητικό πεδίο συνδέεται με την Στατιστική, τις Βάσεις Δεδομένων και την Μηχανική Μάθηση (Witten & Frank, 2005). Τεχνικά, μπορούμε να πούμε πώς η εξόρυξη δεδομένων είναι η διαδικασία εύρεσης σχέσεων ή προτύπων ανάμεσα σε πολλά δεδομένα σε μία μη δομημένη βάση δεδομένων. Αφορά την επεξεργασία των δεδομένων που εξήχθησαν από αλγόριθμους ανακάλυψης γνώσης καθώς και την ερμηνεία των αποτελεσμάτων τους. Επιτρέπει στους χρήστες να αναλύσουν τα δεδομένα από διάφορες οπτικές γωνίες, να τα κατηγοριοποιήσουν και να συνοψίσουν τις σχέσεις τις οποίες βρήκαν οι αλγόριθμοι.

Ο όρος εξόρυξη δεδομένων είναι νέος όρος, ενώ ως η διαδικασία προϋπήρχε μέσω της τεχνολογίας (Witten & Frank, 2005). Πολλές εταιρίες χρησιμοποιούσαν υπερ-υπολογιστές για την ανάλυση των δεδομένων από τα scanner των supermarkets για την ανάλυση των προτιμήσεων των καταναλωτών. Για παράδειγμα, το σύστημα θα αναγνωρίσει και θα παρουσιάσει ένα ανάλογο αποτέλεσμα: ένας πελάτης αγοράζει ψωμί, τότε κατά μια πιθανότητα θα αγοράσει και γάλα, επειδή αναγνωρίστηκε ένα

πρότυπο στην σχέση του συνδυασμού ψωμί-γάλα. Η πληροφορία μπορεί να μετατραπεί σε γνώση από τα πρότυπα των δεδομένων. Για παράδειγμα η ανάλυση των πωλήσεων ενός supermarket μπορεί να αναδείξει την αγοραστική συμπεριφορά του καταναλωτή και έτσι να βοηθήσει τους κατασκευαστές ή τους μεταπωλητές να αποφασίσουν ποια προϊόντα πρέπει να διαφημίσουν περισσότερο ή ποια πρέπει να συνδυάσουν μαζί.



Σχήμα 1: Διαδικασία ανακάλυψης γνώσης

Οι εφαρμογές της εξόρυξης δεδομένων ποικίλλουν: οικονομικό μάρκετινγκ, πολιτική επικοινωνία, ιατρική κλπ. Η παρούσα έρευνα χρησιμοποιεί την εξόρυξη δεδομένων για την εύρεση προτύπων στο περιεχόμενο διαφόρων τύπων ιστοσελίδων έτσι ώστε οι αλγόριθμοι που θα χρησιμοποιηθούν να κατηγοριοποιήσουν τις ιστοσελίδες στη σωστή κατηγορία. Το σύστημα θα πρέπει να μαθαίνει αυτόματα και να βελτιώνεται μέσα από την εμπειρία του με τα δεδομένα. Μπορεί είτε να εκπαιδεύεται από πριν με δεδομένα που είναι επισημειωμένα με την σωστή ετικέτα / κατηγορία είτε να ομαδοποιεί αυτόματα τα δεδομένα βρίσκοντας κοινά πρότυπα στις διάφορες ομάδες (βλέπε Σχήμα 1).

Ανάκτηση Πληροφορίας

Η ανάκτηση πληροφορίας είναι η διαδικασία εύρεσης πληροφοριών σε αδόμητα δεδομένα. Αποτελεί ένα ευρύ επιστημονικό πεδίο, με το οποίο ασχολείται εκτενώς η ακαδημαϊκή κοινότητα των επιστημόνων πληροφορικής και μαθηματικών. Επηρεάζεται από πολλά άλλα επιστημονικά πεδία όπως η Επεξεργασία Φυσικής Γλώσσας, οι Σχεσιακές Βάσεις Δεδομένων και άλλα πεδία παρόμοιου ενδιαφέροντος.

Μια ευρέως γνωστή εφαρμογή της ανάκτησης πληροφορίας στην οποία εφαρμόζονται τεχνικές και μηχανισμοί του πεδίου, είναι οι μηχανές αναζήτησης. Οι διαδικτυακές μηχανές αναζήτησης χρησιμοποιούν πολλούς μηχανισμούς τόσο από την ανάκτηση πληροφορίας όσο και από την εξόρυξη δεδομένων έτσι ώστε μέσα από μία πολύπλοκη διαδικασία[1] να επιστρέψει χρήσιμα αποτελέσματα στον χρήστη. Ένα λειτουργικό στοιχείο μιας μηχανής αναζήτησης είναι ο ανιχνευτής. Ο ανιχνευτής (ή αράχνη) είναι ένας μηχανισμός που διαβάζει ιστοσελίδες από τον παγκόσμιο ιστό κινούμενος από μια ιστοσελίδα σε άλλη, μέσω των υπερσυνδέσμων που τις συνδέουν. Μία άλλη κατηγορία των ανιχνευτών είναι οι εστιασμένοι ανιχνευτές όπου επικεντρώνονται σε ειδικές θεματικές περιοχές και έτσι η μηχανή αναζήτησης περιορίζει το περιεχόμενο της έτσι ώστε να αυξήσει την ποιότητα αναζήτησης (Croft et. al., 2010). Τα αποτελέσματα της παρούσας έρευνας μπορούν να συνδυαστούν αργότερα με ένα μηχανισμό εστιασμένης ανίχνευσης ιστοσελίδων έτσι ώστε να επιλέγονται αυτόματα ιστοσελίδες που εμπίπτουν μόνο σε ειδικούς τύπους ιστοσελίδων, συγκεκριμένα πορνογραφικών ιστοσελίδων (Tsukada et. al., 2001).

1 Η διαδικασία αυτή παίρνει επιπρόσθετα στοιχεία από το επιστημονικό πεδίο επεξεργασία φυσικής γλώσσας

Το πεδίο ανάκτησης πληροφορίας επίσης ασχολείται με αλγόριθμους και τεχνικές που αφορούν την ανάκτηση και κατάταξη ιστοσελίδων που επιστρέφονται από μια μηχανή αναζήτησης. Οι αλγόριθμοι και οι τεχνικές αυτές είναι γενικότερα γνωστές ως Μοντέλα Ανάκτησης (retrieval models) και Μοντέλα Κατάταξης (ranking models).

Μοντέλα Ανάκτησης

Τα μοντέλα ανάκτησης χρησιμοποιούνται για την ανάκτηση εγγράφων σχετικών με το ερώτημα που υποβλήθηκε από τον χρήστη. Δηλαδή σε μία μηχανή αναζήτησης, το μοντέλο αυτό θα επιστρέψει από όλη την συλλογή εγγράφων, μόνο τα έγγραφα που σχετίζονται με το ερώτημα του χρήστη. (Croft et. al., 2010). Για την εύρεση της συνάφειας ενός εγγράφου με ένα ερώτημα χρειάζεται να οριστούν κατάλληλες αναπαραστάσεις τόσο του ερωτήματος όσο και του εγγράφου αλλά και μετρικές μέσω των οποίων υπολογίζεται η συνάφεια με βάση τις αναπαραστάσεις αυτές.

Όμως, τα Μοντέλα Ανάκτησης δεν μπορούν να ξεχωρίσουν ανάμεσα στην πληθώρα αποτελεσμάτων του αλγόριθμου ποια έγγραφα είναι περισσότερο συναφή με το ερώτημα από κάποια άλλα, έτσι ώστε τα πιο συναφή να εμφανίζονται πρώτα στην κατάταξη των αποτελεσμάτων που παρουσιάζονται στον χρήστη. Την εργασία αυτή την αναλαμβάνουν τα Μοντέλα Κατάταξης.

Μοντέλα Κατάταξης

Εφόσον βρεθούν τα συναφή με το ερώτημα έγγραφα χρειάζεται να καταταχθούν με φθίνουσα σειρά του βαθμού συνάφειας έτσι ώστε να αυξηθεί η αποτελεσματικότητα αναζήτησης μέσω της ικανοποίησης του χρήστη. Είναι γνωστό ότι σε μια αναζήτηση οι χρήστες πολύ σπάνια ελέγχουν τα έγγραφα που εμφανίζονται μετά την $5^{η}$ θέση στη λίστα των αποτελεσμάτων. Η διαδικασία κατάταξης των εγγράφων βασίζεται σε μαθηματικούς κανόνες (Croft et al., 2010).

Τόσο τα Μοντέλα Ανάκτησης όσο και τα Μοντέλα Κατάταξης μπορεί να βελτιώσουν την αποτελεσματικότητα τους με την βοήθεια της κατηγοριοποίησης ιστοσελίδων διότι μέσω αυτής τους δίνεται μια ξεκάθαρη εικόνα ως προς το τι αντιπροσωπεύει η ιστοσελίδα (Qi & Davison, 2009).

Μηχανική Μάθηση

Η εξόρυξη δεδομένων χρησιμοποιεί εκτεταμένα τεχνικές από το πεδίο της μηχανικής μάθησης. Η μηχανική μάθηση μπορεί να απαντήσει στο εξής ερώτημα: Πώς μπορούμε να προγραμματίσουμε ένα σύστημα το οποίο να μαθαίνει αυτόματα και να βελτιώνεται με την εμπειρία; Οι τεχνικές μηχανικής μάθησης διακρίνονται σε τεχνικές με επίβλεψη[2] και χωρίς επίβλεψη (Βλαχάβας, Ι. κ.ά., 2006).

Μηχανική Μάθηση με επίβλεψη – Ταξινόμηση δεδομένων

Η μηχανική μάθηση με επίβλεψη είναι μια διαδικασία όπου το σύστημα προγραμματίζεται έτσι ώστε να εκπαιδεύεται πώς να ταξινομεί τα δεδομένα σωστά. Στην συνέχεια αφού εκπαιδευτεί δημιουργείται ένα μοντέλο. Ακολούθως το μοντέλο αυτό χρησιμοποιείται για να ταξινομήσει δεδομένα των οποίων τα χαρακτηριστικά γνωρίζουμε αλλά όχι την κατηγορία (ή τάξη ή κλάση) τους (Βλαχάβας, Ι. κ.ά., 2006). Για παράδειγμα ας υποθέσουμε ότι ως δεδομένα εκπαίδευσης δοθούν τα ακόλουθα:

Α/Α Στοιχείου	Ηλικία	Κλάση
1	10	Νεαρός
2	13	Νεαρός
3	16	Νεαρός
4	40	Ενήλικας
5	53	Ενήλικας
6	56	Ενήλικας

Πίνακας 1: Υποθετικά δεδομένα εκπαίδευσης

2 Μάθηση με επίβλεψη ή αλλιώς ταξινόμηση

Στο πιο πάνω παράδειγμα το χαρακτηριστικό είναι η Ηλικία και υπάρχουν δύο κατηγορίες (κλάσεις): Νεαρός και Ενήλικας. Εφόσον το σύστημα εκπαιδευτεί με ένα από τους αλγόριθμους μάθησης με επίβλεψη (οι οποίοι ποικίλουν), θα μπορούσαμε γενικά να πούμε πως το μοντέλο που έχει δημιουργηθεί θα έχει τους κανόνες:

Αν Ηλικία >=10 και Ηλικία <=16 Τότε η Κλάση είναι Νεαρός

Αν Ηλικία >=40 και Ηλικία <=56 Τότε η Κλάση είναι Ενήλικας

Φυσικά η πραγματική διαδικασία είναι αρκετά πιο πολύπλοκη: Τα χαρακτηριστικά που χρησιμοποιούμε στην ταξινόμηση είναι πολύ περισσότερα από ένα και για τη διαδικασία δημιουργίας του μοντέλου εφαρμόζονται μαθηματικές συναρτήσεις (Βλαχάβας, κ.ά., 2006). Εφόσον έχουμε το μοντέλο, εάν του δώσουμε την Ηλικία->11 χωρίς να του δώσουμε την κλάση, βάσει της εκπαίδευσης του θα μας δώσει το αποτέλεσμα “Νεαρός”.

Επομένως, η κλάση είναι η πληροφορία που θα ζητήσουμε από το σύστημα να μας βρει όταν εκπαιδευτεί. Στην πραγματικότητα τα δεδομένα εκπαίδευσης δεν είναι τόσο ξεκάθαρα ως προς την κατηγορία στην οποία ανήκουν. Στο πιο πάνω παράδειγμα θα μπορούσε να έχουμε δύο στοιχεία με: Ηλικία => 19, Κλάση => Νεαρός και Ηλικία => 18, Κλάση => Ενήλικας. Σε αυτή την περίπτωση το μοντέλο θα οδηγεί σε μερικές περιπτώσεις (ηλικίες 18-20) σε λάθος κατηγοριοποίηση. Το σφάλμα αυτό ονομάζεται σφάλμα ταξινόμησης και είναι ένα από τα βασικά κριτήρια ελέγχου της επίδοσης των αλγορίθμων μηχανικής μάθησης.

Στην παρούσα εργασία θα προσπαθήσουμε μέσα από μια διαδικασία που εξηγείται στο κεφάλαιο “Μεθοδολογία” να επιλέξουμε τον αλγόριθμο με το μικρότερο σφάλμα ταξινόμησης.

Μηχανική Μάθηση χωρίς επίβλεψη – Ομαδοποίηση δεδομένων

Η μηχανική μάθηση χωρίς επίβλεψη[3] δεν χρησιμοποιεί δεδομένα εκπαίδευσης, δηλαδή δεν χρειάζεται να εκπαιδεύσουμε το σύστημα. Οι αλγόριθμοι που χρησιμοποιούνται μπορούν να ξεχωρίζουν αυτόματα τις κλάσεις από τα χαρακτηριστικά τους αφού βασίζονται σε μοτίβα που παρατηρούνται στα δεδομένα.

Στο κεφάλαιο “Θεωρητική Τεκμηρίωση” συνοψίσαμε τη θεωρία της εξόρυξης δεδομένων και αναφέραμε το παράδειγμα του Supermarket. Με τους σύγχρονους αλγόριθμους μηχανικής μάθησης η διαδικασία αυτή γίνεται πολύ αποτελεσματικά, ειδικά με αλγόριθμους μάθησης χωρίς επίβλεψη. Ωστόσο, για να μπορέσει ένα σύστημα να μάθει, χρειάζεται να του δοθούν κάποια δεδομένα (ακόμη και χωρίς πληροφορία κατηγορίας), έτσι ώστε να τα επεξεργαστεί. Αναλόγως του τι θέλουμε να μάθει το σύστημα, χρησιμοποιούμε διαφορετικούς αλγόριθμους για την εξαγωγή των χαρακτηριστικών τα οποία δίνονται ως δεδομένα εκπαίδευσης και αξιολόγησης του μοντέλου. Για παράδειγμα θα χρησιμοποιήσουμε διαφορετικό αλγόριθμο εξαγωγής χαρακτηριστικών εάν θέλουμε το σύστημα μας να μαθαίνει να ξεχωρίζει πρόσωπα από ότι αν θέλουμε το σύστημα μας να κατηγοριοποιεί ιστοσελίδες. Στην πρώτη περίπτωση ένας ιδεατός αλγόριθμος θα μπορούσε για την εξαγωγή των χαρακτηριστικών προσώπου να αναγνωρίζει τα μάτια και το περίγραμμα της κεφαλής και να παίρνει ως χαρακτηριστικά την απόσταση ματιών με μύτη και απόσταση ματιών με στόμα. Αυτός ο αλγόριθμος θα δίνει διαφορετικά χαρακτηριστικά και θα επεξεργάζεται διαφορετικά το περιεχόμενο από ένα αλγόριθμο ο οποίος αναγνωρίζει την ύπαρξη ενός καρκινικού κυττάρου σε μια ανάλυση αίματος.

Είναι ξεκάθαρο λοιπόν πως όσο πιο αποτελεσματικός αλγόριθμος χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών αλλά και όσο πιο αποδοτικός αλγόριθμος χρησιμοποιηθεί στην μηχανική μάθηση και

3 Μάθηση χωρίς επίβλεψη ή αλλιώς ομαδοποίηση

ταξινόμηση δεδομένων τόσο πιο ορθά και ακριβή αποτελέσματα θα έχουμε.

Η παρούσα έρευνα δεν εστιάζει στην βελτίωση των αλγορίθμων ταξινόμησης και ομαδοποίησης αφού αυτοί έχουν δοκιμαστεί επί σειρά ετών με επιτυχία και η περαιτέρω βελτίωση τους είναι σχεδόν αδύνατη. Εστιάζουμε στους αλγόριθμους εξαγωγής χαρακτηριστικών από ιστοσελίδες πορνογραφίας αφού στον τομέα αυτό δεν έχουν δημοσιευθεί έρευνες οι οποίες να βασίζονται στο περιεχόμενο της ιστοσελίδας, τόσο στο κείμενο όσο και στην δομή. Οι μόνες έρευνες που ασχολήθηκαν με αυτό, μέχρι την συγγραφή αυτής της έρευνας, έχουν εστιάσει μόνο στα πολυμεσικά αντικείμενα της ιστοσελίδας και όχι στην δομή της ιστοσελίδας ή στο κείμενο της, που είναι ο τρόπος που προσεγγίζουμε τις ιστοσελίδες στην παρούσα έρευνα.

Συνοψίζοντας, παρατηρούμε πως ένας συνδυασμός τεχνικών από διάφορες περιοχές είναι απαραίτητος για την εκπόνηση της παρούσας έρευνας. Από την ανάκτηση πληροφορίας θα χρειαστούμε τεχνικές για να εξαγάγουμε σημαντικά δεδομένα από τις ιστοσελίδες, όπως για παράδειγμα η συχνότητα εμφάνισης των λέξεων, από την εξόρυξη δεδομένων θα χρειαστούμε τεχνικές εύρεσης μοτίβων που συνδέουν τα δεδομένα και από τη μηχανική μάθηση θα χρειαστούμε τις τεχνικές για την δημιουργία μοντέλου για αυτόματη κατηγοριοποίηση των ιστοσελίδων.

Επισκόπηση Βιβλιογραφίας

Η έρευνα αυτή έχει ως επίκεντρο την εξαγωγή χαρακτηριστικών από ιστοσελίδες έτσι ώστε να δημιουργηθούν μοντέλα για κάθε μια από τις κατηγορίες των ιστοσελίδων που θα εξετάσουμε και οι ιστοσελίδες να μπορούν να κατηγοριοποιούνται αυτόματα. Η αυτόματη κατηγοριοποίηση ιστοσελίδων δεν είναι καινούριο ερευνητικό πεδίο, ωστόσο υπάρχουν πολλοί παράγοντες που χρήζουν βελτίωσης (Qi & Davison, 2009). Δεν υπάρχουν προϋπάρχουσες έρευνες για την αναπαράσταση μιας ιστοσελίδας σχετικής με πορνογραφία όσο αφοράν το κείμενο και την δομή της ιστοσελίδας έτσι θα εξετάσουμε πιο ευρύτερες έρευνες που ασχολούνται με την κατηγοριοποίηση ιστοσελίδων γενικότερα. Οι προϋπάρχουσες έρευνες που ασχολούνται με την αυτόματη κατηγοριοποίηση ιστοσελίδων εστιάζουν κυρίως στο περιεχόμενο (Qi & Davison, 2009), δηλαδή το τι περιέχει η ιστοσελίδα, και την δομή της ιστοσελίδας (Asirvatham & Ravi, 2001), δηλαδή το πώς είναι ο κώδικας που είναι γραμμένη η ιστοσελίδα. Επίσης έχουν μελετηθεί οι μέθοδοι επιλογής των χαρακτηριστικών από τις ιστοσελίδες αφού αυτό είναι ένα κρίσιμο στάδιο που επηρεάζει σημαντικά την ποιότητα των αποτελεσμάτων. Στην συνέχεια συνοψίζουμε τις πιο σχετικές με το αντικείμενο της παρούσας μελέτης έρευνες που έχουν πραγματοποιηθεί στο παρελθόν.

Οι Asirvatham και Ravi (2001), εστιάζουν στην εξαγωγή οπτικών χαρακτηριστικών απο ιστοσελίδες ενώ αναφέρονται μερικώς και στην διαδικασία κατηγοριοποίησης κειμένου. Η κατηγοριοποίηση, βασισμένη σε χαρακτηριστικά κειμένου, γίνεται με δεδομένα εκπαίδευσης Corpus (συλλογές κειμένων) από συγκεκριμένες κατηγορίες. Όταν εφαρμοσθεί stopping[4] και εξαχθούν μόνο οι λέξεις-κλειδιά γίνεται κατηγοριοποίηση μέσω

4 Stopping είναι η διαδικασία αφαίρεσης των πολύ συχνών λέξεων όπως για παράδειγμα: και, το, θα, η

μηχανικής μάθησης με επίβλεψη. Ο αλγόριθμος κατηγοριοποίησης που χρησιμοποιήθηκε είναι ο K-Nearest Neighbor (K-NN). Το μοντέλο που δημιουργήθηκε παρουσίαζε μεγάλη επιτυχία στην κατηγοριοποίηση στα δεδομένα εκπαίδευσης, αλλά η αξιοπιστία της κατηγοριοποίησης στα δεδομένα ελέγχου διαφέρει ανάλογα με τον τύπο των ιστοσελίδων. Μερικά χρήσιμα συμπεράσματα της συγκεκριμένης μελέτης ήταν: (1) Το 94.65% των ιστοσελίδων περιέχει λιγότερες από 500 διαφορετικές λέξεις. (2) Η μέση συχνότητα εμφάνισης κάθε λέξης είναι λιγότερη από δύο. Επομένως η διανυσματική αναπαράσταση των ιστοσελίδων που βασίζεται στη συχνότητα εμφάνισης των λέξεων (tf-idf) δεν είναι ιδανική για κατηγοριοποίηση ιστοσελίδων. (3) Κάποια tags χρησιμοποιούνται σε όλες τις κατηγορίες, όμως τείνουν να έχουν διαφορετική συχνότητα εμφάνισης σε κάποιες κατηγορίες εγγράφων.

Στη συγκεκριμένη έρευνα η ταξινόμηση έγινε σε τρεις κατηγορίες: “Personal Page”, “Informational Page”, “Research Page”. Τα αποτελέσματα έδειξαν ότι οι Informational Pages έχουν περισσότερους υπερσυνδέσμους[5] από τις άλλες δύο κατηγορίες. Οι Research Pages είχαν περισσότερο κείμενο από τις υπόλοιπες κατηγορίες και περιέχουν γραφικές παραστάσεις και γραφήματα. Οι γραφικές παραστάσεις διαχωρίζονται από τις εικόνες με βάση το ιστόγραμμα τους. Χρησιμοποιούνται λίγα χρώματα (ένα μικρό φάσμα του ιστογράμματος), τα οποία μπορούν να αναπαρασταθούν με 4 bit, έχοντας περίπου 4000 χρώματα. Οι Informational Pages έχουν πιο πλούσια χρώματα από τις Personal Pages και αυτές με την σειρά τους πιο πλούσια χρώματα από τις Research Pages, αφού στις Research Pages οι εικόνες είναι γραφήματα, δηλαδή συνθετικές εικόνες. Οι συνθετικές εικόνες έχουν λιγότερα χρώματα ενώ οι φυσικές εικόνες πολύ περισσότερα χρώματα αφού οι διάφορες τιμές χρωμάτων απλώνονται σε όλο το φάσμα των χρωμάτων. Τα Personal Pages εμφανίζουν κοινό layout. Το όνομα, η διεύθυνση και φωτογραφία συνήθως βρίσκονται στο πάνω μέρος της σελίδας ενώ στο κάτω μέρος βρίσκονται σύνδεσμοι προς αγαπημένους προορισμούς διαδικτυακών

5 Υπερσυνδέσμους εννοούμε την ετικέτα στο HTML που μεταφέρει τον χρήστη σε άλλη ιστοσελίδα

χώρων, ή προσωπικές δημοσιεύσεις.

Η εξαγωγή χαρακτηριστικών από τις ιστοσελίδες χωρίζεται σε δύο μέρη. Την εξαγωγή πληροφορίας από το κείμενο, που όμως διαφέρει από την κλασική ανάλυση κειμένου, και την εξαγωγή χαρακτηριστικών από εικόνες. Στην εξαγωγή των χαρακτηριστικών από το κείμενο έλαβαν υπόψη μόνο τη ποσότητα του κειμένου και την οπτική τοποθεσία των συνδέσμων μέσα στην ιστοσελίδα. Στην εξαγωγή των χαρακτηριστικών από εικόνες λήφθηκε υπόψη το ιστόγραμμα. Στην εν λόγω έρευνα, η σωστή κατηγοριοποίηση ιστοσελίδων ήταν 87.83%, ωστόσο δεν αναφέρεται η μέθοδος αξιολόγησης των αποτελεσμάτων, ούτε επεξηγείται η ακριβής διαδικασία κατηγοριοποίησης παρά μόνο δίνονται κάποια σχόλια σε μαθηματικές συναρτήσεις για πολλαπλασιασμό μαθηματικών πινάκων (Matrix).

Οι ερευνητές Tsukada, Washio και Motoda, (2001) έχουν παρατηρήσει ότι συγκεκριμένα ουσιαστικά τείνουν να εμφανίζονται σε ιστοσελίδες που ανήκουν σε συγκεκριμένες κατηγορίες. Χρησιμοποιήθηκε ένας αλγόριθμος εξαγωγής χαρακτηριστικών όπου αφού διαγραφούν όλα τα HTML tags (“<a href>” και “<img>”), εφαρμόζεται μορφολογική ανάλυση και εξάγονται όλα τα ουσιαστικά. Η μορφολογική ανάλυση είναι μια τεχνική με την οποία μια πρόταση αναλύεται σε μέρη του λόγου (ουσιαστικά, επίθετα και επιρρήματα). Για την επεξεργασία κειμένου εφαρμόσθηκε stemming[6] και stopping ενώ για την κατηγοριοποίηση χρησιμοποιήθηκε δυαδική ταξινόμηση (binary classification) με τη βοήθεια δένδρων απόφασης (decision tree algorithms). Επίσης η ταξινόμηση βασίζεται σε decision tree algorithm που θεωρήθηκε ως η καλύτερη επιλογή. Η έρευνα εστιάζει στο κειμενικό περιεχόμενο της ιστοσελίδας και χρησιμοποιεί basket analysis με μαθηματικές συναρτήσεις και τους αλγόριθμους association analysis από το πεδίο τεχνητής νοημοσύνης. Η αξιολόγηση των αποτελεσμάτων έγινε με την βοήθεια των Error rate, Precision και Recall. Error rate: Μεταξύ 8% και 16%, Precision: ~80%, Recall:~45% (βλ. Πίνακα 2)

6 Stemming είναι η διαδικασία αναγωγής των όρων στην ρίζα τους

data	attribute	Minsup %	*“Arts & Humanities”*			*“Business & Economy”*		
			Error rate	*Recall*	*Precision*	*Error rate*	*Recall*	*Precision*
Sup10	823	10	12.6	50.5	79.2	14.3	56.5	67.6
Sup20	78	20	13.3	44.0	80.8	15.0	45.5	69.6
Sup30	19	30	13.9	32.0	95.3	13.6	45.5	77.4

data	*“Education”*			*“Government”*			*“Health”*		
	Error rate	*Recall*	*Precision*	*Error rate*	*Recall*	*Precision*	*Error rate*	*Recall*	*Precision*
Sup10	8.30	69.0	86.7	13.8	45.5	76.1	8.90	65.0	87.2
Sup20	10.9	65.2	77.5	14.2	38.5	80.4	16.1	46.6	64.7
Sup30	10.4	57.5	86.6	14.5	32.5	86.7	15.3	29.3	83.0

Πίνακας 2: Τιμές μέσων για παραδείγματα ταξινόμησης δέντρων απόφασης

Arts & Humanities	*Business & Economy*	*Education*	*Government*	*Health*
Illustration	Enterprise	Success	Society	Research
renewal	business	classroom	politics	life
image	guide	school	policy	age
reproduction	month	learning	election	environment
without-notice	information	education	opinion	medical
{without-notice, reproduction}	{month, information}	{learning, education}	activity	health
...	...	...	...	...

Πίνακας 3: Κατηγορίες και λέξεις που τις αντιπροσωπεύουν

Στον ίδιο πίνακα παρουσιάζονται χαρακτηριστικές λέξεις για κάθε κατηγορία όπως βρέθηκαν στη συγκεκριμένη μελέτη. (βλέπε Πίνακα 3)

Σύμφωνα με τους Golub και Ardo (2005) ένα ξεκάθαρο χαρακτηριστικό που

υπάρχει στην γλώσσα HTML είναι τα tags, τα οποίο δεν εμφανίζονται στα απλά έγγραφα κειμένου. Έχει αποδειχτεί ότι η χρήση πληροφοριών των tags μπορεί να βοηθήσει στην αποτελεσματικότητα της ταξινόμησης ιστοσελίδων σε κατηγορίες. Στην συγκεκριμένη έρευνα χρησιμοποιήθηκαν τέσσερα στοιχεία: title, headings, metadata και κυρίως κείμενο. Σύμφωνα με τους συγγραφείς τα καλύτερα αποτελέσματα προκύπτουν κατάλληλο συνδυασμό των τεσσάρων αυτών στοιχείων.

Οι Kwon και Lee (Kwon & Lee, 2000), τροποποίησαν τον αλγόριθμο k-Nearest Neighbor για την κατηγοριοποίηση ιστοσελίδων με βάση τα HTML tags. Οι εν λόγω ερευνητές, παρατήρησαν πως, οι όροι που βρίσκονται σε διαφορετικά tags παίρνουν διαφορετικό βάρος. Οι συγγραφείς χωρίζουν τα HTML tags σε τρεις ομάδες όπου δίνουν σε κάθε ομάδα ένα συγκεκριμένο βάρος.

Οι Shen et. al. (2004), προτείνουν την εύρεση της κατηγορίας μιας ιστοσελίδας από την περίληψη εγγράφου του περιεχομένου της. Αν εξαχθεί με κατάλληλο τρόπο η περίληψη του εγγράφου, τότε μπορεί να αναπαραστήσει με ακρίβεια το κύριο θέμα της ιστοσελίδας. Σύμφωνα με τους συγγραφείς η κατηγοριοποίηση της ιστοσελίδας με βάση την περίληψη εγγράφου είναι 10% πιο ακριβής από άλλους ταξινομητές βασισμένους στο περιεχόμενο.

Συμπερασματικά, οι προηγούμενες έρευνες έχουν αναδείξει προβλήματα αλλά και πλεονεκτήματα αλγόριθμων και τεχνικών που χρησιμοποιούνται για την αυτόματη κατηγοριοποίηση ιστοσελίδων. Για παράδειγμα όταν ως χαρακτηριστικό χρησιμοποιείται η συχνότητα εμφάνισης των λέξεων στην ιστοσελίδα τότε τα αποτελέσματα σε πολλές περιπτώσεις δεν είναι ικανοποιητικά. Αν όμως, το χαρακτηριστικό αυτό, συνδυαστεί με άλλα (όπως π.χ. με δομικά χαρακτηριστικά της ιστοσελίδας) τα αποτελέσματα βελτιώνονται σημαντικά. Ακόμη, παρατηρείται κατά την λειτουργική

κατηγοριοποίηση στο περιεχόμενο πολλών ιστοσελίδων υπάρχουν μοτίβα τα οποία βοηθούν στην ταξινόμηση. Ομοίως, σε θεματική κατηγοριοποίηση, συγκεκριμένα ουσιαστικά τείνουν να εμφανίζονται περισσότερο σε συγκεκριμένες κατηγορίες ιστοσελίδων. Επίσης σε μερικές περιπτώσεις αναλόγως του πώς χρησιμοποιείται η περίληψη κειμένου μπορεί να βοηθήσει στην αυτόματη κατηγοριοποίηση ιστοσελίδων. Η απόδοση βάρους στα επιμέρους χαρακτηριστικά αλλά και η μέθοδος set-of-words μπορούν να επηρεάσουν σημαντικά τα αποτελέσματα της ταξινόμησης. Ακόμη, συγκεκριμένα tags και συγκεκριμένες λέξεις που χαρακτηρίζουν συγκεκριμένους τύπους ιστοσελίδων μπορούν να χρησιμοποιηθούν για καλύτερη ταξινόμηση ιστοσελίδων. Όσον αφορά τις τεχνικές μηχανικής μάθησης, δεν μπορούμε να παραλείψουμε ότι οι προϋπάρχουσες έρευνες χρησιμοποιούν την ταξινόμηση γιατί επιτυγχάνει καλύτερα αποτελέσματα από την ομαδοποίηση. Οι αλγόριθμοι εξαγωγής χαρακτηριστικών που χρησιμοποιήθηκαν σε άλλες έρευνες διαφέρουν από τον αλγόριθμο που χρησιμοποιείται σε αυτή την έρευνα αφού οι παράγοντες και ο συνδυασμός τους με διάφορες επιλεγμένες τεχνικές δίνουν μια διαφορετική οπτική στο σύνολο των δεδομένων.

Ερευνητικά Ερωτήματα και Λειτουργικοποίηση

Το βασικό ερευνητικό ερώτημα της παρούσας μελέτης θα μπορούσε να διατυπωθεί ως εξής:

Μπορούν να αναπτυχθούν τεχνικές αυτόματης ταξινόμησης πορνογραφικών ιστοσελίδων με αποτελέσματα συγκρίσιμα με αυτά της χειρωνακτικής μεθόδου δηλαδή της ταξινόμησης των σελίδων από ανθρώπους;

Επιπλέον για διευκόλυνση της επιστημονικής επαλήθευσης και απάντησης στο πιο πάνω ερώτημα μπορούν να διατυπωθούν οι πιο κάτω υποθέσεις:

Εναλλακτική υπόθεση 1: Η αυτόματη ταξινόμηση πορνογραφικών ιστοσελίδων οδηγεί σε συγκρίσιμα αποτελέσματα με αυτά της χειρωνακτικής ταξινόμησης.

Μηδενική υπόθεση 1: Η αυτόματη ταξινόμηση πορνογραφικών ιστοσελίδων δεν οδηγεί σε συγκρίσιμα αποτελέσματα με αυτά της χειρωνακτικής ταξινόμησης.

Στην πιο πάνω θεώρηση η εξαρτημένη μεταβλητή είναι η επίδοση ταξινόμησης ιστοσελίδων.

Ανεξάρτητες μεταβλητές είναι τα χαρακτηριστικά περιγραφής των ιστοσελίδων καθώς και οι αλγόριθμοι μηχανικής μάθησης, αφού αμφότερα επηρεάζουν την ταξινόμηση.

Λειτουργικοποίηση

Η αυτόματη ταξινόμηση ιστοσελίδων αναφέρεται ως η διαδικασία κατά την οποία ένας αλγόριθμος, από το πεδίο της μηχανικής μάθησης, παίρνει αποτελέσματα από ένα αλγόριθμο εξαγωγής χαρακτηριστικών από ιστοσελίδες σαν είσοδο και στην συνέχεια, παράγει ως έξοδο την κατηγορία στην οποία οι ιστοσελίδες ταξινομήθηκαν.

Ο αλγόριθμος εξαγωγής χαρακτηριστικών αναφέρεται ως ένας αλγόριθμος (δηλαδή η σειρά εκτέλεσης συγκεκριμένων βημάτων για την ολοκλήρωση ενός στόχου) που εξάγει δεδομένα και βασίζεται στο περιεχόμενο μιας ιστοσελίδας, τόσο στο κείμενο αλλά και ως προς την δομή. Δομή εννοούμε τον κώδικα XHTML μέσω του οποίου καθορίζεται η οπτική αναπαράσταση της.

Χαρακτηριστικά στις ιστοσελίδες είναι:

- Η κανονικοποιημένη συχνότητα εμφάνισης επιλεγμένων και προκαθορισμένων λέξεων του κειμένου
- Η κανονικοποιημένη συχνότητα εμφάνισης εικόνων
- Η κανονικοποιημένη συχνότητα εμφάνισης μεγάλων εικόνων[7]
- Η κανονικοποιημένη συχνότητα εμφάνισης των υπερσυνδέσμων.
- Η κανονικοποιημένη μορφή μέσου όρου του τύπου tf-df

Συγκρίσιμα εννοούμε τα αποτελέσματα που δεν εμφανίζουν κάποια στατιστικά σημαντική διαφορά.

7 80x80 Εικονοστοιχείων

Μεθοδολογία

Η παρούσα έρευνα αποσκοπούσε στην εξερεύνηση της πραγματικότητας αυστηρά βασισμένη σε αριθμούς και στατιστικά στοιχεία, και όχι σε κατανόηση φαινομένων. Ως εκ τούτου, η μελέτη που εκπονήθηκε ήταν πειραματική. Η ομάδα ελέγχου αποτέλεσε το σύνολο ιστοσελίδων που ταξινόμησαν χειρωνακτικά οι άνθρωποι, και η πειραματική ομάδα το σύνολο ιστοσελίδων που ταξινομήθηκαν από το σύστημα (παρουσιάζεται εκτενώς στην ενότητα “Μεθοδολογία υλοποίησης”) και χρησιμοποιήθηκαν στην εκπαίδευση του συστήματος.

Θεωρητική Τεκμηρίωση

Πριν προχωρήσουμε αναλυτικά στη μεθοδολογία υλοποίησης της παρούσας μελέτης είναι χρήσιμο να διευκρινίσουμε κάποιες θεωρητικές έννοιες.

Η κατηγοριοποίηση ιστοσελίδων μπορεί να γίνει με βάση τρεις οπτικές:

Θεματική κατηγοριοποίηση (subject classification): Εστιάζει στην θεματική της ιστοσελίδας (π.χ. arts, business, sports).

Λειτουργική κατηγοριοποίηση (functional classification): Εστιάζει στον ρόλο που παίζει μια ιστοσελίδα (π.χ. Personal web page, Course Page, Admission Page).

Κατηγοριοποίηση με βάση την άποψη (sentiment classification): Εστιάζει στην στάση του συγγραφέα – δημιουργού της ιστοσελίδας – για ένα συγκεκριμένο θέμα.

Υπάρχουν δύο κατηγορίες τεχνικών μηχανικής μάθησης για να δημιουργηθούν μοντέλα για τις κατηγορίες ιστοσελίδων:

Δυαδική ταξινόμηση (binary classification[8]): Ο ταξινομητής εκπαιδεύεται για να κατατάσσει το στοιχείο (ιστοσελίδα) σε μια από δύο κλάσεις. Για

8 Επίσης αναφέρεται και ως μέθοδος one-against-all

παράδειγμα μια η κλάση ονομάζεται “porn page” και η άλλη κλάση “not porn page”.

Ταξινόμηση σε πολλαπλές κλάσεις (multiclass classification): Ο ταξινόμητης εκπαιδεύεται για να κατατάσσει το στοιχείο (ιστοσελίδα) σε μια κλάση από ένα σύνολο κλάσεων που περιλαμβάνει περισσότερες από δύο κλάσεις (π.χ. να κατατάξει το στοιχεία σε μία από τις ακόλουθες κλάσεις: Arts, Business, Museum, Sport, Computers)

Αναλόγως του αριθμού κλάσεων (κατηγοριών) μπορούμε να πούμε πως κάθε στοιχείο μπορεί να ανήκει σε περισσότερες από μία κλάσεις. Εάν χωρίζουμε τα στοιχεία σε ακριβώς μια τάξη το καθένα τότε έχουμε ταξινόμηση μοναδικής ετικέτας (single-label classification), δηλαδή το κάθε στοιχείο υπάγεται σε μια μόνο κατηγορία. Αν επιτρέψουμε το ίδιο στοιχείο να ταξινομηθεί σε περισσότερες από μια κατηγορίες τότε έχουμε ταξινόμηση πολλαπλής ετικέτας (multilabel classification)

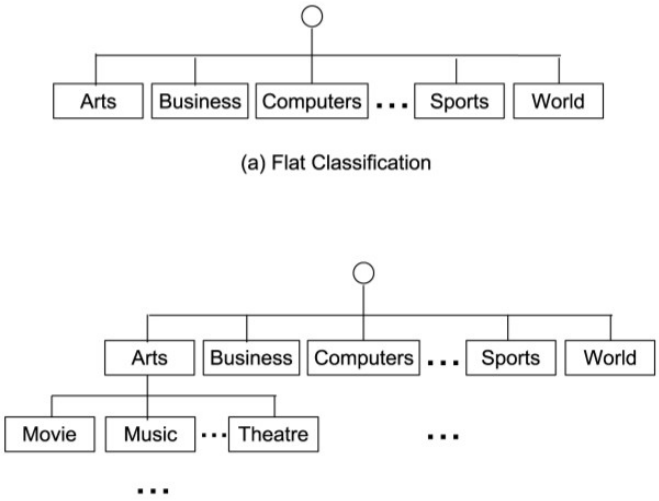
Με βάση τη βεβαιότητα με την οποία ταξινομούμε ένα στοιχείο σε μια κατηγορία διακρίνουμε δύο τύπους ταξινόμησης:

Αυστηρή κατηγοριοποίηση (hard classification): Το στοιχείο είτε ανήκει είτε δεν ανήκει σε μια κλάση, χωρίς ενδιάμεσες τιμές.

Χαλαρή ταξινόμηση (soft classification): Για κάθε στοιχείο υπολογίζεται η πιθανότητα ($0<p<1$) να ανήκει σε μία κατηγορία. Το στοιχείο υπάρχει μια πιθανότητα να ανήκει στην συγκεκριμένη κλάση.

Με βάση τα επίπεδα ταξινόμησης μπορούμε να διακρίνουμε δύο τρόπους ταξινόμησης (βλέπε Σχήμα 2): την Επίπεδη (flat classification) και την ιεραρχική (hierarchical classification). Στην επίπεδη ταξινόμηση οι κατηγορίες είναι ιδίας σημαντικότητας και παράλληλες, δηλαδή δεν υπάρχει κάποια ιεραρχία. Στην ιεραρχική ταξινόμηση υπάρχει ιεραρχία δέντρου, όπου

μια γενικότερη κατηγορία μπορεί να έχει ένα αριθμό ειδικότερων υποκατηγοριών.



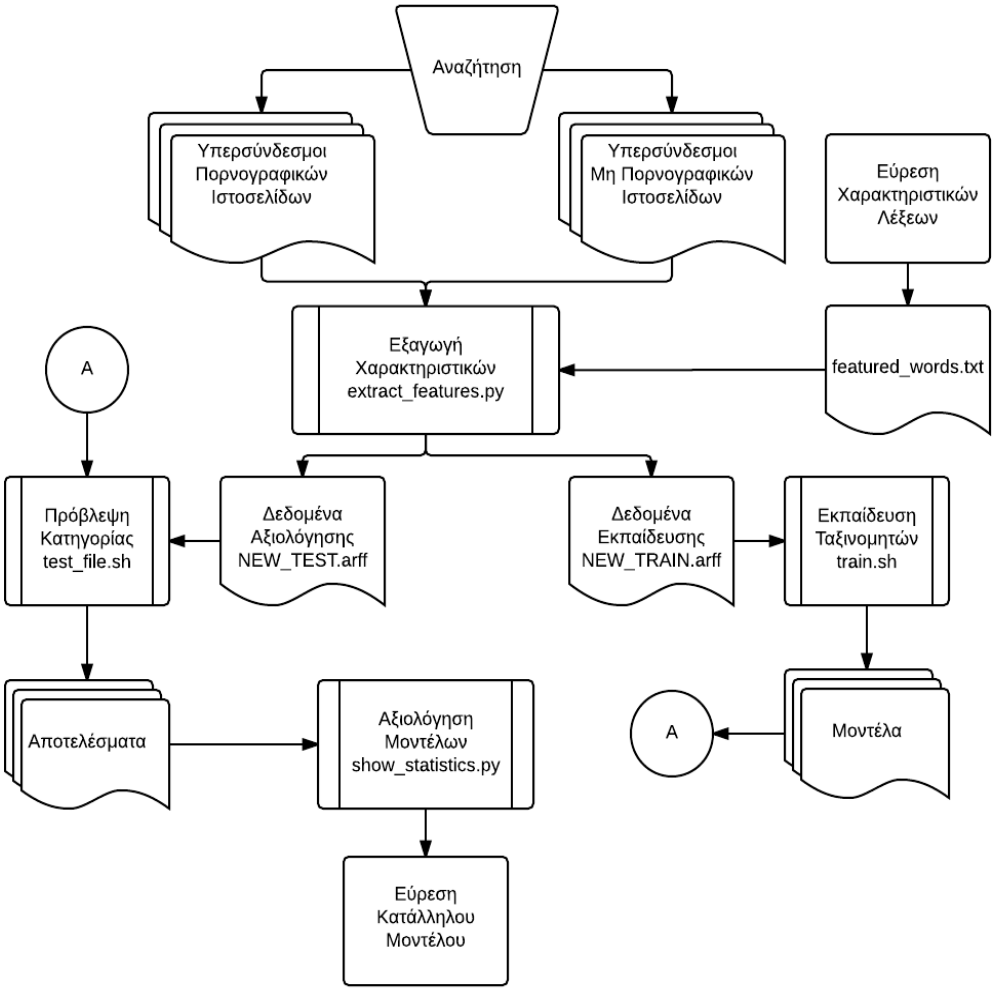
Σχήμα 2: Επίπεδη και ιεραρχική ταξινόμηση

Ευρήματα

Η υλοποίησης μεθοδολογίας συνοψίζεται στο Σχήμα 3 και τα επιμέρους βήματα αναλύονται στη συνέχεια.

Δειγματοληψία

Το δείγμα της έρευνας περιλάμβανε συνολικά 284 υπερσυνδέσμους ιστοσελίδων. Συλλέχθηκαν ιστοσελίδες έτσι ώστε να υπάρχει επαρκής αριθμός ανά κατηγορία και να μπορεί το σύστημα να εκπαιδευτεί αποτελεσματικά μέσω παραδειγμάτων και να πραγματοποιηθεί αξιολόγηση των μοντέλων εκπαίδευσης. Η έρευνα διεξήχθη από τον Ιανουάριο 2015 μέχρι τον Μάιο του 2015 έτσι οι ιστοσελίδες που χρησιμοποιούνται αφορούν το συγκεκριμένο χρονικό διάστημα.



Σχήμα 3: Διάγραμμα Ροής που απεικονίζει τη Μεθοδολογία Υλοποίησης

Αναζήτηση

Η αναζήτηση ιστοσελίδων όσο και η χειρωνακτική κατηγοριοποίηση τους έγινε με την βοήθεια ανθρώπινου παράγοντα. Η αναζήτηση των ιστοσελίδων πραγματοποιήθηκε με την βοήθεια της διαδικτυακής μηχανής αναζήτησης Google. Η προσέγγιση αυτής της έρευνας στοχεύει την αποκλειστική μελέτη των ιστοσελίδων που είναι βελτιστοποιημένες για μηχανές αναζήτησης. Η

επιλογή αυτή έγινε με βάση τη λογική ότι οι περισσότεροι χρήστες χρησιμοποιούν τις διαδικτυακές μηχανές αναζήτησης για να ψάξουν ιστοσελίδες στο διαδίκτυο. Με αυτό τον τρόπο θα επιτρεπόταν η μελέτη των ιστοσελίδων που προκύπτουν από τις πρώτες σελίδες αποτελεσμάτων αναζήτησης της Google, το μέσο δηλαδή που θα χρησιμοποιούσε ένας μέσος χρήστης διαδικτύου. Τα ερωτήματα που χρησιμοποιήσαμε για την αναζήτηση ιστοσελίδων τόσο για εκπαίδευση όσο και για αξιολόγηση περιλάμβαναν τα εξής: “porno”, “porno sites”, “blowjob”, ”porns”, “threesome”, “nude tits”, “rapes” και “sex”. Τα ερωτήματα αναζήτησης για τις ιστοσελίδες εύρεσης λέξεων που χρησιμοποιήθηκαν για την δημιουργία χαρακτηριστικών εκπαίδευσης και αξιολόγησης περιλάμβαναν τα εξής: “rapes” και “sex”. Οι περισσότεροι υπερσύνδεσμοι ιστοσελίδων οδηγούν στην ρίζα της ιστοσελίδας διότι εκεί είναι το κυρίως κείμενο που θέλουμε να εντοπίσουμε όσο και τα ιχνογραφήματα αφού είναι η πρώτη ιστοσελίδα που επισκέπτεται ο χρήστης για να περιηγηθεί στις επόμενες ιστοσελίδες του ιστότοπου (για την περίπτωση των πορνογραφικών ιστοσελίδων η πρώτη σελίδα λειτουργεί και σαν ευρετήριο προς τον χρήστη).

Κριτήριο Αξιολόγησης Χειρωνακτικής Κατηγοριοποίησης

Πορνογραφικές ιστοσελίδες (βλέπε Παράρτημα Α.2) θεωρούμε τις ιστοσελίδες που περιέχουν άσεμνου τύπου πολυμεσικού περιεχομένου με γυμνές γυναίκες ή άνδρες. Ιστοσελίδες οι οποίες επεξηγούν είτε αναφέρονται σε άσεμνο περιεχόμενο μόνο με κείμενο (π.χ. Ιστορίες έρωτα) δεν θεωρούνται πορνογραφικές ιστοσελίδες.

Προετοιμασία Δεδομένων

Οι ιστοσελίδες που επέστρεψε η μηχανή αναζήτησης με τα ερωτήματα στην ενότητα “Αναζήτηση” και ήταν κατηγορίας “πορνογραφικές ιστοσελίδες” δεν συλλέγονταν απαραίτητα όλες. Η συλλογή των δειγμάτων εκπαίδευσης ήταν επιλεκτική. Η λανθασμένη ή τυχαία επιλογή τέτοιων ιστοσελίδων θα είχε ως αποτέλεσμα ο ταξινομητής μας να δημιουργούνταν με λανθασμένο πρότυπο ιστοσελίδων[9] πορνογραφίας. Κατά συνέπεια, έχουμε επιλέξει μερικές ιστοσελίδες (για τον σύνολο ιστοσελίδων εκπαίδευσης και αξιολόγησης) της κατηγορίας “Μη Πορνογραφικές Ιστοσελίδες” να έχουν μερικές λέξεις οι οποίες εμπεριέχονται στα χαρακτηριστικά εκπαίδευσης.

Χειρωνακτική κατηγοριοποίηση

Το πρώτο στάδιο της έρευνας ξεκίνησε με τη συλλογή των ιστοσελίδων στις οποίες θα διεξάγονταν τα πειράματα. Αρχικά επιλέχθηκαν οι κατηγορίες που θέλαμε να γίνει η κατηγοριοποίηση. Καθώς εφαρμόσαμε δυαδική κατηγοριοποίηση οι κατηγορίες είναι “ιστοσελίδες πορνογραφίας” και “ιστοσελίδες μη πορνογραφίας”. Όπως αναφέρθηκε, ήδη οι κατηγορίες ιστοσελίδων μπορούν να συγκεκριμενοποιηθούν ανάλογα με την κατηγοριοποίηση που θα ακολουθηθεί (θεματική ταξινόμηση, λειτουργική ταξινόμηση και κατηγοριοποίηση με βάση την άποψη όπως επεξηγήθηκαν νωρίτερα). Στην παρούσα έρευνα χρησιμοποιήθηκε η θεματική κατηγοριοποίηση. Δημιουργήσαμε μια λίστα για κάθε κατηγορία η οποία περιλάμβανε δύο στήλες, τα URL των ιστοσελίδων και την κατηγορία των ιστοσελίδων. Η διαδικασία αυτή (κατηγοριοποίηση) έγινε χειρωνακτικά, από έναν χρήστη. Θα ήταν πιο συνετό αν επιλέγαμε περισσότερους χρήστες, καθώς είναι πιθανόν οι χρήστες να βάζουν μερικές ιστοσελίδες σε διαφορετικές κατηγορίες. Ωστόσο, ο χρόνος δεν μας επέτρεψε αυτή την

9 πχ. Η ιστοσελίδα είχε μικρού μήκους και πλάτους εικόνες και το περιεχόμενο κειμένου περιείχε λάθος λέξεις από τις επιλεγμένες.

ενέργεια. Σε κάθε ιστοσελίδα δόθηκε μια και μοναδική ετικέτα κατηγορίας (single-label classification). Για το σύνολο ιστοσελίδων εκπαίδευσης έχουν επιλεχθεί 88 ιστοσελίδες πορνογραφικού περιεχομένου και 88 ιστοσελίδες μη πορνογραφικού περιεχομένου. Για το σύνολο ιστοσελίδων αξιολόγησης έχουν επιλεχθεί 54 ιστοσελίδες πορνογραφικού περιεχομένου και 54 ιστοσελίδες μη πορνογραφικού περιεχομένου (βλ. Πίνακα 4).

Πορνογραφικές Ιστοσελίδες Συνόλου Εκπαίδευσης	88
Μη Πορνογραφικές Ιστοσελίδες Συνόλου Εκπαίδευσης	88
Πορνογραφικές Ιστοσελίδες Συνόλου Αξιολόγησης	54
Μη Πορνογραφικές Ιστοσελίδες Συνόλου Αξιολόγησης	54

Πίνακας 4: Διαχωρισμός Συνόλου Ιστοσελίδων ανά κατηγορία

Περιορισμός χαρακτηριστικών

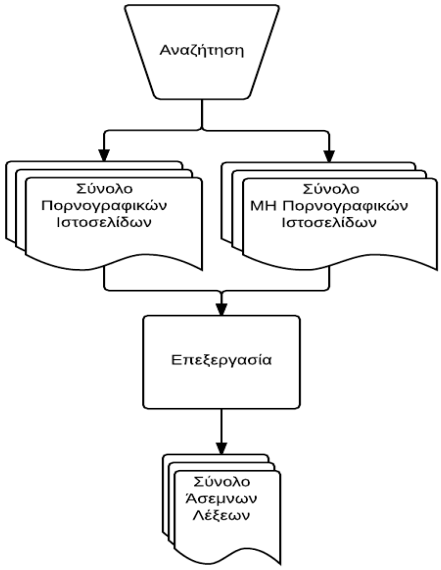
Ένας περιορισμός των χαρακτηριστικών είναι να μην αναλυθούν τα meta tags, ως προς το κείμενο του attribute τους. Ειδικότερα, με βάση την έρευνα των Asirvatham και Ravi (Asirvatham & Ravi, 2001), ο δημιουργός της ιστοσελίδας, ενδέχεται να βάζει λέξεις που να μην αντικατοπτρίζουν την ιστοσελίδα και συνεπώς υπάρχει ο κίνδυνος να οδηγηθούμε σε λάθος συμπεράσματα. Για την αναπαράσταση του κειμένου χρησιμοποιήθηκαν χαρακτηριστικά τύπου bag-of-words (δεν χρησιμοποιήσαμε τη θέση των λέξεων – set-of-words - παρά μόνο αν εμφανίζονται στο κείμενο και πόσες φορές εμφανίζονται).

Χαρακτηριστικά Ιστοσελίδων

Είναι αναγκαίο να επιλεγούν χαρακτηριστικά αναπαράστασης ιστοσελίδων έτσι ώστε να μπορούμε αργότερα να εκπαιδεύσουμε τον ταξινομητή να προβλέπει τον τύπο κατηγορίας της ιστοσελίδας (“Ιστοσελίδες πορνογραφίας” ή “Ιστοσελίδες μη πορνογραφίας”).

Επιλογή λέξεων αναπαράστασης ιστοσελίδων

Σε αυτό το σημείο έπρεπε να επιλέξουμε τις λέξεις που θα αναπαριστούν τις πορνογραφικές ιστοσελίδες όσον αφορά το μέρος του περιεχομένου κειμένου. Για τον εντοπισμό των λέξεων χρησιμοποιήσαμε ένα πρόγραμμα σε python (βλέπε Κώδικας 1, Σχήμα 4.0).



Σχήμα 4: Διαδικασία εξαγωγής λέξεων αναπαράστασης πορνογραφικών ιστοσελίδων

Με την χρήση του αρχείου *“common.py”* (βλέπε Κώδικας 1.0) που χρησιμοποιήσαμε μπορούμε να παρατηρήσουμε ποιες λέξεις των πορνογραφικών ιστοσελίδων και μη πορνογραφικών ιστοσελίδων εμφανίζονται από κοινού. Επίσης μπορούμε να δούμε ποιες λέξεις υπάρχουν μόνο στις πορνογραφικές ιστοσελίδες. Ταυτόχρονα μπορούμε να δούμε και τις λέξεις που εμφανίζονται μόνο στις μη πορνογραφικές ιστοσελίδες. Το αρχείο *“common.py”* εκτελεί το αρχείο “*methods.py*” (βλέπε Κώδικας 1.1). Μας βοηθά να μάθουμε ποιες λέξεις περιέχει κάθε ιστοσελίδα και σε ποια κατηγορία ανήκει. Αυτό θα μας δώσει μια ευρύτερη εικόνα των λέξεων που τείνουν να εμφανίζονται στην κατηγορία πορνογραφικών ιστοσελίδων. Από τις λέξεις που υπήρχαν από κοινού επιλέξαμε κυρίως τις λέξεις που είναι

άσεμνες και έχουν υψηλή διανυσματική αναπαράσταση tf-df και υψηλό αριθμό σε εμφάνιση πορνογραφικών ιστοσελίδων. Ένα άλλο κριτήριο επιλογής των λέξεων ήταν: ο αριθμός εμφάνισης της κοινής[10] λέξης να είναι κυρίως μεγαλύτερος του δεκαπέντε. Υπήρχαν περισσότερες λέξεις προς επιλογή, όμως ο μεγάλος αριθμός χαρακτηριστικών τείνει να δημιουργεί αναξιόπιστα μοντέλα ταξινόμησης, έτσι επιλέξαμε μόνο 44 λέξεις οι οποίες θεωρούμε πως χαρακτηρίζουν και αντιπροσωπεύουν τις πορνογραφικές ιστοσελίδες. Η συχνότητα εμφάνισης των λέξεων που παρουσιάζονται στον Πίνακα 5 είναι μέρος του διανύσματος χαρακτηριστικών όπου θα δοθεί για την εκπαίδευση του ταξινομητή.

10 Να υπάρχει και στις δύο λίστες – πορνογραφικές ιστοσελίδες και μη πορνογραφικές ιστοσελίδες

Λέξη	Κατηγορία					
	Πορνογραφικές Ιστοσελίδες			Μη Πορνογραφικές Ιστοσελίδες		
	Αρ. Εμφάνισης Λέξης	Tf-df	Συχν. Εμφάνισης σε tag “<a>”	Αρ. Εμφάνισης Λέξης	Tf-df	Συχν. Εμφάνισης σε tag “<a>”
amateur	59	0.006	292	4	0.000	6
anal	74	0.007	563	49	0.002	93
asian	64	0.003	218	8	0.000	7
ass	67	0.004	246	29	0.001	20
babe	52	0.002	101	3	0.000	0
bisexual	28	0.000	34	21	0.000	11
blonde	60	0.003	186	9	0.000	1
blowjob	50	0.002	138	17	0.001	17
boobs	37	0.001	96	4	0.000	1
cock	64	0.004	244	19	0.001	7
dick	43	0.001	75	14	0.000	2
fingering	24	0.000	32	15	0.000	16
forced	39	0.009	409	8	0.000	1
fuck	76	0.008	478	26	0.001	4
fucked	60	0.003	194	8	0.000	3
fucking	51	0.002	162	21	0.000	7
fucks	45	0.001	95	2	0.000	0
hairy	40	0.001	108	2	0.000	0
handjob	35	0.001	57	13	0.000	15
hardcore	52	0.002	113	22	0.000	26
horny	36	0.001	50	13	0.000	16
huge	39	0.001	94	11	0.000	2
japanese	47	0.002	161	5	0.000	9
latina	38	0.001	60	6	0.000	6

lesbian	58	0.003	265	30	0.001	54
masturbating	25	0.000	41	8	0.000	5
masturbation	24	0.001	32	40	0.002	140
milf	59	0.003	156	3	0.000	3
nude	29	0.001	74	10	0.000	7
porno	46	0.002	138	3	0.000	2
pornos	1	0.000	1	0	0.000	0
pornostars	1	0.000	1	0	0.000	0
pussy	63	0.004	283	16	0.000	3
russian	43	0.002	124	2	0.000	1
sexy	53	0.002	109	25	0.001	12
slut	37	0.001	61	8	0.000	3
sperm	17	0.000	17	11	0.000	12
stripper	14	0.000	15	0	0.000	0
suck	31	0.001	37	7	0.000	1
sucking	35	0.001	64	12	0.000	2
threesome	37	0.001	51	17	0.001	17
tits	52	0.004	317	11	0.000	1
tube	66	0.016	1260	5	0.000	1
xxx	50	0.005	355	17	0.000	2

Πίνακας 5: Λέξεις αναπαράστασης πορνογραφικών ιστοσελίδων

Υπάρχουν μερικές λέξεις οι οποίες βάσει των αριθμών δεν θα έπρεπε να επιλεχθούν όπως είναι για παράδειγμα οι λέξεις “masturbation” και “anal”. Η υψηλή συχνότητα των λέξεων αυτών σε μη πορνογραφικές ιστοσελίδες οφείλεται στο ότι έχουμε επιλέξει σκόπιμα μη πορνογραφικές ιστοσελίδες οι οποίες περιέχουν κείμενο το οποίο σχετίζεται με ιστορίες έρωτα. Προσπαθήσαμε να χρησιμοποιήσουμε την ψηλή εμφάνισης του “<a>” tag (κυρίως σε πορνογραφικές ιστοσελίδες) ως ξεχωριστό χαρακτηριστικό για κάθε χαρακτηριστική λέξη που είχε αρνητικά αποτελέσματα.

Εξαγωγή χαρακτηριστικών

Όπως έχει παρατηρηθεί και στις προηγούμενες έρευνες που έχουν αναφερθεί, οι ιστοσελίδες μπορούν να αναπαρασταθούν με δύο τρόπους. Με βάση το περιεχόμενο τους, δηλαδή από το κείμενο που περιέχουν, και με βάση τη δομή τους, δηλαδή από τα tags και τον HTML κώδικα που περιέχουν.

Η ανάλυση του περιεχομένου που περιέχει μια ιστοσελίδα είναι μια πολύ εύκολη διαδικασία για τον άνθρωπο και μπορεί να γίνει πολύ αποτελεσματικά για τις πλείστες ιστοσελίδες. Για τον υπολογιστή όμως είναι μια πολύπλοκη διαδικασία με διαφορετικά αποτελέσματα αναλόγως της μεθόδου (ή του συνδυασμού πολλών μεθόδων) που χρησιμοποιείται. Η δομή συγκεκριμένων κατηγοριών ιστοσελίδων τείνει να κρύβει κάποια χαρακτηριστικά που την αντιπροσωπεύουν. Όσον αφορά την δομή, ο στόχος μας στην φάση εξαγωγής χαρακτηριστικών ήταν να βρούμε αυτά τα χαρακτηριστικά για την κάθε ιστοσελίδα.

Όσον αφορά την εξαγωγή χαρακτηριστικών από το κείμενο χρησιμοποιήσαμε διάφορες απλές τεχνικές από το πεδίο της επεξεργασίας φυσικής γλώσσας και ανάκτησης πληροφορίας. Ακόμη, έχουν εξεταστεί τα χαρακτηριστικά από την έρευνα των Golub και Ardo (2005), τα χαρακτηριστικά title, metadata και headings και τα οποία δεν βοήθησαν στην ταξινόμηση. Έτσι αποφασίσαμε να μην τα τοποθετήσουμε στα χαρακτηριστικά για καλύτερη επίδοση των ταξινομητών.

Ο αλγόριθμος εξαγωγής χαρακτηριστικών είναι πολύ σημαντικός για την κατηγοριοποίηση ιστοσελίδων αφού αυτός θα δώσει τα στοιχεία εισόδου του αλγορίθμου ταξινόμησης. Όσο καλύτερος είναι ο αλγόριθμος εξαγωγής χαρακτηριστικών και όσο καλύτερα συνδυαστεί με τον κατάλληλο αλγόριθμο ταξινόμησης, (εφόσον δοθούν κατάλληλα δεδομένα εκπαίδευσης) τόσο καλύτερα αποτελέσματα θα έχουμε. Ένας περιορισμός των χαρακτηριστικών είναι να μην αναλυθούν τα meta tags, ως προς το κείμενο του attribute τους,

γιατί ο δημιουργός της ιστοσελίδας, βάσει της έρευνας των Asirvatham και Ravi (Asirvatham & Ravi, 2001), υπάρχει η πιθανότητα να βάζει λέξεις που να μην αντικατοπτρίζουν την ιστοσελίδα, και υπάρχει κίνδυνος να οδηγούμαστε σε λάθος συμπεράσματα. Για την αναπαράσταση του κειμένου θα χρησιμοποιηθούν χαρακτηριστικά τύπου bag-of-words (δεν θα χρησιμοποιήσουμε τη θέση των λέξεων –set -of words- παρά μόνο αν εμφανίζονται στο κείμενο και πόσες φορές εμφανίζονται) και διανυσματική αναπαράσταση τύπου tf-df, μια παραλλαγή του tf-idf (Croft et al., 2010).

Υπολογισμός Χαρακτηριστικών

Τα χαρακτηριστικά αναπαράστασης ιστοσελίδας που επιλέξαμε είναι:

Ο μέσος όρος της διανυσματικής αναπαράστασης τύπου tf-df

Κανονικοποιημένη συχνότητα εμφάνισης των tags “<a>”

Κανονικοποιημένη συχνότητα εμφάνισης εικόνων

Κανονικοποιημένη συχνότητα εμφάνισης εικόνων συγκεκριμένου μεγέθους

Κανονικοποιημένες συχνότητες εμφάνισης των λέξεων αναπαράστασης

Μέσος όρος διανυσματικής αναπαράστασης τύπου tf-df

Ο τύπος tf-df είναι η συχνότητα εμφάνισης της λέξης προς τον συνολικό αριθμό λέξεων ιστοσελίδας επί τον αριθμό ιστοσελίδων που περιέχουν την λέξη προς τον συνολικό αριθμό ιστοσελίδων. Ο μέσος όρος του tf-df είναι το άθροισμα του tf-df των χαρακτηριστικών λέξεων προς τον συνολικό αριθμό χαρακτηριστικών λέξεων. Ο τύπος tf-df είναι μια παραλλαγή του τύπου tf-idf.

Όπως έχουμε παρατηρήσει στον Πίνακα 5 το να περιλαμβάνεται μια λέξη στα tags “<a>” δεν παίζει ιδιαίτερο ρόλο (εφόσον υπάρχει χαμηλή συχνότητα εμφάνισης των λέξεων σε αυτά τα tags). Έτσι αποφασίσαμε να μην συμπεριλάβουμε την κάθε περίπτωση (“<a>” tag – λέξη) ως ξεχωριστό χαρακτηριστικό στην αναπαράσταση πορνογραφικών ιστοσελίδων αλλά ως σύνολο.

Η κανονικοποιημένη συχνότητα εμφάνισης των tags “<a>” είναι ο αριθμός

εμφάνισης των tags “<a>” που έχουν ως anchor text την λέξη αναπαράστασης προς τον συνολικό αριθμό των tags “<a>”

Η κανονικοποιημένη συχνότητα εμφάνισης εικόνων είναι ο αριθμός εμφάνισης του tag “<img>” προς τον συνολικό αριθμό tags.

Η κανονικοποιημένη συχνότητα εμφάνισης εικόνων συγκεκριμένου μεγέθους είναι ο αριθμός εμφάνισης tag “<img>” με τις ιδιότητες πλάτους και ύψους μεγαλύτερου των ογδόντα εικονοστοιχείων. Ο λόγος που επιλέξαμε ογδόντα εικονοστοιχεία για πλάτος και ύψος είναι επειδή είναι μικρότερος αριθμός από το πλάτος και ύψος ιχνογραφήματος των πορνογραφικών ιστοσελίδων.

Η κανονικοποιημένη συχνότητα εμφάνισης της λέξης είναι η συχνότητα εμφάνισης της λέξης προς τον συνολικό αριθμό των λέξεων της ιστοσελίδας.

Εργαλεία ανάλυσης δεδομένων

Για την εκπόνηση της παρούσας μελέτης θα χρησιμοποιήσουμε μια σειρά από εργαλεία λογισμικού τα οποία παρατίθενται συνοπτικά πιο κάτω:

Python[11]

Η python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου. Θα χρησιμοποιηθεί κυρίως για τον προγραμματισμό των αλγορίθμων εξαγωγής χαρακτηριστικών.

Natural Language Toolkit[12]

Αποτελεί μια βιβλιοθήκη συναρτήσεων Python και χρησιμοποιείται για επεξεργασία φυσικής γλώσσας. Σε αυτή την έρευνα θα χρησιμοποιηθεί για την διευκόλυνση της διανυσματικής αναπαράστασης του περιεχομένου των ιστοσελίδων.

11 https://www.python.org/
12 http://www.nltk.org/

BeautifulSoup[13]

Αποτελεί μια βιβλιοθήκη συναρτήσεων Python και χρησιμοποιείται για την ανάλυση της δομής (αλλά και του περιεχομένου) ιστοσελίδων. Θα χρησιμοποιήσουμε τη βιβλιοθήκη αυτή για συλλογή στοιχείων που αφορούν τη δομή των ιστοσελίδων.

Weka[14]

Το Weka είναι μια πλατφόρμα με αλγορίθμους αναγνώρισης προτύπων και μηχανικής μάθησης υλοποιημένους σε Java. Παρέχει τις δυνατότητες δημιουργίας μοντέλων αλλά και ταξινόμησης δεδομένων με βάση τα μοντέλα αυτά (δέχεται ως είσοδο τα χαρακτηριστικά και τα επεξεργάζεται βάση συγκεκριμένων αλγόριθμων μηχανικής μάθησης και έξοδος του είναι τα ταξινομημένα δεδομένα εισόδου). Επίσης η αποτελεσματικότητα της ταξινόμησης του βασίζεται σε ερευνητικά αποδεκτές μεθόδους αξιολόγησης. Στην παρούσα εργασία το Weka χρησιμοποιήθηκε για δημιουργία των μοντέλων για τις κατηγορίες των ιστοσελίδων μέσω διαφόρων αλγορίθμων μηχανικής μάθησης που παρέχει. Επίσης η αξιολόγηση της δημιουργίας μοντέλων εκπαίδευσης του συστήματος επιτυγχάνεται μέσω των μετρικών Mean Absolute Error, Root mean squared error, Relative absolut error, Root relative squared error, Kappa statistic από το Weka.

Matlab

Η Matlab είναι μια πλατφόρμα που περιέχει πολλές μαθηματικές συναρτήσεις για προσομοίωση συστημάτων σε πολλά ερευνητικά πεδία. Το πεδίο που μας ενδιαφέρει είναι η μηχανική μάθηση. Χρησιμοποιήθηκε στην τελική αξιολόγηση του συστήματος μέσω του t-test.

---

13 http://www.crummy.com/software/BeautifulSoup/
14 http://www.cs.waikato.ac.nz/ml/weka/

Αλγόριθμοι εκπαίδευσης

Αλγόριθμος J48

Ο J48 αλγόριθμος αποτελεί υλοποίηση ανοικτού κώδικα σε Java του αλγόριθμου C4.5 μέσα στο εργαλείο εξόρυξης δεδομένων weka. Ο C4.5 είναι ένα πρόγραμμα που δημιουργεί ένα δέντρο απόφασης βασισμένο σε ένα σύνολο από δεδομένα εισόδου με ετικέτες. Αυτός ο αλγόριθμος κατασκευάστηκε από τον Ross Quinlan. Τα δέντρα αποφάσεων που δημιουργούνται από τον C4.5, μπορούν να χρησιμοποιηθούν για ταξινόμηση. Για τον λόγο αυτό αναφέρεται συχνά σαν ένας στατιστικός ταξινομητής.

Αλγόριθμος Bayes Net

Ένα δίκτυο Bayes είναι ένας τρόπος να αναπαριστά την από κοινού κατανομή ενός συνόλου μεταβλητών με έναν τρόπο που είναι ιδιαίτερα χρήσιμος για την αναπαράσταση γνώσης. Για παράδειγμα η περιγραφή ενός σερβίτσιου για ένα γεύμα. Όλα τα αντικείμενα, δηλαδή το πιάτο, το σκεύος, η χαρτοπετσέτα, το μπολ, η κούπα και τα υλικά που μπορεί να είναι φτιαγμένα τα πιο πάνω, είναι μεταβλητές. Τα κύρια αντικείμενα υποδηλώνουν τα τμήματα της ρύθμισης του χώρου. Η μεταβλητή ρύθμιση έχει τέσσερις πιθανές τιμές (πρωινό, γεύμα, δείπνο, επιδόρπιο) που δηλώνουν τα αντίστοιχα είδη των γευμάτων. Ένα πιάτο μπορεί να είναι χάρτινο ή κεραμικό. Επίσης, υπάρχει πιθανότητα, τα αντικείμενα αυτά να έχουν σχέση μεταξύ τους. Για παράδειγμα, μια χαρτοπετσέτα αναμένεται σε κάθε ένα από τα πιθανά γεύματα (Rimey & Brown, 1992).

Αλγόριθμος SMO

Διαδοχική ελάχιστη βελτιστοποίηση (SMO) είναι ένας αλγόριθμος για την αποτελεσματική επίλυση του προβλήματος βελτιστοποίησης που προκύπτει κατά τη διάρκεια της εκπαίδευσης των μηχανών με διανύσματα

υποστήριξης. Ο SMO σπάει το πρόβλημα σε μια σειρά από μικρότερα επιμέρους προβλήματα, τα οποία στη συνέχεια λύνονται αναλυτικά. Εφευρέθηκε από τον John Platt το 1998. Χρησιμοποιείται ευρέως σε μηχανές εκπαίδευσης διανυσμάτων.

Ταξινόμηση με δέντρα απόφασης

Τα δέντρα αποφάσεων είναι εποπτευόμενοι αλγόριθμοι που χωρίζουν τα δεδομένα αναδρομικά, βασισμένοι στα χαρακτηριστικά των δεδομένων , μέχρι να ικανοποιηθεί μια τερματική συνθήκη. Ο ταξινομητής δένδρου απόφασης είναι μια από τις πιθανές προσεγγίσεις για πολυεπίπεδη λήψη αποφάσεων. Το πιο σημαντικό στοιχείο του ταξινομητή δένδρου απόφασης είναι η ικανότητα διάσπασης της διαδικασίας απόφασης σε μια συλλογή από πιο απλές αποφάσεις, με τέτοιο τρόπο, έτσι ώστε να παρέχουν μια λύση που είναι συνήθως πιο εύκολο να μεταφραστεί. (Βλαχάβας, κ.α., 2006)

Αλγόριθμος Naive Bayes

Είναι ένας πιθανοτικός ταξινομητής που βασίζεται στο θεώρημα Bayes. Επίσης εξετάζει μια ισχυρή παραδοχή ανεξαρτησίας. Ο ταξινομητής αυτός θεωρεί ότι όλα τα χαρακτηριστικά ανεξάρτητα συμβάλλουν στην πιθανότητα μιας συγκεκριμένης απόφασης. Λαμβάνοντας υπόψη τη φύση του υποκείμενου μοντέλου πιθανοτήτων, ο αφελής ταξινομητής Bayes μπορεί να εκπαιδευτεί πολύ αποτελεσματικά σε επιτηρούμενο περιβάλλον μάθησης και να εργάζεται πολύ καλύτερα σε πολλές σύνθετες καταστάσεις του πραγματικού κόσμου. Επειδή οι μεταβλητές θεωρούνται ανεξάρτητες, μόνο οι διακυμάνσεις των μεταβλητών για κάθε κλάση πρέπει να προσδιορίζονται και όχι ολόκληρη η μήτρα συνδιακύμανσης (Aruna, Rajagopalan & Nandakishore, 2011).

Αλγόριθμος K-Star

Ο αλγόριθμος K-Star είναι ένας ταξινομητής βασισμένος σε στιγμιότυπα, κι έτσι η κλάση ενός δοκιμαστικού στιγμιότυπου βασίζεται στην κλάση παρόμοιων στιγμιότυπων εκπαίδευσης, όπως καθορίζεται με κάποια συνάρτηση ομοιότητας. Διαφέρει από άλλους ταξινομητές βασισμένους σε στιγμιότυπα στο ότι χρησιμοποιεί μια συνάρτηση απόστασης βασισμένη στην εντροπία (Cleary & Trigg, 1995).

Τεχνητά νευρωνικά δίκτυα

Ο αλγόριθμος αυτός βασίζεται στην θεωρία των νευρωνικών δικτύων πρόσθιας τροφοδότησης. Επιπρόσθετα βασίζεται στην μέθοδο ανάστροφη μετάδοση λάθους. Η ανάστροφη μετάδοση λάθους είναι γνωστή μέθοδος για την εκπαίδευση νευρωνικών δικτύων πολλών επιπέδων. Βασίζεται στον γενικευμένο κανόνα δέλτα. Η διαδικασία του κύκλου εκπαίδευσης του MLP αποτελείται από δύο στάδια. Εισάγονται στην είσοδο δεδομένα (από ένα διάνυσμα εκπαίδευσης) και τροφοδοτούν ένα κρυφό επίπεδο. Η έξοδος του κρυφού επίπεδο είναι η τροφοδοσία ενός άλλους κρυφού επίπεδου και η διαδικασία επαναλαμβάνεται διαδοχικά μέχρι το επίπεδο εξόδου. (Βλαχάβας, κ.α., 2006).

Αλγόριθμος Multilayer Perceptron (MLP)

Ο αλγόριθμος multilayer perceptron ανήκει στην κατηγορία τεχνητών νευρωνικών δικτύων. Παρέχει ένα εύκολο τρόπο για την εκμάθηση αριθμητικών και διανυσματικών συναρτήσεων. Χρησιμοποιούνται τόσο για παρεμβολή όσο και για ταξινόμηση. Το πλεονέκτημα των νευρωνικών δικτύων (και συνεπώς των αλγόριθμων που βασίζονται στην θεωρία αυτή) είναι η μεγάλη ανοχή τους σε δεδομένα εκπαίδευσης με λανθασμένες τιμές. Τα νευρωνικά δίκτυα χωρίζονται σε τρία κύρια επίπεδα. Το επίπεδο εισόδου, το κρυφό επίπεδο και το επίπεδο εξόδου. Το επίπεδο εισόδου χρησιμοποιείται

για την εισαγωγή των δεδομένων. Στο κρυφό επίπεδο μπορούν να υπάρχουν ένα ή περισσότερα επίπεδα (δηλαδή πολυεπίπεδο νευρωνικό δίκτυο). Επίσης σε αυτό το επίπεδο μπορούν να υπάρχουν κρυφά επίπεδα. (Βλαχάβας, Ι., κ.α., 2006)

Αλγόριθμος εξαγωγής χαρακτηριστικών

Ο αλγόριθμος εξαγωγής χαρακτηριστικών είναι πολύ σημαντικός για την κατηγοριοποίηση ιστοσελίδων αφού αυτός θα δώσει τα στοιχεία εισόδου του αλγόριθμου ταξινόμησης. Όσο καλύτερος είναι ο αλγόριθμος εξαγωγής χαρακτηριστικών και όσο καλύτερα συνδυαστεί με τον κατάλληλο αλγόριθμο ταξινόμησης[15], τόσο πιο ποιοτικά αποτελέσματα θα υπάρξουν.

Επιλογή Λέξεων

Η εξαγωγή των λέξεων από μια ιστοσελίδα γίνεται με τον εξής τρόπο:

Αρχικά, αφαιρούνται τα tags απο το κείμενο. Έπειτα, παράγονται tokens από τον διαχωρισμό των κενών του κειμένου. Στη συνέχεια το κάθε token δίνεται ως όρισμα στην συνάρτηση onlyWords() (βλ. Κώδικα 1.1) και επιστρέφονται μόνο οι χαρακτήρες του αγγλικού αλφαβήτου, σε μορφή κεφαλαίων ή πεζών. Εαν το μήκος της λέξης που επιστρέφεται είναι ίσο ή μεγαλύτερο από τρία, τότε η λέξη προστίθεται στην λίστα του συνόλου των λέξεων της ιστοσελίδας.

Αρχείο εξαγωγής χαρακτηριστικών “extract_features.py”

Εφόσον εκτελέσουμε το αρχείο extract_features.py (βλέπε Κώδικας 3.0) με όρισμα “True” ξεκινά η διαδικασία λήψης των ιστοσελίδων και εξαγωγής των χαρακτηριστικών των ιστοσελίδων αξιολόγησης σε μορφή arff (αποθηκεύεται στο αρχείο “NEW_TEST.arff”). Όταν δοθεί όρισμα “False” τότε ξεκινά η διαδικασία εξαγωγής χαρακτηριστικών ιστοσελίδων εκπαίδευσης σε

15 Εφόσον δοθούν ποιοτικά δεδομένα εκπαίδευσης

μορφή arff όπου αποθηκεύεται στο αρχείο “NEW_TRAIN.arff”. Αν δεν δοθεί όρισμα ξεκινά η διαδικασία εξαγωγής χαρακτηριστικών ιστοσελίδων για αξιολόγηση (και δημιουργείται το αρχείο “NEW_TEST.arff”).

Περιγραφή λειτουργίας αρχείου εξαγωγής χαρακτηριστικών

1. Φορτώνεται το αρχείο “methods.py” όπου περιέχει τις απαραίτητες συναρτήσεις
2. Φορτώνονται από τον σκληρό δίσκο οι λέξεις αναπαράστασης πορνογραφικών ιστοσελίδων (βλέπε Κώδικα 10.0) και δημιουργείται μια ενιαία λίστα
3. Στην συνέχεια φορτώνονται τα κατάλληλα αρχεία που περιέχουν τους υπερσύνδεσμους πορνογραφικών και μη πορνογραφικών ιστοσελίδων
4. Επομένως, γίνεται λήψη της κάθε ιστοσελίδας
5. Ακολουθεί εξαγωγή χαρακτηριστικών της κάθε ιστοσελίδας αμέσως μετά την λήψη της
6. Τέλος, δημιουργείται το αρχείο τύπου arff και αποθηκεύεται στον σκληρό δίσκο.

Περιγραφή συναρτήσεων εξαγωγής χαρακτηριστικών

Οι συναρτήσεις βρίσκονται στο αρχείο “methods.py” όπου χρησιμοποιείται από το κυρίως πρόγραμμα στο αρχείο “extract_features.py”.

onlyWords(token): Δέχεται ως όρισμα ένα token και επιστρέφει μόνο τα γράμματα αλφαβήτου και αφαιρεί οποιοδήποτε άλλο χαρακτήρα.

loadFile(path,filename): Δέχεται το σχετικό μονοπάτι ενός αρχείου και το όνομα του αρχείου. Επιστρέφει σε λίστα, τις γραμμές του αρχείου.

loadStopList(): Φορτώνει το αρχείο “stoplist.txt” όπου βρίσκεται στο ίδιο σχετικό μονοπάτι με το αρχείο εκτέλεσης και επιστρέφει μία λίστα με τα stopwords.

ApplyStopping(dic): Δέχεται ως όρισμα ένα λεξικό και αφού καλέσει την loadStopList(), αφαιρεί τα κλειδιά του λεξικού τα οποία είναι στην λίστα

όπου περιέχει τα stopwords.

findDuplicates(dict1, dict2): Δέχεται δύο λεξικά και τυπώνει οποιαδήποτε κλειδιά των λεξικών είναι κοινά.

getImageNum(soup): Δέχεται ως όρισμα ένα αντικείμενο της κλάσης BeautifulSoup και επιστρέφει τον αριθμό εικόνων που βρέθηκαν. Οι εικόνες εντοπίζονται από το tag “<img>”. Η συνάρτηση αυτή εντοπίζει τις ιδιότητες μεγέθους και μήκους. Για να μετρηθεί ως έγκυρη η εικόνα είναι απαραίτητο οι ιδιότητες μήκους και πλάτους να υπάρχουν και τα εικονοστοιχεία του μήκους όσο και του πλάτους να είναι μεγαλύτερα των ογδόντα εικονοστοιχείων.

getImageNum(soup): Δέχεται ως όρισμα ένα αντικείμενο της κλάσης BeautifulSoup και επιστρέφει τον αριθμό των εικόνων (tag “<img>”).

hasVideoPlayer(soup): Δέχεται ως όρισμα ένα αντικείμενο της κλάσης BeautifulSoup και επιστρέφει τον αριθμό των tags “<embed>”. Η συνάρτηση χρησιμοποιήθηκε για να ελέγξουμε αν έχει κάποιο αντίκτυπο στην εκπαίδευση του ταξινομητή όμως είχαμε αρνητικά αποτελέσματα.

generate_tfdf(word,site,wordsIndex,structIndex,dw,n): Παίρνει ως όρισμα την λέξη που θα δημιουργηθεί η αναπαράσταση tfdf, ο υπερσύνδεσμος της ιστοσελίδας, το λεξικό wodsIndex που περιέχει χαρακτηριστικά λέξεων (συχνότητα εμφάνισης), το λεξικό structIndex το οποίο περιέχει χαρακτηριστικά της ιστοσελίδας, ο αριθμός των ιστοσελίδων που περιλαμβάνουν την λέξη που δόθηκε ως όρισμα και ο συνολικός αριθμός ιστοσελίδων. Στην συνέχεια υπολογίζει την αναπαράσταση tfdf όπου και την επιστρέφει.

getARFFData(pornolist, instClass): Δέχεται ως όρισμα ένα λεξικό το οποίο περιέχει τους υπερσύδεσμους ιστοσελίδων πορνογραφικών ή μη πορνογραφικών ιστοσελίδων. Το δεύτερο όρισμα της συνάρτησης είναι η κλάση όπου θα κατατάξει τα παραδείγματα που θα δημιουργήσει από τις πληροφορίες που περιέχονται σε νέα λεξικά που δημιουργεί. Στην συνέχεια επεξεργάζεται τις πληροφορίες που παίρνει κατεβάζοντας μια μια τις

ιστοσελίδες και δημιουργεί τα παραδείγματα (διανύσματα χαρακτηριστικών) όπου τα επιστρέφει με την μορφή arff.

site_total_tfdf(site,tfdf_dict): Δέχεται ως όρισμα ένα υπερσύνδεσμο ιστοσελίδας, και ένα λεξικό όπου θα αποθηκευτούν οι συνολικοί μέσοι όροι των τιμών tfdf. Εσωτερικά χρησιμοποιεί την ενιαία μεταβλητή featured_words_num όπου περιέχει τις λέξεις χαρακτηριστικών. Στην συνέχεια υπολογίζει τους μέσους όρους των αναπαραστάσεων tfdf και τους αποθηκεύει στο λεξικό tfdf_dict. Επιστρέφεται ο μέσος όρος αναπαράστασης tfdf της συγκεκριμένης ιστοσελίδας που δόθηκε.

readSite(siteurl): Δίνεται ως όρισμα ένας υπερσύνδεσμος ιστοσελίδας και επιστρέφεται το κείμενο της ιστοσελίδας σε μορφή αλφαριθμητικού.

getCommonWords(pornolist): Η συνάρτηση αυτή παίρνει μια λίστα με υπερσυνδέσμους ιστοσελίδων και εφόσον κατεβάσει την κάθε ιστοσελίδα βρίσκει τις κοινές λέξεις τους, δημιουργεί ένα λεξικό και το επιστρέφει.

cleanedAtagTokens(soup): Παίρνει ως όρισμα το αντικείμενο της κλάσης BeautifulSoup και στην συνέχεια επιστρέφει σε μία λίστα τις λέξεις που εμπεριέχονται στα “<a>” tags.

getWebsiteWords(soup, txt): Παίρνει ως όρισμα το αντικείμενο της κλάσης BeautifulSoup και το κείμενο της ιστοσελίδας και επιστρέφει σε λίστα τις λέξεις της ιστοσελίδας.

getWordInAtag(soup, word): Δίνεται το αντικείμενο BeautifulSoup και μια λέξη και επιστρέφεται η συχνότητα εμφάνισης της σε “<a>” tags.

getAtagTokens(soup): Δίνεται το αντικείμενο BeautifulSoup και επιστρέφεται ο αριθμός των “<a>” tags.

Κριτήρια αξιολόγησης αλγόριθμων ταξινόμησης

Τα κριτήρια αξιολόγησης της επίδοσης των διαφόρων ταξινομητών που εξετάζονται στα πλαίσια της παρούσας έρευνας είναι τα ακόλουθα: Mean Absolute Error, Root mean squared error, Relative absolut error, Root relative

squared error, Kappa statistic.

Error rate: Το error rate είναι η διαφορά της πραγματικής τιμής (χειρωνακτικής κατηγοριοποίηση – πορνογραφική ιστοσελίδα) από την προβλέψιμη τιμή. (Για παράδειγμα: Αν η πραγματική τιμή είναι 1 και η τιμή πρόβλεψης είναι 0.8 τότε η διαφορά είναι 0.2). Ο τύπος του Error rate δίνεται από την πιο κάτω εξίσωση:

$$ER = 1 - P$$

Όπου το P είναι η τιμή πρόβλεψης.

Mean absolute error: Είναι ο μέσος όρος της απόλυτης τιμής του αθροίσματος του error rate των περιπτώσεων παραδείγματος. Ο τύπος του Mean absolute error δίνεται από την πιο κάτω εξίσωση:

$$\frac{\left(\sum_{i}^{n} 1 - P\right)}{n}$$

Όπου το P είναι η τιμή πρόβλεψης, το i το στοιχείο και το n ο συνολικός αριθμός στοιχείων.

Root mean squared error: Είναι η ρίζα του μέσου όρου του αθροίσματος του error rate στην δύναμη του 2 για κάθε μια περίπτωση παραδείγματος. Ο τύπος του Root mean squared error δίνεται από την πιο κάτω εξίσωση:

$$\sqrt{\frac{\sum_{i}^{n} (1 - P)^2}{n}}$$

Όπου το P είναι η τιμή πρόβλεψης, το i το στοιχείο και το n ο συνολικός αριθμός στοιχείων.

Relative absolute error: Είναι η απόλυτη τιμή του μέσου όρου του αθροίσματος του error rate προς την πραγματική τιμή (χειρωνακτικής κατηγοριοποίησης) των παραδειγμάτων. Ο τύπος του Relative absolute error

δίνεται από την πιο κάτω εξίσωση:

$$\left(\frac{\sum_{i}^{n} 1-P}{n} \right)$$

Όπου το P είναι η τιμή πρόβλεψης, το i το στοιχείο και το n ο συνολικός αριθμός στοιχείων.

Root relative squared error: Είναι η ρίζα του μέσου όρου του αθροίσματος του error rate προς την πραγματική τιμή (χειρωνακτικής κατηγοριοποίησης) των παραδειγμάτων. Ο τύπος του Root relative squared error δίνεται από την πιο κάτω εξίσωση:

$$\sqrt{\frac{\sum_{i}^{n} 1-P}{n}}$$

Όπου το P είναι η τιμή πρόβλεψης, το i το στοιχείο και το n ο συνολικός αριθμός στοιχείων.

Kappa Statistic: Είναι η πιθανότητα το αποτέλεσμα της πρόβλεψης του παραδείγματος (instance) να είναι τυχαίο. Οι τύποι του kappa statistics δίνονται από τις πιο κάτω εξισώσεις:

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy}$$

$$totalAccuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$randomAccuracy = \frac{(TN + FP)(TN + FN) + (FN + TP) * (FP + TP)}{Total * Total}$$

Όπου το TN είναι τα πραγματικά αρνητικά στοιχεία, το TP είναι τα πραγματικά θετικά στοιχεία, το FP είναι τα θετικά στοιχεία πρόβλεψης και το FN είναι τα αρνητικά στοιχεία πρόβλεψης.

Εκπαίδευση συστήματος και δημιουργία μοντέλων

Στο επόμενο στάδιο οι αναπαραστάσεις των ιστοσελίδων μέσω χαρακτηριστικών χρησιμοποιήθηκαν για τη δημιουργία μοντέλων για τις δύο κατηγορίες (πορνογραφικές ιστοσελίδες, μη πορνογραφικές ιστοσελίδες). Για το σκοπό αυτό χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης που διατίθενται μέσω του λογισμικού Weka (Witten & Frank, 2005). Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι: Multilayer Perceptron, Sequential minimal optimization (SMO), BayesNet, NaiveBayes, J48 και KSTAR. Το είδος μηχανικής μάθησης που επιλέξαμε αναφέρεται στη μάθηση με επίβλεψη γιατί με βάση την έρευνα των Qi και Davison (Qi & Davison, 2009) είναι πιο αποτελεσματική για το συγκεκριμένο πρόβλημα (κατηγοριοποίηση ιστοσελίδων). Επίσης σύμφωνα με τους Asirvatham και Ravi (Asirvatham & Ravi, 2001) η ομαδοποίηση – μάθηση χωρίς επίβλεψη – είναι μια διαδικασία με μεγάλο υπολογιστικό κόστος ενώ χρειάζεται εκ των προτέρων να είναι ο γνωστός ο αριθμός των κατηγοριών στις οποίες θα γίνει η ταξινόμηση. Για τη δημιουργία των μοντέλων συλλέχθηκαν συνολικά 176 Ιστοσελίδες (88 για ιστοσελίδες πορνογραφίας και 88 για μη πορνογραφικές ιστοσελίδες). Το Σχήμα 3.0 βοηθά να κατανοήσουμε καλύτερα την διαδικασία εκπαίδευσης και αξιολόγησης του συστήματος.

Εφόσον υπάρχουν πλέον τα διανύσματα χαρακτηριστικών για εκπαίδευση μπορούμε να εκπαιδεύσουμε το σύστημα. Την διαδικασία αυτή την έχει αναλάβει μια σειρά εντολών η οποία καλεί το πρόγραμμα weka για την χρησιμοποίηση έξι διαφορετικών αλγορίθμων ταξινόμησης, τους οποίους χρησιμοποιούμε για να δημιουργήσουμε έξι διαφορετικά μοντέλα (ένα μοντέλο για κάθε αλγόριθμο) τα οποία αποθηκεύουμε στον σκληρό δίσκο[16] (βλέπε Κώδικας 6.0). Οι προκαθορισμένοι παράμετροι για τους αλγόριθμους ταξινόμησης δεν έχουν τροποποιηθεί. Η εκπαίδευση των ταξινομητών έχει δημιουργηθεί με το flag no-cv έτσι ώστε να μην εφαρμόσει cross validation, δηλαδή τα δείγματα να χρησιμοποιηθούν όλα για εκπαίδευση. Η αξιολόγηση

16 http://cis.cut.ac.cy/~Daniel/new_train_models.zip

των μοντέλων εκπαίδευσης με βάση τα δείγματα της ίδιας της εκπαίδευσης είναι αναξιόπιστη αφού τα αποτελέσματα της πρόβλεψης είναι η αντανάκλαση των δεδομένων για τα οποία έχει εκπαιδευτεί (βλέπε Πίνακα 6).

	Multilayer Perceptron	**SMO**	**Bayes Net**	**Naive Bayes**	**J48**	**Kstar**
Correctly Classified Instances	98.3%	97.1 %	96.6 %	98.3%	98.3 %	100%
Incorrectly Classified Instances	1.7 %	2.9 %	3.4 %	1.7 %	1.7 %	0 %
Mean absolute error	0.0283	0.0286	0.0345	0.0171	0.0289	0.0047
Root mean squared error	0.1289	0.169	0.1851	0.1309	0.1202	0.0285
Relative absolute error	5.6 %	5.7 %	6.9 %	3.4 %	5.7 %	0.9 %
Root relative squared error	25.8 %	33.8 %	37.0 %	26.2 %	24.0 %	5.7 %
Kappa statistic	0.9657	0.9428	0.9314	0.9657	0.9657	1

Πίνακας 6: Στατιστικά εκπαίδευσης ταξινομητών

Αυτόματη Ταξινόμηση

Εφόσον έχουμε τα μοντέλα που έχουν εκπαιδευτεί να προβλέπουν και να διακρίνουν τα δεδομένα σε δύο κατηγορίες (“ιστοσελίδες πορνογραφίας” και “ιστοσελίδες μη πορνογραφίας”) μπορούμε στη συνέχεια να αξιολογήσουμε τα μοντέλα μας και να επιλέξουμε το πιο αξιόπιστο μέσω ενός νέου συνόλου διανυσματικών αναπαραστάσεων ιστοσελίδων – το σύνολο ελέγχου, το οποίο είναι διαφορετικό από το σύνολο εκπαίδευσης.

Το σύνολο εκπαίδευσης είναι το σύνολο των παραδειγμάτων για τα οποία η πληροφορία της κατηγορίας στην οποία ανήκει η ιστοσελίδα είναι διαθέσιμη

και χρησιμοποιήθηκε κατά την εκπαίδευση.

Το σύνολο ελέγχου είναι το σύνολο των παραδειγμάτων τα οποία δεν χρησιμοποιήθηκαν στην εκπαίδευση του συστήματος (δημιουργία μοντέλων). Η πληροφορία της κατηγορίας στην οποία ανήκει η ιστοσελίδα είναι διαθέσιμη ώστε να μπορεί να ελεγχθεί κατά πόσο το σύστημα μπορεί να την προβλέψει σωστά (χωρίς προφανώς να την γνωρίζει).

Αξιολόγηση επίδοσης συστήματος

Για την αξιολόγηση της επίδοσης του συστήματος χρησιμοποιήθηκαν συνολικά 108 ιστοσελίδες εκ των οποίων οι 54 ήταν ιστοσελίδες πορνογραφίας και οι υπόλοιπες 54 μη πορνογραφίας. Για τα δεδομένα εκπαίδευσης βλέπε Κώδικα 5.0 και για τα δεδομένα ελέγχου βλέπε Κώδικα 5.1. Για την διαδικασία της πρόβλεψης κατηγορίας ιστοσελίδας χρησιμοποιήσαμε το εργαλείο weka και η διαδικασία αυτοματοποιήθηκε μέσα από μια σειρά εντολών γλώσσας bash (βλ. Κώδικα 7.0). Η πληροφορία της κατηγορίας στην οποία ανήκει η ιστοσελίδα είναι διαθέσιμη ώστε να μπορεί να ελεγχθεί κατά πόσο το σύστημα μπορεί να την προβλέψει σωστά (χωρίς προφανώς να τη γνωρίζει). Με την βοήθεια του μοντέλου της κατηγορίας έγινε πρόβλεψη της κατηγορίας κάθε ιστοσελίδας (ανατίθεται η ετικέτα κατηγορίας που αντιστοιχεί στο μεγαλύτερο σκορ – για παράδειγμα αν έχουμε σκορ 0.1 τότε ανατίθεται η ετικέτα “ιστοσελίδα μη πορνογραφικού περιεχομένου” ενώ για 0.9 ανατίθεται η ετικέτα “ιστοσελίδα πορνογραφικού περιεχομένου”). Από τα αποτελέσματα που προκύπτουν από το αρχείο “test_file.py” χρησιμοποιούμε το αρχείο “show_statistics.py” (βλ. Κώδικα 8.0) έτσι ώστε να αποφασίσουμε ποιος αλγόριθμος και παράλληλα ποιο μοντέλο είναι το κατάλληλο. Ο καλύτερος αλγόριθμος (που έχει τις καλύτερες επιδόσεις στην πρόβλεψη κατηγορίας) είναι ο Bayes Net. Στον Πίνακα 7 παρατηρούμε τις επιδόσεις όλων των αλγορίθμων. Στην συνέχεια, για κάθε μια από τις ιστοσελίδες στο σύνολο ελέγχου (δεδομένα αξιολόγησης) οι προβλέψεις συγκρίνονται με τις κατηγορίες που έδωσε ένας άνθρωπος (χειρωνακτική κατηγοριοποίηση) και τα αποτελέσματα συγκρίνονται με τη

βοήθεια της μεθόδου t-test για να διαπιστώσουμε κατά πόσον υπάρχει ή όχι στατιστικά σημαντική διαφορά μεταξύ τους. Έτσι μπορούμε να υποστηρίξουμε την ερευνητική υπόθεση απορρίπτοντας την μηδενική ή το αντίστροφο. Η διαδικασία αυτή επιτυγχάνεται με την βοήθεια του αρχείου “extract_vectors.py” (βλέπε Κώδικα 9.0) όπου η έξοδος του είναι το αποτέλεσμα της αξιολόγησης (αρχείο BN). Στην συνέχεια δημιουργούμε δύο διανύσματα τα οποία χρησιμοποιούνται στην συνάρτηση ttest(x,y) του εργαλείου matlab για τον υπολογισμό του p-value και της υπόθεσης (H).

Αλγόριθμος	**Ποσοστό Σωστών Προβλέψεων**	**Ποσοστό Λανθασμένων Προβλέψεων**
Multilayer Perceptron	92.5%	7.5%
SMO	95.3%	4.7%
Bayes Net	99.1%	0.9 %
Naive Bayes	98.1%	1.9%
J48	97.2%	2.8%
Kstar	78.5%	21.5%

Πίνακας 7: Επιδόσεις αξιολόγησης αλγόριθμων ταξινόμησης

Είδος ταξινόμησης

Η ταξινόμηση τύπου πολλαπλής ετικέτας (multilabel classification) καθώς και η χαλαρή ταξινόμηση (soft classification) χαρακτηρίζουν καλύτερα το τι πραγματικά συμβαίνει σε μια ιστοσελίδα ως προς την κατηγορία της. Παρόλα αυτά σύμφωνα με τους Qi και Davison (Qi & Davison, 2009) οι περισσότερες έρευνες σε αυτό το πεδίο χρησιμοποιούν αυστηρή ταξινόμηση (hard classification) και ταξινόμηση μοναδικής ετικέτας (single label class) ανά έγγραφο. Επειδή οι δυνατότητες μας στο χρονικό περιθώριο που μας δίνεται δεν μπορούν να προχωρήσουν την έρευνα σε αυτό το επίπεδο έχουμε επιλέξει αυστηρή ταξινόμηση και ταξινόμηση μοναδικής ετικέτας.

Όσον αφορά την ιεραρχία των κατηγοριών, επιλέξαμε επίπεδη ταξινόμηση

(flat classification) γιατί η ιεραρχική ταξινόμηση (hierarchical classification), χρειάζεται μια πολύπλοκη διαδικασία για την ανάλυση κειμένου στις ιστοσελίδες. Η ανάλυση αυτή εμπίπτει στο πεδίο επεξεργασίας φυσικής γλώσσας και δεν είναι το κύριο αντικείμενο μελέτης της παρούσας έρευνας καθώς χρειάζεται περισσότερος χρόνος και περισσότερη εμβάθυνση της έρευνας στον τομέα αυτό.

Έλεγχος Σύγκρισης Μέσων (t-test)

Το t στατιστικό τέστ ανεξάρτητων δειγμάτων (independent measures t-test) αποτελεί μια στατιστική μέθοδο μέσα από την οποία χρησιμοποιούνται δεδομένα από δύο διαφορετικά δείγματα για τον έλεγχο υποθέσεων σε σχέση με την διαφορά των μέσων τιμών που παρουσιάζονται ανάμεσα στους πληθυσμούς των δειγμάτων (Κατσάνος & Αβούρης, 2008).

Η παρούσα έρευνα εξετάζει την πιθανή διαφορά ανάμεσα στις μέσες τιμές που προέρχονται από τον ανθρώπινο παράγοντα και την πρόβλεψη του υπολογιστικού συστήματος. Ως εκ τούτου το εν λόγο t-test κρίθηκε ως το καταλληλότερο για τον έλεγχο αυτής της στατιστικής υπόθεσης.

Χρησιμοποιώντας το εργαλείο Matlab μπορέσαμε να υπολογίσουμε το p-value και το Hypothesis Value (H). Εκτελώντας το αρχείο “extract_vectors.py” (βλέπε Κώδικας 9.0) εξασφαλίσαμε τα διανύσματα της χειρωνακτικής κατηγοριοποίησης ανθρώπου και αυτόματης κατηγοριοποίησης από τον υπολογιστή αφού μας δείχνει την κάθε περίπτωση ξεχωριστά. Στην συνέχεια τα εντάξαμε στο εργαλείο matlab και πήραμε τα εξής αποτελέσματα:

Το Hypothesis value (H) είναι: 0 (η υπόθεση ισχύει) και p-value: 0.1080.

Παρατηρούμε λοιπόν πως τα αποτελέσματα δεν είναι στατιστικά σημαντικά (εφόσον το p-value είναι μεγαλύτερο του 0.05) όμως, η εναλλακτική υπόθεση μας ισχύει. Υπάρχει η πιθανότητα 11% η υπόθεση που υποστηρίζουμε να μην ισχύει.

Συμπεράσματα

Η παρούσα ερεύνα παρουσίασε έναν αλγόριθμο που παρουσιάζει κατά 99.1% επιτυχή αξιολόγηση των πορνογραφικών ιστοσελίδων, μέσω της μηχανικής μάθησης όμως αυτό δεν ισχύσει στατιστικά γιατί το πείραμα είναι κατά 11% αναξιόπιστο. Αυτό σημαίνει ότι η υπόθεση πως ο υπολογιστής μπορεί να εντοπίσει τις ιστοσελίδες πορνογραφικού περιεχομένου το ίδιο καλά όσο ο άνθρωπο ισχύει, όμως υπάρχει 11% πιθανότητα να υποστηρίζουμε μια υπόθεση που δεν ισχύει. Ο αλγόριθμος της παρούσας έρευνας τείνει να εντοπίζει λανθασμένα ένα μικρό ποσοστό (0.9%) πορνογραφικών ιστοσελίδων. Ωστόσο, είναι προτιμότερο ο υπολογιστής να εντοπίζει ένα ελάχιστο ποσοστό ιστοσελίδων πορνογραφικού περιεχομένου λανθασμένα παρά να εντοπίζει ιστοσελίδες μη πορνογραφικού περιεχομένου ως πορνογραφικές ιστοσελίδες. Εν τέλει, ο αλγόριθμος αυτός μπορεί να επαναληφθεί σε πολλές ακαδημαϊκές μελέτες για να εξετάσουν την πιθανότητα επανάληψης αποτελεσμάτων και την αλλαγή των τεχνικών και μεθόδων που χρησιμοποιήθηκαν. Τα ερευνητικά αποτελέσματα ήταν αρκετά για να υποστηρίξουν την εναλλακτική υπόθεση: “Η αυτόματη ταξινόμηση πορνογραφικών ιστοσελίδων οδηγεί σε συγκρίσιμα αποτελέσματα με αυτά της χειρωνακτικής ταξινόμησης”. Ως εκ τούτου μπορεί να απορριφθεί η μηδενική υπόθεση όπου σύμφωνα με την οποία, η αυτόματη ταξινόμηση πορνογραφικών ιστοσελίδων δεν οδηγεί σε συγκρίσιμα αποτελέσματα με αυτά της χειρωνακτικής ταξινόμησης. Με αυτή την έρευνα έχουμε μάθει πως μια μικρή αλλαγή στον αλγόριθμο εξαγωγής χαρακτηριστικών έχει μεγάλο αντίκτυπο στην δημιουργία σωστών μοντέλων πρόβλεψης πορνογραφικών ιστοσελίδων. Επίσης, έχουμε κατανοήσει τη χρησιμότητα απλών τεχνικών της ανάκτησης πληροφορίας και τις βασικές αρχές των αλγόριθμων της μηχανικής μάθησης στην κατηγοριοποίηση ιστοσελίδων πορνογραφικού περιεχομένου. Επιπρόσθετα, έχουμε δει πως η αξιολόγηση του συστήματος με τα ίδια δεδομένα όπου εκπαιδεύτηκε δημιουργεί αναξιόπιστα στατιστικά αποτελέσματα. Η παρούσα έρευνα παρουσίασε τρόπους κατηγοριοποίησης

ιστοσελίδων πορνογραφίας. Παρόλα αυτά, είναι απαραίτητη η διεξαγωγή περισσότερων ερευνών (μεμονωμένα ή σε συνδυασμό) ώστε να δημιουργηθεί μια στερεότερη βάση για την αυτόματη κατηγοριοποίηση ιστοσελίδων πορνογραφίας μέσω κειμένου και δομής της ιστοσελίδας.

Ηθικά Ζητήματα

Είναι σημαντικό να αναφέρουμε οποιαδήποτε ηθικά ζητήματα που αφορούν την παρούσα έρευνα. Ένα από αυτά είναι η αναληθής επισήμανση ενός ποσοστού ιστοσελίδων κάτω από την κατηγορία "πορνογραφικές ιστοσελίδες". Αυτό μπορεί να παρουσιαστεί λόγω της στατιστικής πιθανότητας που υπάρχει, να κατηγοριοποιήσει ο αλγόριθμος λανθασμένα.

Παράλληλα τίθενται κάποιοι προβληματισμοί στα πλαίσια των κριτικών προσεγγίσεων για την συμμετοχή των χρηστών στο διαδίκτυο σήμερα (Beer, 2009). Οι προβληματισμοί αυτοί αφορούν την εφαρμογή των αλγορίθμων και τεχνικών που χρησιμοποιήθηκαν στην παρούσα ερεύνα, στην σύγχρονη κοινωνία. Αυτή η ερεύνα απέδειξε με το πείραμα ότι ο αυτόματος εντοπισμός πορνογραφικών ιστοσελίδων μπορεί να επιτευχθεί με σχεδόν την ίδια ακρίβεια της χειρωνακτικής μεθόδου, δηλαδή τον εντοπισμό των πορνογραφικών ιστοσελίδων από ανθρώπους. Εύλογα, μπορούμε να πούμε ότι τεχνικές από τα πεδία της τεχνητής νοημοσύνης και εξόρυξης άποψης μπορούν να χρησιμοποιηθούν για να εντοπίσουν άλλους είδους περιεχόμενο, όπως για παράδειγμα περιεχόμενο που συνδέονται με συγκεκριμένες ιδεολογίες. Πόσο ηθικό είναι να περιορίζεται η ανάκτηση τέτοιων πληροφοριών, σε μία εποχή όπου η κοινή γνώμη επηρεάζεται εκτενώς από το διαδίκτυο;

Περιορισμοί Έρευνας

Ένας περιορισμός που προκύπτει αποτελεί η έλλειψη εκτενούς χρησιμοποίησης των αλγορίθμων από το πεδίο της επεξεργασίας φυσικής γλώσσας η οποία θα μπορούσε να βοηθήσει σημαντικά στην εύρεση καλύτερων χαρακτηριστικών για την αντιπροσώπευση του περιεχομένου των ιστοσελίδων. Ένας άλλος περιορισμός είναι η αποφυγή του πολυμεσικού περιεχομένου αφού η επεξεργασία του χρειάζεται περισσότερο χρόνο και διαφορετικού τύπου αλγορίθμων εξαγωγής χαρακτηριστικών (για παράδειγμα ανάλυση εικόνων, ανάλυση βίντεο). Ακόμη, η αποφυγή της ταξινόμησης πολλαπλής ετικέτας και της χαλαρής ταξινόμησης περιορίζει την έρευνα καθώς με την επιλογή των παραγόντων αυτών ίσως να επιτυγχάναμε καλύτερα αποτελέσματα, όμως η εμπειρία του ερευνητή στα συγκεκριμένα πεδία καθιστά δύσκολη τη διερεύνηση του. Δεν θα μπορούσαμε να παραλείψουμε ότι η χειρωνακτική κατηγοριοποίηση έχει διεκπεραιωθεί μόνο από έναν άνθρωπο ενώ θα ήταν συνετό να εκπονηθεί από πολλούς ανθρώπους έτσι ώστε να μετρηθεί και η συμφωνία μεταξύ των ανθρώπων. Οι επιδόσεις των αλγορίθμων της έρευνας θα μπορούσαν να ήταν καλύτερες αν χρησιμοποιούνταν αλγόριθμος επεξεργασίας εικόνων έτσι ώστε να εντοπίζεται το χρώμα δέρματος, τα μέρη του σώματος και τα σχήματα που υπάρχουν στην εικόνα. Άλλο περιορισμό αποτελεί η επικέντρωση της έρευνας σε ιστοσελίδες που είναι γραμμένες στην αγγλική γλώσσα, με επακόλουθο να μην είναι δυνατή η σωστή πρόβλεψη ιστοσελίδων που είναι γραμμένες σε άλλες γλώσσες.

Μελλοντικές Έρευνες

Μελλοντικές μελέτες μπορούν να βασιστούν στην παρούσα έρευνα για την επέκταση του αλγορίθμου εξαγωγής χαρακτηριστικών προσθέτοντας και διαφορετικού τύπου περιεχόμενο όπως και flash, video, audio, applets έτσι ώστε η ιστοσελίδα να αναπαρασταθεί καλύτερα. Επίσης, μελλοντικές έρευνες μπορούν να χρησιμοποιήσουν την παρούσα ερεύνα με στόχο να

βρουν όλες τις πορνογραφικές ιστοσελίδες και στη συνέχεια να βρουν πιθανόν παράνομες ιστοσελίδες ή ακόμη να επικεντρωθούν στον εντοπισμό παιδόφιλων σε ενδο-δίκτυα όπως το The Onion Routers[17]. Τα αποτελέσματα της παρούσας έρευνας υποστηρίζουν την αυτόματη κατηγοριοποίηση πορνογραφικών ιστοσελίδων με επιτυχία με ένα μικρό ποσοστό λάθους. Μια μελλοντική ερεύνα θα μπορούσε να ακολουθήσει την ίδια μεθοδολογία για την δημιουργία μοντέλων που κατηγοριοποιούν ιστοσελίδες σε άλλες γλώσσες όπως για παράδειγμα τα Ισπανικά και τα Γαλλικά.

17 https://www.torproject.org/

Βιβλιογραφία

Αγγλόγλωσση Βιβλιογραφία

Aruna,S., Rajagopalan,S.P. & Nandakishore,L.V. (2011). Knowledge based analysis of various statistical tools in detecting breast. In D.C. Wyld, et al. (CS & IT 02, pp. 37–45). CCSEA Retrieved from http://airccj.org/CSCP/vol1/csit1205.pdf

Asirvatham, A. P., & Ravi, K. K. (2001, December). Web page classification based on document structure. In *IEEE National Convention*.

Beer, D. (2009). Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society*, *11*(6), 985-1002

Bergman, M. K. (2001). White paper: the deep web: surfacing hidden value. *Journal of electronic publishing*, 7(1). DOI: http://dx.doi.org/10.3998/3336451.0007.104

Cleary,G.J. & Trigg,E.L (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. *In:12th International Conference on Machine Learning*. pp 108-114.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Reading: Addison-Wesley.

Golub, K. & Ardo, A. (2005). Importance of HTML structural elements and metadata in automated subject classification. *In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*., vol. 3652. Springer, Berlin, Germany. pp 368–378.

Kwon, O.-W. & Lee, J.-H. (2000). Web page classification based on k-nearest neighbor approach. I*n Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL)*. ACM Press, New York, NY. pp 9–15.

NIST. (2007). Text REtrieval Conference (TREC). http://trec.nist.gov/.

Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys (CSUR)*, 41(2), 12.

Rimey, R. D., & Brown, C. M. (1992, February). Where to look next using a Bayes net: An overview. *In Proceedings of 1992 DARPA Image Understanding Workshop.* pp 927-932.

Shen, D., Chen, Z., Yang, Q., Zeng, H.-J., Zhang, B., LU, Y., & MA, W.-Y. (2004).

Web-page classification through summarization. *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York. pp 242–249.

Tsukada, M., Washio, T., & Motoda, H. (2001). Automatic web-page classification by using machine learning methods. In Web Intelligence: Research and Development. *Springer Berlin Heidelberg*. pp 303-313

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Ελληνόγλωσση Βιβλιογραφία

Βλαχάβας, Ι., Κεφάλας, Π., Βασιλειάδης, Ν., Κόκκορας, Φ., & Σακελλαρίου, Η. (2006). *Τεχνητή Νοημοσύνη*. Εκδόσεις: Γκιούρδας Β, Θεσαλλονίκη.

Κατσάνος, Χ., & Αβούρης, Ν. (2008). *Στατιστικές μέθοδοι ανάλυσης πειραματικών δεδομένων συνεργασίας*. Εκδόσεις Κλειδάριθμος, Αθήνα.

Παράρτημα Α.1: Κώδικας Python

Κώδικας 1.0: Αρχείο “common.py” - Κύριο πρόγραμμα για την εύρεση χαρακτηριστικών

```
# -*- coding: utf-8 -*-
from BeautifulSoup import BeautifulSoup
import urllib
import urllib2
import nltk
import json
import sys
execfile('methods.py')
porn=loadFile("lists/","relevant_choose.txt")
print "File: ","Relevant Porn Sites", "urls #", len(porn)
nporn=loadFile("lists/","irrelevant_choose.txt")
print "File: ","Irrelevant Porn Sites", "urls #", len(nporn)

irel = getCommonWords(nporn)
rel  = getCommonWords(porn)

commonWords={}
relWords={}
irelWords={}
for w in irel.keys():
        if w in rel.keys():
                commonWords[w]={'ir':{},'rel':{}}
                commonWords[w]['ir']=irel[w]
                commonWords[w]['rel']=rel[w]
for w in rel:
        if w not in commonWords.keys():
```

```
                relWords[w]=rel[w]

for w in irel:
        if w not in commonWords.keys():
                irelWords[w]=irel[w]

#save to file
fo=open("relevant.dat", "w")
fo.write(json.dumps(relWords))
fo.close()
#save to file
fo=open("irrelevant.dat", "w")
fo.write(json.dumps(irelWords))
fo.close()
#save to file
fo=open("commons.dat", "w")
fo.write(json.dumps(commonWords))
fo.close()
filer_commonDict(commonWords,15)
```

Κώδικας 1.1: Αρχείο “methods.py” - Εκτελείται από το αρχείο “common,py” και από το αρχείο “extract_features.py” - Περιέχει τις μεθόδους εύρεσης χαρακτηριστικών εκπαίδευσης και μεθόδους εξαγωγής χαρακτηριστικών.

```
def onlyWords(token):
        alpha=list("abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ")
        word=""
        letters=list(token)
        for letter in letters:
                if letter in alpha:
                        word+=letter
        return word
```

```
def loadFile(path,filename):
	txt=open(path+filename, 'r')
	return [i.replace('\n','') for i in txt.readlines()]


def loadStopList():
	data=open('stoplist.txt','r').read()
	slist=data.split("\n")
	flist=[]
	for i in slist:
		i=i.replace(" ","") #remove space - if any
		if len(i)>0:
			flist.append(i)
	return flist


def ApplyStopping(dic):
	stoplist=loadStopList()
	for i in dic.keys():
		if i in stoplist:
			del dic[i]


def findDuplicates(dict1, dict2):
	#we check both dictionaries to make sure we checked urls in dictionaries without slash in the end
	for w in dict1:
		if w[len(w)-1:]=='/':
			w=w[:len(w)-1]
		if w in dict2:
			print "WARNING: DOUBLE URL FOUND! ",w
	for w in dict2:
		if w[len(w)-1:]=='/':
			w=w[:len(w)-1]
		if w in dict1:
			print "WARNING: DOUBLE URL FOUND! ",w
```

```
def getImageNum(soup):
        '''find specific images of specific size 80x80'''
        imgs=soup.findAll("img")
        lst=[]
        for i in imgs:
                try:
                        try:
                                width=i._getAttrMap()['width']
                                height=i._getAttrMap()['height']
                                if int(width) > 80 and int(height) > 80: lst.append(i)
                        except ValueError:
                                try:
                                        width=i._getAttrMap()['width'].split("px")[0]
                                        height=i._getAttrMap()['height'].split("px")[0]
                                except ValueError:
                                        pass
                except KeyError:
                        pass
        return len(lst)

def getAllImageNum(soup):
        imgs=soup.findAll("img")
        return len(imgs)

def hasVideoPlayer(soup):
        player=soup.findAll("embed")
        return len(player)

def getAtags(soup):
        return soup.findAll("a")

def getAtagsNum(soup):
```

```
	return len(soup.findAll("a"))

def getAtagTokens(soup):
	#epistrefei ta tokens pu emperiexonte sta atags tis istoselidas

	totalATAGtokens=[]
	#prosthese apla ta tokens pu iparxun sta tags
	atags=soup.findAll("a")
	for tag in atags:
		text=nltk.clean_html(tag.text)
		if len(text)>=3:
			texttokens=nltk.word_tokenize(text)
			#prostheseta sta tokens pu ine mono gia atags
			totalATAGtokens.extend(texttokens)
	return totalATAGtokens

def getWordInAtag(soup, word):
	#epistrefei posa atags perilamvanun tin sigekrimeni lexi stin istoselida

	c=0
	words_in_atags=getAtagTokens(soup)
	for w in words_in_atags:
		if w==word: c=c+1
	return c

def getWebsiteWords(soup, txt):
	#requires soup and clean_text(without html tags)

	#keep only tokens - remove html
	clean=nltk.clean_html(txt)
	#tokenize website's text
	tmptokens = nltk.word_tokenize(clean)
	tokens=[]
```

```
	#create token list - filtrarei ean to token exei megethos panw apo 3
	for toki in range(len(tmptokens)):
		if len(tmptokens[toki])>=3: tokens.append(tmptokens[toki])



	#epelexe mono ta tokens pu einai lexeis		(totaltokens)
	totalwords=[]
	for token in tokens:
		word=onlyWords(token)
		if len(word)>2: totalwords.append(word)
	return totalwords


def getURLWords(url):
	opener = urllib2.build_opener()
	opener.addheaders = [('User-agent', 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:36.0) Gecko/20100101 Firefox/36.0')]
	txt = opener.open(url).read().lower()
	soup=BeautifulSoup(txt)

	#keep only tokens - remove html
	clean=nltk.clean_html(txt)
	#tokenize website's text
	tmptokens = nltk.word_tokenize(clean)
	tokens=[]
	#create token list - filtrarei ean to token exei megethos panw apo 3
	for toki in range(len(tmptokens)):
		if len(tmptokens[toki])>=3: tokens.append(tmptokens[toki])



	#epelexe mono ta tokens pu einai lexeis		(totaltokens)
	totalwords=[]
	for token in tokens:
		word=onlyWords(token)
```

```
            if len(word)>2: totalwords.append(word)
        return totalwords

def getTotalWords(soup,txt):
        return len(getWebsiteWords(soup, txt))

def getWordFreq(txt,soup,word):
        site_words=getWebsiteWords(soup,txt)
        freq=nltk.FreqDist(site_words)[word]
        return freq

def cleanedAtagTokens(soup):
        #atag tokens - kratise kai katharise ta tokens pu ine mesa se atags (realATAGtokens)

        realATAGtokens=[]
        totalATAGtokens=getAtagTokens(soup)
        for token in totalATAGtokens:
            word=onlyWords(token)
            if len(word)>2: realATAGtokens.append(word)
        return realATAGtokens

def getWordsAndWebsitesData(urls,atagsIndex,pornoDist):
        #epistrefei poses selides periexun tin tade lexi, kai ta atag_occurences gia kathe lexi

        indexDict={}

        for url in urls:
            for word in pornoDist[url].keys():
                #find word in <a> tag for that url
                atags=0
                if atagsIndex[url].has_key(word)==True: atags=atagsIndex[url][word]

                if indexDict.has_key(word)==True:
```

```
                                        if indexDict[word].has_key(url)==False:          #add new url to word
                                                indexDict[word][url]={'words_freq':len(pornoDist[url]), 'freq':pornoDist[url][word], 'atag_occurence':atags, 'tfdf':0}
                                        else:
                                                indexDict[word][url]['words_freq']=len(pornoDist[url])
                                                indexDict[word][url]['freq']=pornoDist[url][word]
                                                indexDict[word][url]['atag_occurence']=atags
                                                indexDict[word][url]['tfdf']=0


                                else:   #if word not exist
                                        indexDict[word]={url:{'words_freq':len(pornoDist[url]), 'freq':pornoDist[url][word], 'atag_occurence':atags, 'tfdf':0}}
        return indexDict



def site_total_tfdf(site,tfdf_dict):
        tfdf_sum=0
        featured_words_num=len(feature_set_words)
        for fw in feature_set_words:
                tfdf_sum+=tfdf_dict[site][fw]
        return tfdf_sum/(featured_words_num+0.0)


def readSite(siteurl):
        try:
                opener = urllib2.build_opener()
                opener.addheaders = [('User-agent', 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:36.0) Gecko/20100101 Firefox/36.0')]
                txt = opener.open(siteurl).read().lower()
        except IOError:
                #remove site url - unable to retrieve
                print "Problematic site -removing-",siteurl
                return None
        return txt
```

```
def generate_tfdf(word,site,wordsIndex,structIndex,dw,n):
	#lexiko gia to tfdf lexeon twn istoselidwn
	freq=wordsIndex[site][word]['word_freq']
	total_words=structIndex[site]['total_words']
	wordinWebsites=dw[word]
	return round((freq/(total_words+0.0))*(wordinWebsites/(n+0.0)),3)
#-----------------------------------------------------------------------------------------------------------
#-----------------------------------------------------------------------------------------------------------
def filer_commonDict(commonWords,limit):
	for w in commonWords.keys():
		if commonWords[w]['rel']['total_sites']>limit:
			print w,commonWords[w]

def getCommonWords(pornolist):
	pornos=pornolist
	structIndex={}
	wordsIndex={}
	final_dict={}

	unique_words={}

	for siteurl in pornos:
		txt=readSite(siteurl)
		if txt==None or txt=='':
			#remove site url - unable to retrieve
			print "Problematic site -removing-",siteurl
			continue

		soup=BeautifulSoup(txt)

		#words of website
		website_words=getWebsiteWords(soup,txt)
```

```
        website_distinct_words=list(set(website_words))

        #sinolo lexeon istoselidas
        totalwords=len(getWebsiteWords(soup,txt))



        structIndex[siteurl]={}
        structIndex[siteurl]['total_words']=totalwords
        structIndex[siteurl]['words_list']=website_distinct_words

        #find word frequency for every featured word of this site
        wordsIndex[siteurl]={}

        for w in website_distinct_words:
                print w
                word_frequency=getWordFreq(txt,soup,w)
                wordsIndex[siteurl][w]={}
                wordsIndex[siteurl][w]['word_freq']=word_frequency
                wordsIndex[siteurl][w]['word_atagsum']=getWordInAtag(soup,w)
                if unique_words.has_key(w):
                        unique_words[w]+=1
                else: unique_words[w]=1

        print totalwords,siteurl

#dimiurgia lexiku lexeon olon ton istoselidon
for w in unique_words.keys():
        final_dict[w]={}

#sinolo selidon
n=len(structIndex.keys())

tfdf_dict={}
```

```
	#generate tfdf for each word for each site
	for site in structIndex.keys():
		for w in wordsIndex[site].keys():
			tf=round(wordsIndex[site][w]['word_freq']/(0.0+structIndex[site]
['total_words']),3)
			if tfdf_dict.has_key(w)==True: tfdf_dict[w]+=tf
			else: tfdf_dict[w]=tf

	#generate average_tfdf
	for w in unique_words.keys():
		final_dict[w]['tfdf']=round((tfdf_dict[w]/(unique_words[w]
+0.0))*(unique_words[w]/(n+0.0)),3)

	######################generate total_atags##################
	########initialize atags sum for each word#####
	#for w in unique_words.keys():
	for w in unique_words.keys():
		final_dict[w]['atagsum']=0

	for site in structIndex.keys():
		site_words=structIndex[site]['words_list']
		final_dict[w]['total_sites']=unique_words[w]

	#Apply stopping
	ApplyStopping(final_dict)

	return final_dict



def getARFFData(pornolist, instClass):
	pornos=pornolist

	structIndex={}
```

```
wordsIndex={}


for siteurl in pornos:
	txt=readSite(siteurl)
	if txt==None:
		#remove site url - unable to retrieve
		print "Problematic site -removing-",siteurl
		continue

	soup=BeautifulSoup(txt)

	#total tags number
	tagsnum=len(soup.findAll())



	#get big and normal_imgs
	BIGimgs=getImageNum(soup)

	normal_imgs=getAllImageNum(soup)
	if normal_imgs>0: BIGimgs_normalized=BIGimgs/(normal_imgs+0.0)
	else: BIGimgs_normalized=0
	normalized_normalimgs=normal_imgs/(tagsnum+0.0)


	#find atag_average_of_featured words in <a> tags
	featured_in_atags_sum=0
	for fw in feature_set_words:
		featured_in_atags_sum+=getWordInAtag(soup,fw)

	#poses lexis iparxun se atags
	words_in_atags=len(cleanedAtagTokens(soup))
```

```
                #poses xaraktiristikes lexis vrethikan sta atag / poses lexis sinolika iparxun se
atags
                if words_in_atags>0:

        normalized_featured_atag_normalized=featured_in_atags_sum/words_in_atags
                else: normalized_featured_atag_normalized=0

                #sinolo lexeon istoselidas
                totalwords=len(getWebsiteWords(soup,txt))

                structIndex[siteurl]={}
                structIndex[siteurl]['BIGimgs']=BIGimgs_normalized
                structIndex[siteurl]['imgs']=normalized_normalimgs
                structIndex[siteurl]['avg_atag_words']=normalized_featured_atag_normalized
                structIndex[siteurl]['total_words']=totalwords

                #find word frequency for every featured word of this site
                wordsIndex[siteurl]={}
                for fw in feature_set_words:
                        word_frequency=getWordFreq(txt,soup,fw)
                        wordsIndex[siteurl][fw]={'word_freq':word_frequency}

                print totalwords,siteurl


        #se poses selides aniki i kathe lexi
        dw={}
        for fw in feature_set_words:
                dw[fw]=0
        for fw in feature_set_words:
                for site in structIndex.keys():
                        if wordsIndex[site][fw]['word_freq']>0:
                                dw[fw]+=1
```

```
#sinolo selidon
n=len(structIndex.keys())

tfdf_dict={}
#generate tfdf for each word for each site
for site in structIndex.keys():
	tfdf_dict[site]={}
	for fw in feature_set_words:
		tfdf_dict[site][fw]=generate_tfdf(fw,site,wordsIndex,structIndex,dw,n)



#calculate tfdf MEAN weight
for site in structIndex.keys():
	structIndex[site]['mean_tfdf']=site_total_tfdf(site,tfdf_dict)

#print arff data
arffData=""

#generate ARFF
for site in structIndex.keys():
	instanceLine=""

	for w in sorted(feature_set_words):
		w_tfdf=tfdf_dict[site][w]
		instanceLine+=str(w_tfdf)+","

	BigImg=str(structIndex[site]['BIGimgs'])
	Img=str(structIndex[site]['imgs'])
	porn_atags_score=str(round(structIndex[site]['avg_atag_words'],3))
	avg_tfdf=str(structIndex[site]['mean_tfdf'])

	#put features in arff format
```

```
                instanceLine+=BigImg+","+Img+","+porn_atags_score+","+avg_tfdf
                instanceLine+=","+instClass    #SET CLASS VARIABLE
                instanceLine+="\n"
                arffData+=instanceLine

        return arffData
for w in site_words:
                        final_dict[w]['atagsum']+=wordsIndex[site][w]['word_atagsum']

        for w in unique_words.keys():
                final_dict[w]['total_sites']=unique_words[w]

        #Apply stopping
        ApplyStopping(final_dict)

        return final_dict



def getARFFData(pornolist, instClass):
        pornos=pornolist

        structIndex={}
        wordsIndex={}


        for siteurl in pornos:
                txt=readSite(siteurl)
                if txt==None:
                        #remove site url - unable to retrieve
                        print "Problematic site -removing-",siteurl
                        continue
```

```
                soup=BeautifulSoup(txt)

                #total tags number
                tagsnum=len(soup.findAll())



                #get big and normal_imgs
                BIGimgs=getImageNum(soup)

                normal_imgs=getAllImageNum(soup)
                if normal_imgs>0: BIGimgs_normalized=BIGimgs/(normal_imgs+0.0)
                else: BIGimgs_normalized=0
                normalized_normalimgs=normal_imgs/(tagsnum+0.0)



                #find atag_average_of_featured words in <a> tags
                featured_in_atags_sum=0
                for fw in feature_set_words:
                        featured_in_atags_sum+=getWordInAtag(soup,fw)

                #poses lexis iparxun se atags
                words_in_atags=len(cleanedAtagTokens(soup))

                #poses xaraktiristikes lexis vrethikan sta atag / poses lexis sinolika iparxun se
atags
                if words_in_atags>0:

        normalized_featured_atag_normalized=featured_in_atags_sum/words_in_atags
                else: normalized_featured_atag_normalized=0

                #sinolo lexeon istoselidas
                totalwords=len(getWebsiteWords(soup,txt))
```

```
        structIndex[siteurl]={}
        structIndex[siteurl]['BIGimgs']=BIGimgs_normalized
        structIndex[siteurl]['imgs']=normalized_normalimgs
        structIndex[siteurl]['avg_atag_words']=normalized_featured_atag_normalized
        structIndex[siteurl]['total_words']=totalwords

        #find word frequency for every featured word of this site
        wordsIndex[siteurl]={}
        for fw in feature_set_words:
                word_frequency=getWordFreq(txt,soup,fw)
                wordsIndex[siteurl][fw]={'word_freq':word_frequency}

        print totalwords,siteurl



#se poses selides aniki i kathe lexi
dw={}
for fw in feature_set_words:
        dw[fw]=0
for fw in feature_set_words:
        for site in structIndex.keys():
                if wordsIndex[site][fw]['word_freq']>0:
                        dw[fw]+=1

#sinolo selidon
n=len(structIndex.keys())

tfdf_dict={}
#generate tfdf for each word for each site
for site in structIndex.keys():
        tfdf_dict[site]={}
        for fw in feature_set_words:
                tfdf_dict[site][fw]=generate_tfdf(fw,site,wordsIndex,structIndex,dw,n)
```

```
#calculate tfdf MEAN weight
for site in structIndex.keys():
	structIndex[site]['mean_tfdf']=site_total_tfdf(site,tfdf_dict)

#print arff data
arffData=""

#generate ARFF
for site in structIndex.keys():
	instanceLine=""

	for w in sorted(feature_set_words):
		w_tfdf=tfdf_dict[site][w]
		instanceLine+=str(w_tfdf)+","

	BigImg=str(structIndex[site]['BIGimgs'])
	Img=str(structIndex[site]['imgs'])
	porn_atags_score=str(round(structIndex[site]['avg_atag_words'],3))
	avg_tfdf=str(structIndex[site]['mean_tfdf'])

	#put features in arff format
	instanceLine+=BigImg+","+Img+","+porn_atags_score+","+avg_tfdf
	instanceLine+=","+instClass    #SET CLASS VARIABLE
	instanceLine+="\n"
	arffData+=instanceLine

return arffData
```

Κώδικας 2.0: Σύνολο stopping words

I

i

a

aboard

about

above

across

after

afterwards

against

agin

ago

agreed-upon

ah

alas

albeit

all

all-over

almost

along

alongside

altho

although

amid

amidst

among

amongst

an

and

another

any

anyone

anything

around

as

aside
astride
at
atop
avec
away
back
be
because
before
beforehand
behind
behynde
below
beneath
beside
besides
between
bewteen
beyond
bi
both
but
by
ca.
de
des
despite
do
down
due
durin
during

each
eh
either
en
every
ever
everyone
everything
except
far
fer
for
from
go
goddamn
goody
gosh
half
have
he
hell
her
herself
hey
him
himself
his
ho
how
however
i
if
in

inside
insofar
instead
into
it
its
itself
la
le
les
lest
lieu
like
me
minus
moreover
my
myself
near
near-by
nearer
nearest
neither
nevertheless
next
no
nor
not
nothing
notwithstanding
o
o'er
of

off
on
once
one
oneself
only
onto
or
other
others
otherwise
our
ours
ourselves
out
outside
outta
over
per
rather
regardless
round
se
she
should
since
so
some
someone
something
than
that
the

their
them
themselves
then
there
therefore
these
they
thine
this
those
thou
though
through
throughout
thru
till
to
together
toward
towardes
towards
uh
under
underneath
unless
unlike
until
unto
up
upon
uppon
us

via

vis-a-vis

vis-&#65533;-vis

we

well

what

whatever

whatsoever

when

whenever

where

whereas

wherefore

whereupon

whether

which

whichever

while

who

whoever

whom

whose

why

with

withal

within

without

ye

yea

yeah

yes

yet

yonder

```
you
your
yours
yourself
yourselves
wants
```

Κώδικας 3.0: Αρχείο “extract_features.py” - Χρησιμοποιείται για την εξαγωγή χαρακτηριστικών σε μορφή arff

```
# -*- coding: utf-8 -*-
from BeautifulSoup import BeautifulSoup
import urllib
import urllib2
import nltk
import json
import sys

execfile('methods.py')



#feature_set_words is GLOBAL
feature_set_words=loadFile("Data/","featured_words.txt")
print "File: ","featured_words.txt", "words #", len(feature_set_words)




##                      SET TESTING MODE OR TRAINING MODE                    ##
#false for training and true for testing
if len(sys.argv)>1:
        isTest=sys.argv[1]
else:
        isTest="True"

test_set_porn=loadFile("lists/","test_set_pornos.txt")
```

```
print "File: ","test_set_pornos.txt", "urls #", len(test_set_porn)
test_set_nporn=loadFile("lists/","test_set_non_pornos.txt")
print "File: ","test_set_non_pornos.txt", "urls #", len(test_set_nporn)
pornos=loadFile("lists/","pornos_train.txt")
print "File: ","pornos_train.txt", "urls #", len(pornos)
NoNpornos=loadFile("lists/","non_pornos_train.txt")
print "File: ","non_pornos_train.txt", "urls #", len(NoNpornos)

if (isTest=="True"):
        print "Start generating Test Set"
        print
        arffDataTY=getARFFData(test_set_porn, "YES")
        arffDataTN=getARFFData(test_set_nporn, "NO")
        arffDataT=arffDataTY+arffDataTN
        filename="NEW_TEST.arff"
        test_set=test_set_porn+test_set_nporn
        findDuplicates(test_set, pornos+NoNpornos)
else:
        print "Start generating Training Set"
        print
        arffDataTY=getARFFData(pornos, "YES")
        arffDataTN=getARFFData(NoNpornos, "NO")
        arffDataT=arffDataTY+arffDataTN
        filename="NEW_TRAIN.arff"



#save to file
fo=open(filename, "w")
fo.write("@relation categorization\n")
fo.write("\n")
#set to arff attribute names
```

```
for w in sorted(feature_set_words):
        fo.write("@attribute "+w+" real"+"\n")
fo.write("@attribute BIGIMGs real"+"\n")
fo.write("@attribute NormalIMGs real"+"\n")
fo.write("@attribute TotalATAGS real"+"\n")
fo.write("@attribute totalTFIDF real"+"\n")
fo.write("@attribute isPorn {YES,NO}"+"\n")
fo.write("\n")
fo.write("@data\n")
fo.write(arffDataT)
fo.close()
```

Κώδικας 5.0: Αρχείο δεδομένων εκπαίδευσης “NEW_TRAIN.arff”

```
@relation categorization

@attribute amateur real
@attribute anal real
@attribute asian real
@attribute ass real
@attribute babe real
@attribute bisexual real
@attribute blonde real
@attribute blowjob real
@attribute boobs real
@attribute cock real
@attribute dick real
@attribute fingering real
@attribute forced real
@attribute fuck real
@attribute fucked real
@attribute fucking real
@attribute fucks real
@attribute hairy real
@attribute handjob real
@attribute hardcore real
@attribute horny real
```

@attribute huge real

@attribute japanese real

@attribute latina real

@attribute lesbian real

@attribute masturbating real

@attribute masturbation real

@attribute milf real

@attribute nude real

@attribute porno real

@attribute pornos real

@attribute pornostars real

@attribute pussy real

@attribute russian real

@attribute sexy real

@attribute slut real

@attribute sperm real

@attribute stripper real

@attribute suck real

@attribute sucking real

@attribute threesome real

@attribute tits real

@attribute tube real

@attribute xxx real

@attribute BIGIMGs real

@attribute NormalIMGs real

@attribute TotalATAGS real

@attribute totalTFIDF real

@attribute isPorn {YES,NO}

@data

0.002,0.005,0.003,0.004,0.001,0.0,0.003,0.001,0.0,0.006,0.0,0.0,0.0,0.002,0.0,0.001,0.002,0.0,0.0,0.001,0.0,0.003,0.0,0.002,0.001,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.001,0.0,0.0,0.001,0.001,0.001,0.003,0.001,0.0,0.0,0.0680713128039,0.0,0.00113636363636,YES

0.0,0.0,0.004,0.0,0.007,0.0,0.0,0.0,0.0,0.0,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.961538461538,0.0492424242424,0.0,0.000522727272727,YES

0.164,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00372727272727,YES

0.003,0.001,0.001,0.0,0.002,0.001,0.003,0.001,0.0,0.007,0.002,0.001,0.0,0.001,0.002,0.001,0.003,0.001,
0.001,0.003,0.0,0.002,0.001,0.001,0.002,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.004,0.0,0.003,0.0,0.0,0.0,0.001
,0.001,0.001,0.003,0.001,0.003,0.0,0.0464480874317,0.0,0.00134090909091,YES

0.004,0.001,0.001,0.009,0.001,0.0,0.001,0.001,0.001,0.002,0.0,0.0,0.0,0.001,0.004,0.003,0.002,0.001,0.0
01,0.002,0.002,0.001,0.001,0.001,0.002,0.0,0.001,0.005,0.0,0.002,0.0,0.0,0.008,0.001,0.003,0.0,0.0,0.0,0
.001,0.001,0.001,0.002,0.002,0.003,0.102040816327,0.054384017758,0.0,0.00163636363636,YES

0.002,0.005,0.003,0.004,0.001,0.0,0.003,0.001,0.0,0.006,0.0,0.0,0.0,0.002,0.0,0.001,0.002,0.0,0.0,0.001,
0.0,0.003,0.0,0.002,0.001,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.001,0.0,0.0,0.001,0.001,0.001,0.00
3,0.001,0.0,0.0,0.0680713128039,0.0,0.00113636363636,YES

0.0,0.005,0.003,0.0,0.0,0.001,0.005,0.001,0.001,0.004,0.003,0.0,0.0,0.003,0.0,0.001,0.0,0.001,0.0,0.0,0.0
01,0.0,0.0,0.001,0.003,0.0,0.001,0.002,0.0,0.001,0.0,0.0,0.003,0.0,0.003,0.0,0.0,0.0,0.001,0.001,0.001,0.
002,0.001,0.002,0.0,0.0353390639924,0.0,0.00115909090909,YES

0.0,0.011,0.008,0.003,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.012,0.003,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.005,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008,0.0,0.0,0.13048635
8244,0.0,0.00145454545455,YES

0.004,0.011,0.005,0.005,0.002,0.0,0.003,0.001,0.0,0.003,0.002,0.0,0.0,0.002,0.0,0.0,0.002,0.0,0.0,0.003,
0.0,0.001,0.001,0.002,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.001,0.0
,0.002,0.0,0.0812324929972,0.0,0.00122727272727,YES

0.001,0.022,0.001,0.001,0.001,0.001,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.025,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.
0,0.001,0.0,0.002,0.0,0.0,0.003,0.0,0.004,0.0,0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.023,0.01
4,0.99173553719,0.0507656807216,0.0,0.00240909090909,YES

0.001,0.004,0.002,0.002,0.001,0.0,0.001,0.001,0.0,0.003,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.
002,0.001,0.001,0.002,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.001,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.002,0
.011,0.005,0.0140845070423,0.0406294706724,0.0,0.00115909090909,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.0,0.0,0.0,0.005,0.004,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.009,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0849673202614,0.0,0.00
0704545454545,YES

0.002,0.006,0.001,0.002,0.001,0.0,0.001,0.002,0.001,0.004,0.0,0.0,0.001,0.009,0.001,0.002,0.001,0.001,
0.001,0.001,0.0,0.001,0.001,0.001,0.005,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.006,0.017,0.006,0.983606557377,0.0587479935795,0.0,0.00184090909091,YES

0.003,0.002,0.0,0.004,0.0,0.0,0.003,0.0,0.0,0.004,0.0,0.0,0.0,0.002,0.001,0.0,0.001,0.0,0.0,0.0,0.001,0.0,
0.0,0.0,0.001,0.0,0.0,0.003,0.0,0.008,0.0,0.0,0.0,0.0,0.002,0.001,0.0,0.0,0.0,0.0,0.001,0.002,0.0,0.001,0.0
,0.0663983903421,0.0,0.000909090909091,YES

0.007,0.014,0.001,0.009,0.0,0.001,0.001,0.002,0.0,0.005,0.0,0.001,0.0,0.001,0.0,0.001,0.0,0.001,0.001,0.
001,0.0,0.0,0.001,0.001,0.006,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.005,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.
007,0.003,0.0,0.0,0.0530386740331,0.0,0.00165909090909,YES

0.003,0.001,0.001,0.0,0.001,0.001,0.004,0.002,0.0,0.006,0.002,0.001,0.0,0.001,0.002,0.001,0.002,0.001,
0.001,0.003,0.0,0.002,0.001,0.001,0.002,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.003,0.0,0.003,0.0,0.0,0.0,0.001
,0.001,0.001,0.004,0.001,0.003,0.0,0.0467032967033,0.0,0.00131818181818,YES

0.0,0.004,0.003,0.001,0.0,0.0,0.004,0.0,0.0,0.001,0.001,0.0,0.039,0.009,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,
0.003,0.0,0.001,0.0,0.0,0.002,0.0,0.002,0.0,0.0,0.002,0.006,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.02,0.0,0.0,
0.100446428571,0.0,0.00229545454545,YES

0.002,0.0,0.0,0.0,0.0,0.0,0.007,0.002,0.0,0.0,0.0,0.0,0.0,0.014,0.01,0.005,0.005,0.0,0.0,0.0,0.003,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.983606557377,0.08
76436781609,0.0,0.00127272727273,YES

0.002,0.015,0.002,0.001,0.001,0.0,0.002,0.002,0.001,0.002,0.0,0.0,0.013,0.007,0.004,0.002,0.001,0.001,
0.0,0.0,0.001,0.0,0.002,0.0,0.003,0.0,0.0,0.002,0.001,0.0,0.0,0.0,0.004,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0

,0.0,0.03,0.001,1.0,0.167539267016,0.0,0.00231818181818,YES

0.002,0.02,0.002,0.001,0.001,0.0,0.003,0.002,0.0,0.001,0.0,0.0,0.02,0.014,0.003,0.002,0.001,0.0,0.0,0.00
1,0.0,0.0,0.001,0.0,0.003,0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0
13,0.001,1.0,0.171568627451,0.0,0.00222727272727,YES

0.001,0.004,0.001,0.001,0.0,0.0,0.0,0.001,0.001,0.003,0.0,0.0,0.001,0.006,0.001,0.001,0.0,0.001,0.0,0.0,
0.0,0.001,0.0,0.0,0.002,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.004,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.066,0.
006,1.0,0.0512091038407,0.0,0.00247727272727,YES

0.002,0.008,0.002,0.002,0.001,0.0,0.001,0.002,0.0,0.006,0.0,0.0,0.001,0.007,0.002,0.002,0.0,0.001,0.001
,0.001,0.0,0.001,0.001,0.001,0.004,0.0,0.0,0.002,0.0,0.002,0.0,0.0,0.004,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0
.0,0.006,0.075,0.005,0.994475138122,0.0504178272981,0.0,0.00322727272727,YES

0.003,0.007,0.001,0.002,0.001,0.0,0.001,0.001,0.001,0.003,0.0,0.0,0.0,0.004,0.001,0.002,0.0,0.001,0.001
,0.001,0.0,0.001,0.001,0.0,0.003,0.001,0.0,0.002,0.0,0.002,0.0,0.0,0.003,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.005,0.014,0.004,1.0,0.0549282880684,0.0,0.00152272727273,YES

0.001,0.003,0.003,0.002,0.001,0.0,0.001,0.0,0.0,0.002,0.0,0.0,0.01,0.006,0.003,0.002,0.004,0.001,0.0,0.0
01,0.001,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.003,0.0,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0
.0,0.992481203008,0.149943630214,0.0,0.00115909090909,YES

0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
004,0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0587275693312,0.0,0.000318
181818182,YES

0.0,0.004,0.005,0.0,0.0,0.0,0.006,0.0,0.0,0.002,0.001,0.0,0.035,0.009,0.019,0.001,0.0,0.0,0.0,0.0,0.003,0.
001,0.004,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.001,0.004,0.001,0.001,0.0,0.0,0.002,0.0,0.0,0.0,0.01,0
.0,0.0,0.266666666667,0.0,0.00252272727273,YES

0.001,0.006,0.003,0.005,0.001,0.0,0.003,0.001,0.0,0.001,0.0,0.0,0.013,0.01,0.006,0.004,0.003,0.0,0.0,0.0
01,0.0,0.0,0.004,0.0,0.001,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
992481203008,0.149943630214,0.0,0.00156818181818,YES

0.003,0.005,0.002,0.006,0.001,0.0,0.004,0.005,0.0,0.003,0.001,0.0,0.001,0.002,0.0,0.001,0.001,0.001,0.0
01,0.007,0.0,0.001,0.001,0.001,0.005,0.001,0.0,0.002,0.001,0.001,0.0,0.0,0.006,0.001,0.002,0.0,0.0,0.0,0
.0,0.003,0.001,0.008,0.017,0.002,0.0,0.0849952516619,0.0,0.00220454545455,YES

0.003,0.011,0.003,0.002,0.001,0.0,0.001,0.001,0.001,0.005,0.0,0.0,0.001,0.014,0.001,0.001,0.0,0.002,0.0
01,0.0,0.0,0.001,0.003,0.001,0.004,0.0,0.0,0.003,0.0,0.002,0.0,0.0,0.002,0.001,0.001,0.0,0.0,0.0,0.003,0.
001,0.0,0.006,0.028,0.011,0.989010989011,0.0829157175399,0.0,0.00261363636364,YES

0.0,0.002,0.001,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.009,0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.017,0.002,0.0,0.069738480697
4,0.0,0.000954545454545,YES

0.0,0.0,0.005,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.019,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.026,0.0,1.0,0.168776371308,0.0,0.00131
818181818,YES

0.001,0.004,0.0,0.004,0.002,0.0,0.008,0.001,0.003,0.005,0.001,0.0,0.002,0.003,0.003,0.002,0.0,0.0,0.0,0.
001,0.002,0.001,0.0,0.001,0.001,0.0,0.0,0.001,0.001,0.007,0.0,0.0,0.003,0.0,0.002,0.002,0.0,0.0,0.0,0.0,0
.0,0.002,0.007,0.001,0.0,0.0671641791045,0.0,0.00161363636364,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.157894736842,0.0673758865248,0.0,0.0,Y
ES

0.005,0.0,0.0,0.0,0.009,0.0,0.0,0.0,0.0,0.011,0.0,0.0,0.0,0.013,0.005,0.0,0.0,0.0,0.006,0.0,0.0,0.0,0.0,0.0,
0.016,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006,0.0,0.0,0.046511627907,0.
0,0.00165909090909,YES

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.015,0.006,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.046,0.007,0.0,0.0883720930233,0.0,
0.00177272727273,YES

0.007,0.018,0.004,0.001,0.0,0.0,0.0,0.002,0.0,0.005,0.001,0.0,0.0,0.002,0.001,0.001,0.001,0.001,0.001,0.
0,0.0,0.001,0.005,0.0,0.011,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.005,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.005,0
.048,0.0,0.96,0.0259929299231,0.0,0.00288636363636,YES

0.001,0.006,0.003,0.013,0.001,0.0,0.002,0.001,0.0,0.001,0.0,0.001,0.0,0.001,0.002,0.002,0.001,0.001,0.0
01,0.001,0.001,0.0,0.001,0.001,0.001,0.0,0.001,0.001,0.0,0.002,0.0,0.0,0.011,0.001,0.002,0.0,0.0,0.0,0.0
01,0.0,0.001,0.003,0.001,0.001,0.0,0.0342706502636,0.0,0.00152272727273,YES

0.004,0.004,0.001,0.004,0.0,0.0,0.002,0.0,0.0,0.002,0.0,0.0,0.0,0.001,0.001,0.002,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.002,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.056029
2326431,0.0,0.000590909090909,YES

0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.001,0.001,0.003,0.001,0.0,0.0,0.002,0.002,0.0,0.001,0.001,0.0,0.001,0.0,0.
002,0.0,0.0,0.001,0.0,0.0,0.0,0.007,0.0,0.0,0.0,0.006,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.006,0.013,0.001,0
.0179640718563,0.20875,0.0,0.00115909090909,YES

0.0,0.012,0.008,0.003,0.001,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.022,0.009,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0
05,0.0,0.002,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.001,0.012,0.0,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.016,0.0,0.0,
0.176,0.0,0.00229545454545,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007,0.625,0.0776699029126,0.0,0.0001590909
09091,YES

0.0,0.009,0.0,0.0,0.007,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.005,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0
04,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.961538461538,0.04966
57115568,0.0,0.000886363636364,YES

0.002,0.008,0.005,0.002,0.001,0.0,0.001,0.001,0.0,0.001,0.001,0.0,0.01,0.012,0.004,0.002,0.0,0.0,0.001,
0.001,0.0,0.001,0.004,0.0,0.001,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.002,0.002,0.002,0.001,0.0,0.0,0.001,0.0,0.
001,0.0,0.001,0.001,0.0,0.129220508545,0.0,0.00163636363636,YES

0.0,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.75,0.0466101694915,0.0,0.000136363
636364,YES

0.004,0.004,0.004,0.0,0.008,0.0,0.009,0.002,0.001,0.006,0.001,0.0,0.0,0.002,0.005,0.0,0.001,0.002,0.001
,0.003,0.002,0.002,0.001,0.002,0.003,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.001,0.003,0.0,0.0,0.0,0.0,0.00
1,0.0,0.004,0.003,0.001,0.127208480565,0.258624628741,0.0,0.00177272727273,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.0,0.003,0.0,0.0,0.001,0.0
02,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.95652173913,0.065714
2857143,0.0,0.000295454545455,YES

0.001,0.011,0.009,0.002,0.001,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.014,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,
0.0,0.0,0.013,0.0,0.0,0.001,0.0,0.007,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.052,0.043,0.01860
46511628,0.131498470948,0.0,0.00361363636364,YES

0.03,0.001,0.001,0.001,0.0,0.0,0.004,0.001,0.0,0.003,0.001,0.0,0.0,0.006,0.003,0.006,0.001,0.001,0.0,0.0
,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.002,0.002,0.001,0.0,0.0,0.003,0.001,0.007,0.002,0.0,0.0,0.0,0.002,0.0,0
.002,0.007,0.001,0.491803278689,0.067217630854,0.0,0.00206818181818,YES

0.004,0.0,0.008,0.003,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.021,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.005,0.0,0.0,0.001,0.001,0.009,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.064,0.04,0.0186046511
628,0.131659522352,0.0,0.00370454545455,YES

0.0,0.003,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.033,0.007,0.0,0.103082851638,0.0,0

.00145454545455,YES

0.0,0.008,0.005,0.001,0.0,0.0,0.006,0.0,0.0,0.001,0.0,0.0,0.012,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
05,0.0,0.0,0.0,0.0,0.005,0.0,0.0,0.0,0.0,0.0,0.005,0.003,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.018,0.001,0.0,0.11
5667718191,0.0,0.00165909090909,YES

0.002,0.006,0.001,0.003,0.001,0.001,0.001,0.001,0.001,0.004,0.001,0.0,0.001,0.011,0.001,0.003,0.001,0.
001,0.0,0.001,0.0,0.002,0.001,0.001,0.004,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.005,0.001,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.006,0.018,0.008,0.983606557377,0.0608175473579,0.0,0.00204545454545,YES

0.0,0.007,0.0,0.0,0.009,0.0,0.0,0.005,0.0,0.012,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0
16,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.006,0.0,0.0,0.0466288594833,0.0,
0.0015,YES

0.002,0.009,0.003,0.002,0.001,0.0,0.006,0.003,0.0,0.003,0.002,0.0,0.0,0.003,0.001,0.004,0.001,0.001,0.0
01,0.002,0.0,0.002,0.0,0.002,0.002,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.0,0.001,0.002,0.0,0.0,0.0,0.0,0.0,
0.002,0.006,0.001,0.0,0.0293847566575,0.0,0.00147727272727,YES

0.005,0.001,0.001,0.008,0.001,0.0,0.002,0.001,0.001,0.003,0.0,0.0,0.0,0.001,0.005,0.003,0.002,0.001,0.0
01,0.003,0.001,0.001,0.001,0.001,0.002,0.0,0.0,0.004,0.0,0.002,0.0,0.0,0.008,0.001,0.002,0.001,0.0,0.0,0
.001,0.001,0.001,0.002,0.002,0.003,0.0980392156863,0.0564784053156,0.0,0.00165909090909,YES

0.002,0.011,0.008,0.007,0.001,0.0,0.002,0.003,0.001,0.004,0.003,0.0,0.0,0.014,0.002,0.003,0.001,0.002,
0.001,0.002,0.0,0.001,0.004,0.0,0.004,0.001,0.0,0.002,0.001,0.001,0.0,0.0,0.006,0.0,0.001,0.001,0.0,0.0,
0.001,0.001,0.0,0.004,0.007,0.005,0.0186046511628,0.0466985230235,0.0,0.00243181818182,YES

0.009,0.012,0.005,0.0,0.0,0.0,0.0,0.0,0.003,0.011,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.004,0.0,0.0,0.004
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.0,0.0,0.0283018867925,0
.0,0.00129545454545,YES

0.001,0.004,0.0,0.0,0.001,0.001,0.001,0.001,0.001,0.001,0.002,0.0,0.0,0.004,0.003,0.0,0.002,0.0,0.0,0.0,
0.001,0.002,0.0,0.0,0.005,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.003,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.001,0.003,
0.004,0.005,0.152542372881,0.0471622701839,0.0,0.00113636363636,YES

0.006,0.015,0.003,0.002,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.003,0.007,0.004,0.001,0.002,0.001,0.001,0.
0,0.001,0.0,0.001,0.0,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0
06,0.0,1.0,0.0802259887006,0.0,0.00143181818182,YES

0.002,0.012,0.005,0.003,0.001,0.0,0.004,0.001,0.0,0.003,0.001,0.0,0.012,0.007,0.007,0.001,0.001,0.0,0.0
,0.0,0.001,0.001,0.003,0.001,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.001,0.002,0.003,0.001,0.0,0.0,0.003,0.
0,0.001,0.001,0.001,0.0,0.0,0.114138438881,0.0,0.00184090909091,YES

0.0,0.01,0.006,0.006,0.002,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.014,0.005,0.004,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.002,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.032,0.0,0.04,0.1237
62376238,0.0,0.002,YES

0.008,0.015,0.003,0.001,0.0,0.0,0.0,0.001,0.0,0.004,0.001,0.0,0.0,0.001,0.0,0.001,0.001,0.002,0.001,0.0,
0.0,0.001,0.004,0.0,0.01,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.004,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.006,0.01
3,0.001,0.975609756098,0.0403675746636,0.0,0.00193181818182,YES

0.003,0.007,0.003,0.002,0.001,0.0,0.001,0.001,0.001,0.006,0.0,0.001,0.001,0.008,0.001,0.002,0.0,0.001,
0.001,0.001,0.0,0.001,0.001,0.001,0.005,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.004,0.001,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.006,0.016,0.005,0.989010989011,0.0607679465776,0.0,0.00190909090909,YES

0.002,0.008,0.003,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007,0.0,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.001
,0.0,0.001,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.002,0.001,0.001,0.001,0.0,0.0,0.0,0.0,0.001,0.001,0.052,0.00
7,0.0,0.0373111316682,0.0,0.00220454545455,YES

0.001,0.005,0.002,0.003,0.0,0.0,0.0,0.002,0.001,0.002,0.0,0.0,0.001,0.01,0.007,0.004,0.004,0.001,0.001,
0.001,0.002,0.0,0.001,0.0,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.0,0.001,0.001,0.0,0.0,0.0,0.001,0.00
1,0.0,0.016,0.001,1.0,0.175901495163,0.0,0.00165909090909,YES

0.0,0.007,0.003,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.013,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.038,0.006,0.0,0.0879396984925,
0.0,0.00168181818182,YES

0.002,0.004,0.0,0.0,0.001,0.001,0.002,0.001,0.001,0.001,0.002,0.0,0.0,0.004,0.002,0.001,0.001,0.0,0.0,0.
0,0.001,0.002,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.003,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.001,0.003,0
.004,0.005,0.152542372881,0.0470119521912,0.0,0.00111363636364,YES

0.002,0.001,0.0,0.003,0.0,0.0,0.013,0.0,0.0,0.004,0.003,0.0,0.017,0.015,0.004,0.0,0.0,0.001,0.0,0.0,0.0,0.
0,0.0,0.0,0.002,0.0,0.0,0.002,0.001,0.0,0.0,0.0,0.006,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.0,0.0,0.0,0.
0768518518519,0.0,0.00181818181818,YES

0.002,0.004,0.0,0.002,0.003,0.0,0.0,0.001,0.0,0.0,0.0,0.001,0.0,0.012,0.001,0.004,0.0,0.0,0.0,0.004,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.002,0.0,0.006,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0225
225225225,0.0,0.001,YES

0.0,0.003,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.015,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.002,0.0,0.004,0.0,0.110817941953,0.
0,0.000818181818182,YES

0.001,0.005,0.001,0.006,0.001,0.001,0.001,0.002,0.0,0.002,0.0,0.0,0.001,0.006,0.0,0.002,0.0,0.001,0.001
,0.001,0.0,0.0,0.001,0.001,0.002,0.0,0.0,0.001,0.0,0.002,0.0,0.0,0.007,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.001,
0.005,0.017,0.001,0.00990099009901,0.0797788309637,0.0,0.00161363636364,YES

0.003,0.013,0.004,0.005,0.002,0.0,0.001,0.001,0.001,0.005,0.0,0.001,0.0,0.004,0.001,0.001,0.0,0.003,0.0
01,0.001,0.0,0.0,0.002,0.0,0.007,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.007,0.002,0.001,0.0,0.0,0.0,0.0,0.0,0.002,
0.006,0.003,0.002,0.0,0.0114285714286,0.0,0.00186363636364,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0703125,0.0,0.0,YES

0.0,0.004,0.0,0.002,0.003,0.0,0.006,0.001,0.0,0.003,0.002,0.0,0.035,0.009,0.019,0.001,0.0,0.0,0.0,0.0,0.0
02,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.003,0.001,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.002,0.
0,0.0,0.268817204301,0.0,0.00227272727273,YES

0.0,0.009,0.007,0.009,0.0,0.002,0.0,0.0,0.0,0.0,0.002,0.0,0.008,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0
03,0.0,0.003,0.0,0.0,0.006,0.0,0.005,0.0,0.0,0.003,0.002,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.001,0.018,0.0,1
.0,0.153225806452,0.0,0.00197727272727,YES

0.008,0.001,0.002,0.0,0.003,0.0,0.001,0.001,0.0,0.006,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.007,0.0,0.001,0.001
,0.002,0.001,0.002,0.001,0.0,0.0,0.0,0.009,0.001,0.0,0.0,0.01,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.001,0.002,0.
004,0.0,0.822222222222,0.0507900677201,0.0,0.00156818181818,YES

0.003,0.008,0.001,0.003,0.001,0.0,0.001,0.001,0.002,0.004,0.0,0.0,0.001,0.011,0.0,0.002,0.001,0.002,0.0
01,0.0,0.0,0.0,0.001,0.001,0.004,0.0,0.0,0.002,0.001,0.003,0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.003,0.123,0.009,0.994475138122,0.0921588594705,0.0,0.00434090909091,YES

0.006,0.004,0.0,0.021,0.0,0.0,0.006,0.015,0.002,0.013,0.006,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.005,0.008,0.0
,0.002,0.0,0.008,0.0,0.0,0.0,0.006,0.0,0.0,0.0,0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.025,0.003,0.0,0
.0,0.0342465753425,0.0,0.00320454545455,YES

0.011,0.019,0.005,0.005,0.0,0.0,0.003,0.012,0.0,0.003,0.001,0.002,0.0,0.0,0.003,0.002,0.001,0.001,0.001
,0.011,0.0,0.0,0.0,0.001,0.002,0.0,0.001,0.009,0.0,0.0,0.0,0.0,0.0,0.001,0.003,0.001,0.0,0.0,0.0,0.0,0.0,0.
018,0.004,0.0,0.0,0.037711313394,0.0,0.00272727272727,YES

0.0,0.007,0.008,0.0,0.01,0.0,0.007,0.0,0.0,0.002,0.0,0.0,0.029,0.038,0.01,0.0,0.0,0.0,0.0,0.002,0.001,0.0,
0.003,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006,0.003,0.003,0.004,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.
115234375,0.0,0.00306818181818,YES

0.002,0.003,0.001,0.004,0.004,0.0,0.005,0.003,0.001,0.008,0.001,0.0,0.0,0.003,0.007,0.002,0.0,0.001,0.0
,0.005,0.002,0.0,0.001,0.001,0.002,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.007,0.0,0.003,0.002,0.0,0.0,0.0,0.0,0.0
01,0.005,0.001,0.0,0.112359550562,0.112090680101,0.0,0.00172727272727,YES

0.003,0.007,0.002,0.002,0.001,0.0,0.001,0.001,0.002,0.006,0.0,0.0,0.001,0.011,0.001,0.002,0.0,0.001,0.0
01,0.001,0.0,0.001,0.001,0.001,0.004,0.0,0.0,0.003,0.0,0.001,0.0,0.0,0.005,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.008,0.026,0.007,1.0,0.0661035622475,0.0,0.00229545454545,YES

0.0,0.004,0.001,0.002,0.001,0.0,0.004,0.0,0.0,0.001,0.0,0.0,0.002,0.001,0.001,0.0,0.0,0.0,0.0,0.001,0.001
,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.002,0.003,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
420168067227,0.278455624269,0.0,0.000636363636364,YES

0.0,0.0,0.002,0.002,0.0,0.0,0.0,0.003,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.00
1,0.004,0.0,0.001,0.003,0.003,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.001,0.0,0.001,0.002,0.0,0.002,0.002,0.006,0
.913043478261,0.21198156682,0.0,0.000977272727273,YES

0.0,0.01,0.004,0.0,0.002,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.018,0.015,0.002,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.00
3,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.002,0.003,0.002,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.005,0.0,0.0,0.116
831683168,0.0,0.00175,YES

0.003,0.007,0.006,0.004,0.001,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.009,0.001,0.001,0.0,0.001,0.0,0.002,
0.0,0.001,0.001,0.0,0.002,0.001,0.0,0.003,0.001,0.004,0.0,0.0,0.003,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.00
1,0.032,0.012,0.997297297297,0.102663706992,0.0,0.00225,YES

0.0,0.005,0.003,0.0,0.0,0.001,0.005,0.001,0.001,0.004,0.003,0.0,0.0,0.003,0.0,0.001,0.0,0.001,0.0,0.0,0.0
01,0.0,0.0,0.001,0.003,0.0,0.001,0.002,0.0,0.001,0.0,0.0,0.003,0.0,0.003,0.0,0.0,0.0,0.001,0.001,0.001,0.
002,0.001,0.002,0.0,0.0353390639924,0.0,0.00115909090909,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.137931034483,0.0483333333333,0.0,0.0,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.285714285714,0.00614574187884,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.516129032258,0.00626262626263,0.0,0.
000272727272727,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00149476831091,0.0,6.8181818
1818e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0,0.0,0.0,2.27272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00911161731207,0.0,2.27272727273
e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.476744186047,0.0845624385447,0.0,0.0,N
O

0.0,0.003,0.0,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.001
,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.00022727272727
3,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0249396621078,0.0,0.0,NO

0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00428082191781,0.0,2.27272727273
e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0208078335373,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00735294117647,0.0,0.00011363636
3636,NO

0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0191972076789,0.
0,6.81818181818e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.030303030303,0.0,6.81818181818e-
05,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.55,0.0455580865604,0.0,2.27272727273
e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0981012658228,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.444444444444,0.00408905043162,0.0,0.0,
NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.166666666667,0.00562324273664,0.0,
4.54545454545e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0195035460993,0.0,0.0,N
O

0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.6875,0.0325203252033,0.0,9.0909090
9091e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.05,0.0545702592087,0.0,0.0,NO

0.0,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.833333333333,0.0821917808219,0.
0,0.000181818181818,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25,0.0227272727273,0.0,4.545454545
45e-05,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.181818181818,0.0203703703704,0.0,4
.54545454545e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0204460966543,0.0,2.2
7272727273e-05,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.603773584906,0.0558482613277,0.0,2.2
7272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00808625336927,0.0,2.27272727273
e-05,NO

0.0,0.001,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0
01,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.0,0.000204545454
545,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00153139356815,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0170357751278,0.0,4.54545454545e-
05,NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.002,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0
01,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.0,0.000295454545
455,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.333333333333,0.00975081256771,0.0,2.
27272727273e-05,NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.001,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.0,0.000227272727
3,NO

0.0,0.004,0.0,0.001,0.0,0.001,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0
01,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0,0.0,0.0,0.000272727272
727,NO

0.0,0.017,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.619047619048,0.00380296993843,0.0,
0.000409090909091,NO

0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.363636363636,0.119565217391,0.0,6.
81818181818e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.005
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0909090909091,0.0196428571429,0.0,
0.000136363636364,NO

0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0754593175853,0.0,9.09090909
091e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00185873605948,0.0,4.545454545
45e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0209523809524,0.0,0.0,N
O

0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.2,0.0106044538706,0.0,9.0909090909
1e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0201465201465,0.0,2.2
7272727273e-05,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0213178294574,0.0,4
.54545454545e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0197841726619,0.0,2.2
7272727273e-05,NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,
0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000113636363636,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00320512820513,0.0,2.27272727273
e-05,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0
01,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.363636363636,0.00450819672131,0
.0,6.81818181818e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.606060606061,0.00997883277895,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0209923664122,0.0,0.0,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00291120815138,0.0,4.54545454545
e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.2,0.0119047619048,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.578313253012,0.0706382978723,0.0,2.2
7272727273e-05,NO

0.0,0.001,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,
0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000159090909091,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0197486535009,0.
0,6.81818181818e-05,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.1,0.0374531835206,0.0,4.5454545454
5e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.642857142857,0.00454988625284,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.011,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01875,0.0,0.00025,NO

0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.1,0.0111482720178,0.0,4.54545454545e-
05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.142857142857,0.00273224043716,0.0,0.0,
NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0201465201465,0.0,4
.54545454545e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.00847457627119,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.273195876289,0.0533260032985,0.0,0.0,N
O

0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.142857142857,0.00932090545939,0.0,
0.000159090909091,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0846153846154,0.0778443113772,0.0,0.0,
NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
01,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000227272727
273,NO

0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0254175744372,0.0,0.000136363636
364,NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,
0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000113636363636,N
O

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,
0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000113636363636,N
O

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,
0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000113636363636,N
O

0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0210325047801,0.
0,9.09090909091e-05,NO

0.0,0.002,0.0,0.002,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.00020454545454
5,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0148867313916,0.0,2.27272727273e-
05,NO

0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.181818181818,0.0213178294574,0.0,2.2
7272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25,0.021897810219,0.0,0.0,NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,
0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0,0.0,0.0,0.000159090909091,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.880597014925,0.0413325107958,0.0,9.0
9090909091e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.049504950495,0.0,0.0,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.666666666667,0.0306539509537,0.0,6
.81818181818e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.009,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.916666666667,0.0112781954887,0.0,0.0
00204545454545,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.00515463917526,0.0,9.090909090
91e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.916666666667,0.0111731843575,0.0,0.0,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.476744186047,0.0805998125586,0.0,9.0
9090909091e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.142857142857,0.0503597122302,0.0,0.0,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0623052959502,0.0,0.0001136363
63636,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.333333333333,0.00603015075377,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0304182509506,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.133603238866,0.113198900092,0.0,0.0,NO

0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001
,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0,0.0,0.0,0.000204545454
545,NO

Κώδικας 5.1: Αρχείο δεδομένων αξιολόγησης “NEW_TEST.arff”

```
@relation categorization

@attribute amateur real
@attribute anal real
@attribute asian real
@attribute ass real
@attribute babe real
@attribute bisexual real
@attribute blonde real
@attribute blowjob real
@attribute boobs real
@attribute cock real
@attribute dick real
@attribute fingering real
@attribute forced real
@attribute fuck real
@attribute fucked real
@attribute fucking real
@attribute fucks real
@attribute hairy real
@attribute handjob real
@attribute hardcore real
@attribute horny real
@attribute huge real
@attribute japanese real
@attribute latina real
@attribute lesbian real
@attribute masturbating real
@attribute masturbation real
@attribute milf real
@attribute nude real
@attribute porno real
@attribute pornos real
@attribute pornostars real
@attribute pussy real
```

@attribute russian real

@attribute sexy real

@attribute slut real

@attribute sperm real

@attribute stripper real

@attribute suck real

@attribute sucking real

@attribute threesome real

@attribute tits real

@attribute tube real

@attribute xxx real

@attribute BIGIMGs real

@attribute NormalIMGs real

@attribute TotalATAGS real

@attribute totalTFIDF real

@attribute isPorn {YES,NO}

@data

0.0,0.002,0.0,0.004,0.002,0.0,0.001,0.001,0.0,0.002,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.00
1,0.0,0.001,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.001,0.002,0
.0,0.064929693962,0.0,0.000590909090909,YES

0.012,0.002,0.001,0.001,0.0,0.0,0.002,0.018,0.0,0.006,0.001,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.001
,0.002,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.002,0.0,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.005,
0.001,0.0,0.128617363344,0.0,0.00145454545455,YES

0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.002,0.0,0.0,0.0,0.01,0.002,0.004,0.002,0.001,0.001,0.0,0.002,
0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.004,0.003,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007,0.001,0.
943396226415,0.113978494624,0.0,0.00113636363636,YES

0.002,0.002,0.0,0.001,0.005,0.0,0.001,0.001,0.001,0.002,0.001,0.0,0.0,0.0,0.002,0.0,0.001,0.001,0.0,0.0,0.0,0.00
1,0.001,0.0,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0
01,0.0,0.979591836735,0.0941101152369,0.0,0.00075,YES

0.0,0.012,0.0,0.004,0.0,0.0,0.0,0.004,0.002,0.0,0.0,0.007,0.0,0.0,0.0,0.0,0.004,0.005,0.008,0.0,0.0,0.0,0.0,0.002,0.002,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.001,0.0,0.0,0.0,0.002,0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.002,0.002,0
.0,0.117001828154,0.0,0.00143181818182,YES

0.006,0.006,0.003,0.001,0.015,0.0,0.011,0.0,0.0,0.005,0.004,0.0,0.0,0.0,0.0,0.011,0.004,0.002,0.001,0.001,0.
002,0.003,0.002,0.001,0.004,0.002,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.006,0.001,0.009,0.0,0.0,0.0,0.0,0.0,0.002,0
.001,0.003,0.001,0.002,0.612244897959,0.0483711747285,0.0,0.00254545454545,YES

0.002,0.007,0.002,0.001,0.001,0.0,0.001,0.001,0.001,0.004,0.001,0.0,0.0,0.004,0.001,0.003,0.001,0.001,
0.001,0.001,0.0,0.001,0.001,0.001,0.004,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.004,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.004,0.013,0.008,0.994565217391,0.0563898253141,0.0,0.00163636363636,YES

0.001,0.003,0.005,0.005,0.0,0.001,0.0,0.0,0.0,0.008,0.003,0.0,0.002,0.003,0.003,0.004,0.001,0.005,0.0,0.
0,0.001,0.002,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.001,0.
001,0.0,0.0479098168154,0.0,0.00118181818182,YES

0.001,0.003,0.0,0.003,0.001,0.0,0.001,0.0,0.002,0.002,0.002,0.0,0.0,0.002,0.002,0.003,0.0,0.0,0.001,0.00
1,0.001,0.002,0.0,0.001,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.004,0
.0,0.002,0.813333333333,0.0632377740304,0.0,0.000886363636364,YES

0.002,0.016,0.003,0.005,0.001,0.0,0.001,0.001,0.001,0.005,0.001,0.0,0.0,0.006,0.001,0.002,0.0,0.001,0.0
01,0.001,0.0,0.001,0.001,0.0,0.004,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.
005,0.012,0.007,1.0,0.0581959262852,0.0,0.00190909090909,YES

0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.002,0.001,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.139784946237,0.0,0.00018
1818181818,YES

0.003,0.003,0.003,0.002,0.0,0.0,0.005,0.003,0.0,0.0,0.001,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.003,0.0,0.0,
0.0,0.004,0.003,0.0,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.001,0.002,0.001,0.0,0.0,0.0,0.0,0.001,0.002,0.002,0.
005,0.453488372093,0.117808219178,0.0,0.00113636363636,YES

0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.004,0.0,0.004,0.0,0.0,0.0,0.011,0.002,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.002,0.004,0.0,0.0,0.003,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.008,0.012,0.320924
261874,0.0,0.00115909090909,YES

0.003,0.007,0.003,0.003,0.001,0.0,0.001,0.001,0.001,0.004,0.0,0.0,0.0,0.001,0.002,0.002,0.0,0.002,0.002
,0.001,0.0,0.002,0.003,0.002,0.005,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.004,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
009,0.002,0.001,0.992,0.0616067028093,0.0,0.00147727272727,YES

0.002,0.009,0.003,0.004,0.001,0.0,0.001,0.0,0.0,0.003,0.0,0.0,0.0,0.003,0.0,0.001,0.0,0.0,0.0,0.003,0.0,0.
0,0.002,0.0,0.003,0.0,0.001,0.003,0.0,0.002,0.0,0.0,0.003,0.001,0.002,0.0,0.001,0.0,0.0,0.0,0.0,0.005,0.0
11,0.006,0.983606557377,0.072705601907,0.0,0.00159090909091,YES

0.009,0.018,0.005,0.001,0.0,0.0,0.0,0.002,0.0,0.004,0.001,0.0,0.0,0.001,0.001,0.001,0.001,0.002,0.003,0.
0,0.0,0.001,0.006,0.001,0.013,0.0,0.001,0.004,0.0,0.0,0.0,0.0,0.005,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.
007,0.002,0.002,0.979591836735,0.0409288339459,0.0,0.00218181818182,YES

0.0,0.004,0.002,0.0,0.001,0.0,0.002,0.002,0.0,0.0,0.001,0.0,0.0,0.004,0.0,0.001,0.002,0.0,0.0,0.0,0.0,0.00
1,0.001,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.002,0.0,0.003,0.9
89130434783,0.116015132409,0.0,0.000795454545455,YES

0.003,0.014,0.003,0.005,0.0,0.0,0.0,0.0,0.0,0.007,0.0,0.001,0.0,0.002,0.001,0.002,0.0,0.003,0.003,0.002,
0.0,0.001,0.004,0.003,0.004,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.003,0.002,0.0,0.0,0.0,0.0,0.001,0.0,0.002,0.00
9,0.002,0.001,0.0263157894737,0.0777891504606,0.0,0.00184090909091,YES

0.001,0.003,0.001,0.001,0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.003,0.0,0.0,0.0
,0.001,0.001,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.003,0.0,0.001,0.95270
2702703,0.149949341439,0.0,0.000477272727273,YES

0.0,0.0,0.0,0.004,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.003,0.0,0.0,0.0,0.0,0.003,0.002,0.0,0.001,0.
0,0.0,0.0,0.001,0.01,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.125,0.0,0.00068
1818181818,YES

0.004,0.013,0.006,0.007,0.004,0.0,0.001,0.003,0.001,0.005,0.0,0.001,0.0,0.001,0.006,0.002,0.001,0.001,
0.001,0.004,0.003,0.0,0.003,0.002,0.002,0.0,0.001,0.003,0.0,0.001,0.0,0.0,0.009,0.001,0.003,0.0,0.0,0.0,
0.0,0.0,0.001,0.005,0.001,0.0,0.0,0.0614203454894,0.0,0.00218181818182,YES

0.002,0.008,0.002,0.002,0.001,0.0,0.002,0.001,0.001,0.003,0.0,0.0,0.0,0.01,0.001,0.003,0.001,0.002,0.0,
0.0,0.0,0.002,0.001,0.0,0.004,0.0,0.0,0.002,0.001,0.001,0.0,0.0,0.006,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,
0.007,0.023,0.017,0.989010989011,0.0619679945523,0.0,0.00238636363636,YES

0.006,0.003,0.007,0.005,0.001,0.0,0.008,0.001,0.002,0.001,0.0,0.0,0.0,0.009,0.013,0.004,0.005,0.003,0.0
01,0.001,0.002,0.002,0.006,0.003,0.0,0.0,0.0,0.013,0.0,0.0,0.0,0.0,0.0,0.004,0.007,0.002,0.0,0.0,0.0,0.0,0
.002,0.001,0.015,0.008,0.988636363636,0.0845341018252,0.0,0.00306818181818,YES

0.003,0.002,0.001,0.001,0.001,0.0,0.001,0.001,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.0,0.001,0.001,0.00
3,0.001,0.0,0.001,0.0,0.002,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.001,0.001,0.001,0.002,0.001,0.0,0.0,0.0,0.001,

0.002,0.004,0.013,0.0,0.0619559651818,0.0,0.00118181818182,YES

0.007,0.016,0.004,0.002,0.001,0.0,0.0,0.002,0.0,0.004,0.001,0.0,0.0,0.002,0.001,0.001,0.001,0.002,0.002,0.0,0.0,0.001,0.005,0.001,0.012,0.0,0.001,0.004,0.0,0.0,0.0,0.0,0.0,0.005,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.006,0.048,0.0,0.96,0.0390381011868,0.0,0.003,YES

0.0,0.003,0.003,0.007,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.008,0.0,0.0,0.007,0.0,0.01,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0652173913043,0.0,0.00122727272727,YES

0.019,0.012,0.003,0.002,0.001,0.0,0.003,0.001,0.0,0.002,0.002,0.0,0.0,0.007,0.002,0.005,0.001,0.0,0.001,0.002,0.001,0.0,0.0,0.001,0.001,0.0,0.001,0.003,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.005,0.002,0.0,0.0,0.0,0.0,0.001,0.001,0.002,0.003,0.004,0.0,0.0460063897764,0.0,0.00202272727273,YES

0.0,0.0,0.0,0.002,0.002,0.0,0.001,0.0,0.001,0.003,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.002,0.003,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.002,0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.004,0.0,0.0,0.0,0.039312039312,0.0,0.000681818181818,YES

0.007,0.008,0.008,0.008,0.0,0.0,0.008,0.0,0.0,0.007,0.0,0.0,0.0,0.0,0.0,0.013,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.006,0.013,0.0,0.0,0.007,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.007,0.0,0.0,0.0,0.0387096774194,0.0,0.00209090909091,YES

0.002,0.003,0.003,0.005,0.003,0.0,0.007,0.002,0.0,0.002,0.006,0.0,0.0,0.0,0.0,0.001,0.001,0.0,0.001,0.001,0.003,0.0,0.001,0.003,0.002,0.001,0.0,0.001,0.002,0.0,0.0,0.0,0.0,0.0,0.002,0.001,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.007,0.0,0.0,0.0,0.952380952381,0.0478359908884,0.0,0.00143181818182,YES

0.002,0.003,0.002,0.002,0.001,0.0,0.001,0.001,0.0,0.003,0.0,0.0,0.0,0.0,0.008,0.001,0.002,0.0,0.002,0.001,0.001,0.0,0.001,0.0,0.001,0.002,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.021,0.01,0.989010989011,0.105385060799,0.0,0.00161363636364,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.079,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.0,0.0,0.0,0.0,0.0,0.0,0.008,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01,1.0,0.0927947598253,0.0,0.00240909090909,YES

0.009,0.038,0.012,0.029,0.006,0.002,0.025,0.01,0.0,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.005,0.0,0.0,0.002,0.004,0.006,0.0,0.0,0.01,0.013,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.016,0.002,0.01,0.0,0.00789622109419,0.0,0.00477272727273,YES

0.0,0.002,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.002,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.001,0.0,0.0,0.0529240539825,0.0,0.000295454545455,YES

0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.002,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.014,0.0,0.182648401826,0.0,0.000659090909091,YES

0.002,0.001,0.002,0.001,0.002,0.0,0.006,0.0,0.001,0.004,0.001,0.0,0.0,0.004,0.002,0.002,0.0,0.001,0.001,0.001,0.002,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.001,0.0,0.0,0.0,0.007,0.001,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.015,0.007,0.992248062016,0.0570039770217,0.0,0.00159090909091,YES

0.003,0.008,0.003,0.003,0.001,0.0,0.007,0.003,0.0,0.003,0.003,0.0,0.0,0.003,0.002,0.005,0.001,0.001,0.001,0.003,0.001,0.002,0.0,0.002,0.002,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.006,0.002,0.0,0.0,0.0293847566575,0.0,0.00168181818182,YES

0.001,0.001,0.001,0.001,0.003,0.0,0.004,0.001,0.0,0.004,0.002,0.0,0.0,0.0,0.003,0.001,0.004,0.002,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.001,0.001,0.0,0.001,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.0270676691729,0.0,0.000931818181818,YES

0.0,0.0,0.005,0.002,0.002,0.0,0.004,0.001,0.0,0.002,0.0,0.0,0.0,0.002,0.0,0.002,0.001,0.002,0.001,0.0,0.001,0.0,0.003,0.0,0.001,0.0,0.0,0.001,0.0,0.002,0.0,0.0,0.0,0.004,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.012,0.991596638655,0.0569377990431,0.0,0.00118181818182,YES

0.0,0.006,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012,0.0,0.0,0.0,0.024,0.0,0.0,0.0,0.0,0.006
,0.0,0.0,0.012,0.0,0.006,0.0,0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.012,0.031,0.0,0.063829787234,
0.0,0.00275,YES

0.007,0.017,0.004,0.002,0.001,0.0,0.0,0.002,0.0,0.004,0.001,0.0,0.0,0.001,0.001,0.001,0.001,0.002,0.002
,0.0,0.0,0.001,0.005,0.001,0.012,0.0,0.001,0.004,0.0,0.0,0.0,0.0,0.005,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.002,
0.007,0.004,0.001,0.96,0.0260037445392,0.0,0.00206818181818,YES

0.005,0.007,0.005,0.012,0.0,0.001,0.003,0.003,0.001,0.003,0.001,0.0,0.0,0.0,0.0,0.004,0.0,0.004,0.003,0.
004,0.0,0.001,0.002,0.002,0.005,0.0,0.002,0.005,0.001,0.0,0.0,0.0,0.008,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0
02,0.007,0.018,0.008,0.379310344828,0.0193268910363,0.0,0.00272727272727,YES

0.012,0.008,0.008,0.015,0.001,0.0,0.001,0.002,0.002,0.004,0.002,0.001,0.0,0.004,0.002,0.003,0.0,0.006,
0.001,0.001,0.0,0.002,0.002,0.003,0.004,0.0,0.001,0.008,0.0,0.0,0.0,0.0,0.016,0.001,0.001,0.001,0.0,0.0,
0.0,0.001,0.0,0.007,0.001,0.004,1.0,0.0443237060337,0.0,0.00284090909091,YES

0.002,0.005,0.002,0.0,0.0,0.0,0.002,0.002,0.0,0.003,0.0,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.001,0.002,0.0,0.
0,0.0,0.002,0.002,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.001,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.002,0.001,0.
001,0.0,0.174334140436,0.0,0.000863636363636,YES

0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.002,0.0,0.0,0.0,0.0,0.006,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.002,
0.004,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.002,0.95652173913,0.06
57142857143,0.0,0.000568181818182,YES

0.003,0.007,0.003,0.006,0.002,0.0,0.001,0.002,0.0,0.004,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.003,0.003,
0.0,0.002,0.002,0.002,0.004,0.0,0.002,0.013,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.005,
0.005,0.0,0.985074626866,0.174934725849,0.0,0.00170454545455,YES

0.002,0.021,0.002,0.006,0.001,0.001,0.001,0.001,0.001,0.006,0.001,0.0,0.0,0.004,0.001,0.003,0.001,0.00
1,0.001,0.001,0.0,0.002,0.001,0.001,0.004,0.0,0.0,0.002,0.0,0.001,0.0,0.0,0.004,0.001,0.0,0.0,0.0,0.0,0.0,
0.0,0.001,0.007,0.013,0.008,1.0,0.0714002380008,0.0,0.00227272727273,YES

0.0,0.0,0.0,0.003,0.0,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.009,0.008,0.003,0.002,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.003,0.0,0.969696969697,0.0
558375634518,0.0,0.000954545454545,YES

0.002,0.007,0.001,0.005,0.0,0.0,0.001,0.001,0.001,0.003,0.0,0.0,0.0,0.002,0.0,0.001,0.001,0.001,0.001,0.
001,0.0,0.002,0.001,0.0,0.004,0.001,0.0,0.001,0.0,0.002,0.0,0.0,0.007,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
005,0.036,0.002,1.0,0.0374181478017,0.0,0.00204545454545,YES

0.006,0.0,0.004,0.002,0.003,0.0,0.004,0.002,0.001,0.004,0.001,0.0,0.0,0.004,0.001,0.001,0.001,0.0,0.0,0.
001,0.001,0.001,0.001,0.002,0.003,0.0,0.001,0.0,0.002,0.0,0.0,0.0,0.001,0.001,0.002,0.0,0.0,0.0,0.0,0.0,0
.001,0.005,0.006,0.001,0.943396226415,0.111814345992,0.0,0.00143181818182,YES

0.003,0.0,0.005,0.0,0.002,0.0,0.003,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.003,0.001,0.001,0.001,0.0,0.001,0.0,0.
0,0.002,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.01,0.004,0.0,0.3
16498316498,0.0,0.001,YES

0.003,0.003,0.003,0.006,0.0,0.0,0.003,0.002,0.083,0.001,0.002,0.0,0.0,0.003,0.0,0.003,0.0,0.002,0.002,0.
003,0.0,0.003,0.0,0.002,0.0,0.0,0.002,0.003,0.0,0.0,0.0,0.0,0.013,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.004,0
.0,0.0,0.978260869565,0.158894645941,0.0,0.00334090909091,YES

0.002,0.002,0.001,0.003,0.002,0.0,0.006,0.0,0.0,0.0,0.004,0.0,0.0,0.004,0.003,0.002,0.0,0.0,0.0,0.0,0.002
,0.001,0.001,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.003,0.0,0.001,0.0,0.0,0.0,0.0,0.001,0.001,0.003,0.001,
0.001,0.970873786408,0.114827201784,0.0,0.00102272727273,YES

0.006,0.0,0.002,0.0,0.009,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.004,0.0,0.0,0.0,0.0
,0.0,0.0,0.006,0.0,0.0,0.0,0.0,0.0,0.0,0.005,0.001,0.0,0.0,0.0,0.0,0.002,0.002,0.0,0.004,0.80701754386,0.
0633333333333,0.0,0.000977272727273,YES

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.024,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0750670241287,0.0,0.000545454545455,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0306905370844,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.368421052632,0.0250990752972,0.0,2.27272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.00383141762452,0.0,0.0,NO

0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0413625304136,0.0,2.27272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00990099009901,0.0970220941402,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0192837465565,0.0,0.0,NO

0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0661971830986,0.0,6.81818181818e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.571428571429,0.00406032482599,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.00461893764434,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00895316804408,0.0,2.27272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0978723404255,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.738219895288,0.10346695558,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00271370420624,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.571428571429,0.0100430416069,0.0,2.27272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0132450331126,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.571428571429,0.0142566191446,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0752688172043,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0151515151515,0.0482456140351,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.625,0.0337552742616,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.001,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0340557275542,0.0,4.5454545454
5e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0234192037471,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0357142857143,0.0996441281139,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.444444444444,0.00674662668666,0.0,2.
27272727273e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5625,0.00332571191021,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.002,0.0,0.0,0.0,0.0,0.0,0.001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00947867298578,0.0,6.818181818
18e-05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.882352941176,0.0625,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0229645093946,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.681818181818,0.00842266462481,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.00353356890459,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0782422293676,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0232558139535,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0235732009926,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.00272851296044,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.545454545455,0.0417118093174,0.0,0.0,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.803278688525,0.152882205514,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.804347826087,0.0383014154871,0.0,0.0,N
O

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.375,0.0666666666667,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.428571428571,0.00849514563107,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.002,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.2,0.025641025641,0.0,4.54545454545e-
05,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.819672131148,0.152882205514,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0555555555556,0.0363636363636,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.005,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0172413793103,0.0901320901321,0.0,0.0,
NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.40625,0.0637450199203,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.10824742268,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.000920245398773,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.018,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.263636363636,0.15514809591,0.0,0.000
409090909091,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.060162601626,0.0,0.0,NO

0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25,0.00344530577089,0.0,0.0,NO

Κώδικας 6.0: Αρχείο εκτέλεσης διαδικασίας εξαγωγής εκπαιδευμένων μοντέλων “train.sh”

```
testpath="python/NEW_TRAIN.arff"


echo "MultilayerPerceptron"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.functions.MultilayerPerceptron -t $testpath -d new_train_models/nnet.model  -no-cv -v


echo "SMO TRAIN"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.functions.SMO -t $testpath -d new_train_models/smo.model -no-cv -v


echo "BayesNet TRAIN"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.bayes.BayesNet -t $testpath -d new_train_models/bayesnet.model  -no-cv -v


echo "NaiveBayes TRAIN"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.bayes.NaiveBayes -t $testpath -d new_train_models/naive.model  -no-cv -v


echo "J48 TRAIN"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.trees.J48 -t $testpath -d new_train_models/j48.model  -no-cv -v


echo "KSTAR TRAIN"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.lazy.KStar -t $testpath -d new_train_models/lazy_KSTAR.model  -no-cv -v
```

Κώδικας 7.0: Αρχείο εκτέλεσης διαδικασίας πρόβλεψης κατηγορίας “test_file.sh”

```
testpath="python/NEW_TEST.arff"

range="0"


echo "MultilayerPerceptron"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.functions.MultilayerPerceptron -T $testpath -l new_train_models/nnet.model -o -p $range > "results/MP"

echo "SMO"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.functions.SMO -T $testpath -l new_train_models/smo.model -o -p $range > "results/SMO"

echo "BayesNet"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.bayes.BayesNet -T $testpath -l new_train_models/bayesnet.model -o -p $range > "results/BN"

echo "NaiveBayes"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.bayes.NaiveBayes -T $testpath -l new_train_models/naive.model -o -p $range > "results/NB"

echo "J48"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.trees.J48 -T $testpath -l new_train_models/j48.model -o -p $range > "results/J48"

echo "KStar"

java -classpath "/usr/share/java/weka.jar" weka.classifiers.lazy.KStar -T $testpath -l new_train_models/lazy_KSTAR.model -o -p $range > "results/KS"
```

Κώδικας 8.0: Αρχείο “show_statistics.py” - Εμφάνιση αξιολόγησης μοντέλων εκπαίδευσης

```
def positions(target, source):
	'''Produce all positions of target in source'''
	pos = -1
	lst=[]
	try:
		while True:
			pos = source.index(target, pos + 1)
			lst.append(pos)
	except ValueError:
		return lst


algorithms={"MultilayerPerceptron":"MP", "NaiveBayes":"NB", "BayesNet":"BN", "Tree J48(c4)":"J48", "KStar":"KS", "SMO":"SMO"}
folder="results/"


print "\n"*3


for algorithm in algorithms.keys():
	filename=algorithms[algorithm]
	f=open(folder+filename,'r').read()
	instances_num=len(f.split("\n"))-7
	wrong_inst=len(positions('+',f))
	print algorithm
	total_inst=wrong_inst+instances_num
	print "Total Instances",total_inst
	if total_inst==0:
		print "No Instances given"
		exit()
```

```
if wrong_inst==0:
        corr="100"
        wrong="0"
else:
        corr=str(round(100-(wrong_inst*100/(instances_num+0.0)),3))
        wrong=str(round((wrong_inst*100/(instances_num+0.0)),3))
print "Correct by: ",corr+"%"
print "Wrong by: ",wrong+"%"
print "*"*30
print
```

Κώδικας 9.0: Αρχείο “extract_vectors.py” - Εκτύπωση διανυσμάτων κατηγοριοποίησης

```
def positions(target, source):
	'''Produce all positions of target in source'''
	pos = -1
	lst=[]
	try:
		while True:
			pos = source.index(target, pos + 1)
			lst.append(pos)
	except ValueError:
		return lst
txt=open('results/BN').read()
ar=txt.split(")")
for line in ar:
	print line
```

Κώδικας 10.0: Αρχείο “featured_words.txt”

```
asian
porno
japanese
blonde
amateur
milf
huge
fucked
tube
fucks
hairy
forced
russian
babe
horny
latina
hardcore
dick
fucking
bisexual
pussy
fuck
sperm
fingering
sucking
masturbating
tits
ass
xxx
boobs
slut
blowjob
sexy
```

suck

handjob

lesbian

nude

threesome

cock

anal

masturbation

pornos

pornostars

stripper

Παράρτημα Α.2: Κατάλογος Ιστοσελίδων

Ιστοσελίδες Πορνογραφικού Περιεχομένου (για την εκπαίδευση ταξινομητή)

http://www.xvideos.com	http://www.porndig.com	http://www.tubexclips.com
http://www.youporn.com	http://www.xnxx.com	http://www.pornhub.com
http://y-suck.com	http://arionmovies.com	http://pornsharing.com/cock-sucking_c
http://www.porn.com	http://www.redtube.com	http://www.youngpornvideos.com
http://o-suck.com	http://www.pornmd.com	http://www.bps-europe.net
http://xxxfuckporn.com	http://www.outfuck.com	http://mrs-porn.com
http://local732.org	http://cumlouder.com	http://theluckylonely.com
http://m.pornsharing.com	http://m.spankbang.com	http://m.xhamster.com
http://hardfuckgirls.biz	http://m.xcafe.com	http://edenfantasys.com
http://tube4us.com	http://moiporn.com	http://fineretroporn.com
http://nicexxxtube.com	http://pornmount.com	http://tubethrill.com
http://magicaltube.com	http://tubegolf.com	http://largefucktube.com
http://enjoyfuck.com	http://21hub.com	http://nudevista.co

		m
http://la-xxx.com	http://dailybasis.com	http://www.porn365.com/Nude.html
http://nudeinfrance.com	http://www.celeb-porn.us/	http://multi.xnxx.com/mode1/p-1
http://www.pornozot.com/		

http://www.porndig.com/	http://rape-video.com	http://forcedfuck.org
http://raped.ws	http://rapedtube.org	http://rape-xxx.com
http://thescreams.net	http://rapeporn.tv	http://therape.net
http://rapeinass.com	http://forcedsextube.org	http://forcedtaboo.com
http://painedporn.com	http://forcedsexvideo.net	http://rapesex.tv
http://forced-tube.net	http://rape-portal.biz	http://pornrapetube.com
http://rapeme.biz	http://rapetube.tv	http://incest-uncensored.com
http://rape-tube.net	http://rudetaboo.com	http://rapefantasytube.com
http://youporn.com	http://anyporn.com	http://familyincest.name
http://sexmomfuck.com	http://familyincesttube.com	http://rudeextreme.com
http://xvideos.com	http://pornhub.com	http://www.swapsmut.com
http://alohatube.com	http://fuq.com	http://dinotube.com
http://xnxx.com	http://tube8.com	http://hornbunny.com
http://yes.xxx	http://xxxmomtube.com	http://xvideohard.com

Ιστοσελίδες μη Πορνογραφικού Περιεχομένου (για εκπαίδευση ταξινομητή)

http://en.wikipedia.org/wiki/Sexual_arousal	http://en.wikipedia.org/wiki/Deep-throating_%28sexual_act%29
http://en.wikipedia.org/wiki/Oral_sex	http://en.wikipedia.org/wiki/Orgasm_control
http://en.wikipedia.org/wiki/Anal_sex	http://en.wikipedia.org/wiki/Child_sexuality
http://en.wikipedia.org/wiki/Masturbation	http://en.wikipedia.org/wiki/Sexual_penetration
http://en.wikipedia.org/wiki/Feminist_sex_wars	http://en.wikipedia.org/wiki/History_of_erotic_depictions
http://en.wikipedia.org/wiki/Sexual_revolution	http://www.saidit.org/archives/jul01/mediaglance.html
http://en.wikipedia.org/wiki/Pornography	http://en.wikipedia.org/wiki/Amateur_pornography
http://en.wikipedia.org/wiki/Sexual_fetishism	http://www.plannedparenthood.org/teens/sex/whats-sex
http://en.wikipedia.org/wiki/Theft_of_fire	http://axxomovies.org/?s=sex
http://en.wikipedia.org/wiki/Prometheus	http://www.nhs.uk/chq/Pages/975.aspx?CategoryID=54
http://spie.org/x41303.xml	http://www.youthoria.org/home/life/sex/safe-sex/1238761856.871/
https://www.lushstories.com/	http://sexstories-xxx.com/14479_jim-s-teenage-daughter-part-3
http://www.humandigest.com/blog/	http://sexstories-xxx.com/13851_mary-s-marvelous-family-chapter-2
http://sexstories-xxx.com/13841_year-thirteen	http://sexstories-xxx.com/13849_sarah-and-daddy-part-3
http://sexstories-xxx.com/13863_the-	http://sexstories-xxx.com/14485_my-

academy-3	daddy-s-cock
http://sexstories-xxx.com/13858_titcage-chapter-41	http://sexstories-xxx.com/13880_jim-s-teenage-daughter
http://sexstories-xxx.com/13883_vixens-lust	http://sexstories-xxx.com/14671_cocksucking-ex-girlfriends
http://www.freechatnow.com/	http://sexstories-xxx.com/14675_hope-s-stratching-day
http://www.reddit.com/r/sexstories/	http://sexstories-xxx.com/14668_pizza-or-tits-your-choice
http://www.12chats.com/	http://sexstories-xxx.com/13952_fun-for-the-wife-and-cindy
http://www.midomi.com/	http://thoughtcatalog.com/justin-alexander/2014/10/supermarket-sex/
http://www.positive.org/JustSayYes/safesex.html	http://thoughtcatalog.com/adrienne-west/2014/09/392221/
http://thoughtcatalog.com/tag/sex-stories/	http://www.femalefirst.co.uk/relationships/Masturbation-41.html
http://www.sexstories.com/	http://www.wikihow.com/Have-Safer-Sex

http://www.humandigest.com/blog/2015/03/sex_during_the_bangalore_hyderabad_bus_trip_ii.html
http://www.thehealthsite.com/sexual-health/7-sinful-sex-positions-for-deeper-penetration-sex-guide-for-beginners/
http://forums.na.leagueoflegends.com/board/showthread.php?t=1372513&page=10
http://forum.com2us.com/forum/main-forum/summoner-s-war/general-ab/623185-toa-80-fuck-you-seara-xd
http://bedsider.org/features/247-how-to-make-sex-safer-in-4-simple-steps
http://www.betterhealth.vic.gov.au/bhcv2/bhcarticles.nsf/pages/Safe_sex?open
http://womenshealth.gov/hiv-aids/preventing-hiv-infection/practice-safer-sex.html

http://www.yourtango.com/experts/sean-jameson/how-masturbate-women
http://www.womenshealthmag.com/sex-and-relationships/masturbation-fun
http://www.womenshealthmag.com/sex-and-relationships/types-of-female-orgasm
http://www.huffingtonpost.com/2015/01/14/reasons-women-should-masturbate_n_6172092.html
http://prettyladysmiles.com/masturbation-techniques-for-women/
http://sexuality.about.com/od/anatomyresponse/ht/masturbatewomen.htm
http://jezebel.com/5107639/92-of-women-masturbate-but-how-often-do-they-do-it
http://std.about.com/od/sextips/a/How-To-Have-Sex-Safely.htm
http://www.projectx.net.au/Sexual-Health-News/to-fuck-or-be-fucked-that-is-the-question-myths-about-anal-sex-that-help-the-spread-of-hiv/menu-id-121.html
http://the3bromigos.com/2013/12/09/5-ways-sex-without-condom/
http://www.sexualityandu.ca/sexual-health/pregnancy/when-is-it-safe-to-have-sex
http://www.babycentre.co.uk/x536429/is-it-safe-to-have-sex-during-pregnancy
https://answers.yahoo.com/question/index?qid=20080818213128AAEO1ig
http://citysweetmermaid.tumblr.com/post/104871687485/girly-made-it-safely-to-west-bubble-fuck-and-i
http://www.irishmanabroad.com/2012/02/drive-safely-slow-the-fuck-down/
http://www.plannedparenthood.org/health-info/stds-hiv-safer-sex/safer-sex
http://thoughtcatalog.com/nikki-hunter/2015/01/heres-what-happened-when-i-stayed-after-work-one-night-with-my-hot-male-underling/
http://thoughtcatalog.com/adrienne-west/2015/01/here-is-the-sexy-game-that-brought-my-boyfriend-and-i-closer-together-while-we-avoided-the-freezing-weather-outside/
http://thoughtcatalog.com/zach-armstrong/2014/12/my-boss-made-the-mistake-of-sending-me-and-my-cute-co-worker-on-an-out-of-town-business-trip/
http://thoughtcatalog.com/melanie-berliet/2014/11/5-real-sex-stories-that-will-make-you-really-horny/
http://thoughtcatalog.com/adrienne-west/2014/10/my-boyfriend-asked-me-to-fulfill-a-very-controversial-fantasy-heres-what-happened-when-i-did/

http://thoughtcatalog.com/adrienne-west/2014/10/how-to-punish-a-very-very-bad-girl/
http://thoughtcatalog.com/adrienne-west/2014/10/how-to-be-the-kind-of-guy-women-love-having-sex-with/
http://thoughtcatalog.com/adrienne-west/2014/09/i-seduced-my-high-school-english-teacher/
http://thoughtcatalog.com/nupur-saraswat/2014/09/i-had-an-orgasm-in-the-middle-of-class-this-is-what-i-was-thinking-about/
http://thoughtcatalog.com/erin-cossetta/2014/09/27-men-share-what-made-the-most-unforgettable-blow-job-of-my-life/
http://thoughtcatalog.com/alexis-caputo/2014/09/10-sexual-fantasies-men-have-that-they-never-tell-their-girlfriends-about/
http://thoughtcatalog.com/alexis-caputo/2014/09/19-women-talk-about-the-surprising-sexual-fantasies-they-would-never-ever-tell-their-boyfriends-about/
http://thoughtcatalog.com/adrienne-west/2014/09/my-best-friends-dad-gave-me-the-weirdest-sex-toy-in-the-world/
http://www.lovepanky.com/women/how-to-tips-and-guide-for-women/how-to-talk-dirty-to-a-guy
http://www.buzzfeed.com/caseygueren/how-women-talk-about-sex
http://www.buzzfeed.com/augustafalletta/what-was-the-sex-talk-like-for-you
https://eagleman6788.wordpress.com/2012/03/29/how-to-talk-dirty-50-examples-that-will-make-you-blush/

Ιστοσελίδες Πορνογραφικού Περιεχομένου (για αξιολόγηση ταξινομητή)

http://bj-bitches.com	http://www.gaybin.com	http://www.worldsex.com
http://alohatube.com	http://fapdu.com	http://xpornking.com
http://brownxxxtube.com	http://analxxxx.com	http://www.3pornstarmovies.com
http://manporn.xxx	http://www.kindgirls.com	http://www.funsexporn.com
http://givemeyoung.com	http://maturevideos.xxx	https://www.yesporn.xxx
http://www.giga.xxx	http://hd.xtapes.tv	http://www.sunporno.com
http://tube18.xxx	http://momxxx.co	http://pornxxxtubes.com
http://www.mporn.xxx	http://www.pornmummy.com	http://www.santasporngirls.com
http://oixxx.com/en/	http://www.wowporn.xxx	http://www.yourdailypornstars.com
http://www.pornburst.xxx	http://www.pornhd.com	http://www.youramateurporn.com
http://topxlive.com	http://www.xxx.xxx	http://girlsavenue.com
http://xxxpawn.xxx	http://www.8teenxxx.com	http://www.tube8.com
http://www.4tube.com	http://www.69style.com	http://www.gonzoxxxmovies.com
http://yea.xxx	http://www.xxx.com	http://www.milfpornl

		overs.com
http://xxxdessert.com	http://www.gayfuror.com	http://www.pureblowjob.com
http://www.youx.xxx	http://www.cumlouder.com	http://www.bigboobspornpics.com
http://xkeezmovies.com	http://xxxsexanal.com	http://www.perfectgirls.net
http://fck.xxx	http://brandporno.com	http://www.yourlust.com

Ιστοσελίδες μη Πορνογραφικού Περιεχομένου (για αξιολόγηση ταξινομητή)

http://langs.eserver.org/linell/chapter01.html	https://frontendmasters.com/courses/functional-javascript/
http://en.wikipedia.org/wiki/Vagina	http://www.latimes.com/la-fi-adelphia2feb02-story.html
http://www.photoeye.com/gallery/	http://www.askmen.com/dating/love_tip_250/251_love_tip.html
http://www.hbo.com/sex-and-the-city	http://en.wikipedia.org/wiki/Sex_and_the_City
http://www.marriageheat.com/	http://en.wikipedia.org/wiki/List_of_programs_broadcast_by_HBO
http://www.ubuntu.com/	http://en.wikipedia.org/wiki/Islamic_views_on_oral_sex
http://en.wikipedia.org/wiki/Human_teeth	https://www.pureromance.com/shop/Adult-Sex-Toys
http://www.modelmayhem.com/list/186838	https://www.goodreads.com/story/show/244197-poetry?chapter=2
http://de.wikipedia.org/wiki/Oral	https://www.fanfiction.net/s/10821073/1/Sex-Scene-in-the-Car
http://en.wikipedia.org/wiki/Condom	https://www.fanfiction.net/s/6005163/1/Indescribable-love
http://simple.wikipedia.org/wiki/Sex_organ	http://www.angelfire.com/ga/wkb/titanic.html
http://indosex.biz	http://en.wikipedia.org/wiki/List_of_organs_of_the_human_body
http://prometheuscode.net/	http://shop.oreilly.com/product/9780596005689.do
http://bazaraki.com/	http://briff.me/2015/03/09/hollywood-sex-scenes/

http://alterego.ae/bodynude/	http://www.bremencapital.com/keyra-augustina/
http://www.pbs.org/wnet/nature/	http://www.scottyancey.com/hot-black-models/
http://www.nature.org/	https://www.fanfiction.net/s/6646043/1/Jack-and-Rose-The-Sex-Chronicles
http://simple.wikipedia.org/wiki/Penis	http://feminismandreligion.com/rosemary-radford-ruether-on-feminism/
http://xaxor.com/engine	http://en.wikipedia.org/wiki/Male_reproductive_system
http://xaxor.com/wallpapers	http://dictionary.reference.com/browse/racism
http://collections.vam.ac.uk	http://xaxor.com/female/3205-girls-in-sexy-underwear-.html
http://www.photographersgallery.com/	http://folk.uio.no/thomas/po/whatisfeminism.html
http://www.vam.ac.uk/content/galleries/level-3/room-100-photographs-gallery/	
https://www.bodywearstore.com/brand/joe-snyder/?gclid=CNrk-tC07sQCFWfJtAodKg0Aig	
http://www.couriermail.com.au/news/opinion/big-tv-wants-to-show-hardcore-programming-earlier-to-compete-with-netflix-and-the-like/story-fnihsr9v-1227251796979?nk=e48aebaf16abbe3e1e3bea50c985f6bb	
http://www.freep.com/story/opinion/columnists/brian-dickerson/2015/04/07/rethinking-sex-registry/25434171/	
http://www.independent.co.uk/news/world/middle-east/yazidi-sex-slaves-gangraped-in-public-by-isis-fighters-harrowing-accounts-reveal-10166875.html	
http://www.dreamstime.com/photos-images/sexy-underwear-female-model.html	
http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html	
http://www.nydailynews.com/entertainment/movies/50-sexiest-movie-scenes-article-1.2114249	
http://www.therichest.com/rich-list/most-popular/10-female-body-parts-men-find-the-sexiest/	
https://www.Google.com/search?q=best+photographs+2015&client=browser-	

ubuntu&hs=OfI&channel=fe&hl=en&source=lnms&tbm=isch&sa=X&ei=_ykpVfzcII
bfaqCRgMgM&ved=0CAcQ_AUoAQ&biw=1366&bih=563

Ιστοσελίδες πορνογραφικού περιεχομένου (για εύρεση χαρακτηριστικών λέξεων)

http://www.xvideos.com	http://forcedfuck.org
http://www.porndig.com	http://raped.ws
http://www.tubexclips.com	http://rapedtube.org
http://www.youporn.com	http://rape-xxx.com
http://www.xnxx.com	http://thescreams.net
http://www.pornhub.com	http://rapeporn.tv
http://y-suck.com	http://therape.net
http://arionmovies.com	http://rapeinass.com
http://pornsharing.com/cock-sucking_c	http://forcedsextube.org
http://www.porn.com	http://forcedtaboo.com
http://www.youngpornvideos.com	http://painedporn.com
http://www.redtube.com	http://forcedsexvideo.net
http://o-suck.com	http://rapesex.tv
http://www.bps-europe.net	http://forced-tube.net
http://www.pornmd.com	http://rape-portal.biz
http://xxxfuckporn.com	http://rapeme.biz
http://www.outfuck.com	http://rapetube.tv
http://mrs-porn.com	http://rape-tube.net
http://thingsthatmakeyouwet.tumblr.com	http://pornrapetube.com
http://m.xbef.com	http://rudetaboo.com
http://theluckylonely.com	http://incest-uncensored.com
http://cumlouder.com	http://rapefantasytube.com
http://local732.org	http://familyincest.name
http://xxwet-kinky-fuckeryxx.tumblr.com	http://youporn.com
http://uncoverbitch.tumblr.com	http://anyporn.com
http://m.pornsharing.com	http://sexmomfuck.com

http://fuckmesuperhard.tumblr.com	http://familyincesttube.com
http://m.spankbang.com	http://rudeextreme.com
http://m.xhamster.com	http://xvideos.com
http://hardfuckgirls.biz	http://www.swapsmut.com
http://m.xcafe.com	http://pornhub.com
http://longhardfucking.tumblr.com	http://alohatube.com
http://edenfantasys.com	http://fuq.com
http://zeze2014.tumblr.com	http://dinotube.com
http://tube4us.com	http://xnxx.com
http://moiporn.com	http://tube8.com
http://phevosd.tumblr.com	http://hornbunny.com
http://fineretroporn.com	http://yes.xxx
http://nicexxxtube.com	http://xxxmomtube.com
http://pornmount.com	http://xvideohard.com
http://tubethrill.com	http://onlysexhere.tumblr.com
http://magicaltube.com	http://nudeinfrance.com
http://tubegolf.com	http://www.celeb-porn.us/
http://largefucktube.com	http://multi.xnxx.com/mode1/p-1
http://enjoyfuck.com	http://www.porn365.com/Nude.html
http://nudevista.com	http://www.pornozot.com/
http://la-xxx.com	http://www.porndig.com/
http://dailybasis.com	http://rape-video.com

Ιστοσελίδες μη πορνογραφικού περιεχομένου (για εύρεση χαρακτηριστικών λέξεων)

http://en.wikipedia.org/wiki/Sexual_arousal	https://answers.yahoo.com/question/index?qid=20080818213128AAEO1ig
http://en.wikipedia.org/wiki/Oral_sex	http://citysweetmermaid.tumblr.com/post/104871687485/girly-made-it-safely-to-west-bubble-fuck-and-i
http://en.wikipedia.org/wiki/Deep-throating_%28sexual_act%29	http://www.irishmanabroad.com/2012/02/drive-safely-slow-the-fuck-down/
http://en.wikipedia.org/wiki/Orgasm_control	http://www.positive.org/JustSayYes/safesex.html
http://en.wikipedia.org/wiki/Anal_sex	http://www.plannedparenthood.org/health-info/stds-hiv-safer-sex/safer-sex
http://en.wikipedia.org/wiki/Child_sexuality	http://thoughtcatalog.com/nikki-hunter/2015/01/heres-what-happened-when-i-stayed-after-work-one-night-with-my-hot-male-underling/
http://en.wikipedia.org/wiki/Masturbation	http://thoughtcatalog.com/adrienne-west/2015/01/here-is-the-sexy-game-that-brought-my-boyfriend-and-i-closer-together-while-we-avoided-the-freezing-weather-outside/
http://en.wikipedia.org/wiki/Sexual_penetration	http://thoughtcatalog.com/zach-armstrong/2014/12/my-boss-made-the-mistake-of-sending-me-and-my-cute-co-worker-on-an-out-of-town-business-trip/
http://en.wikipedia.org/wiki/Feminist_sex_wars	http://thoughtcatalog.com/melanie-berliet/2014/11/5-real-sex-stories-that-will-make-you-really-horny/
http://en.wikipedia.org/wiki/History_of_erotic_depictions	http://thoughtcatalog.com/adrienne-west/2014/10/my-boyfriend-asked-me-to-fulfill-a-very-controversial-fantasy-heres-what-happened-when-i-did/
http://en.wikipedia.org/wiki/Sexual_revolution	http://thoughtcatalog.com/justin-alexander/2014/10/supermarket-sex/
http://www.saidit.org/archives/jul01/mediaglance.html	
http://en.wikipedia.org/wiki/Pornography	
http://en.wikipedia.org/wiki/Amateur_pornography	
http://en.wikipedia.org/wiki/Sexual_fetishism	
http://www.plannedparenthood.org/teens/sex/whats-sex	
http://www.thehealthsite.com/sexual-health/7-sinful-sex-positions-for-deeper-penetration-sex-guide-for-beginners/	

http://forums.na.leagueoflegends.com/board/showthread.php?t=1372513&page=10 http://forum.com2us.com/forum/main-forum/summoner-s-war/general-ab/623185-toa-80-fuck-you-seara-xd http://www.nhs.uk/chq/Pages/975.aspx?CategoryID=54 http://en.wikipedia.org/wiki/Prometheus http://en.wikipedia.org/wiki/Theft_of_fire http://axxomovies.org/?s=sex http://spie.org/x41303.xml http://www.midomi.com/ http://www.youthoria.org/home/life/sex/safe-sex/1238761856.871/ http://www.wikihow.com/Have-Safer-Sex http://bedsider.org/features/247-how-to-make-sex-safer-in-4-simple-steps http://www.betterhealth.vic.gov.au/bhcv2/bhcarticles.nsf/pages/Safe_sex?open http://womenshealth.gov/hiv-aids/preventing-hiv-infection/practice-safer-sex.html http://www.yourtango.com/experts/sean-jameson/how-masturbate-women http://www.womenshealthmag.com/sex-and-relationships/masturbation-fun http://www.womenshealthmag.com/sex-and-relationships/types-of-female-orgasm http://www.huffingtonpost.com/2015/01/14/reasons-women-should-masturbate_n_6172092.html http://www.femalefirst.co.uk/relationships/Masturbation-41.html http://prettyladysmiles.com/masturbation-techniques-for-women/	http://thoughtcatalog.com/tag/sex-stories/ http://thoughtcatalog.com/adrienne-west/2014/10/how-to-punish-a-very-very-bad-girl/ http://thoughtcatalog.com/adrienne-west/2014/10/how-to-be-the-kind-of-guy-women-love-having-sex-with/ http://thoughtcatalog.com/adrienne-west/2014/09/392221/ http://thoughtcatalog.com/adrienne-west/2014/09/i-seduced-my-high-school-english-teacher/ http://thoughtcatalog.com/nupur-saraswat/2014/09/i-had-an-orgasm-in-the-middle-of-class-this-is-what-i-was-thinking-about/ http://thoughtcatalog.com/erin-cossetta/2014/09/27-men-share-what-made-the-most-unforgettable-blow-job-of-my-life/ http://thoughtcatalog.com/alexis-caputo/2014/09/10-sexual-fantasies-men-have-that-they-never-tell-their-girlfriends-about/ http://thoughtcatalog.com/alexis-caputo/2014/09/19-women-talk-about-the-surprising-sexual-fantasies-they-would-never-ever-tell-their-boyfriends-about/ http://thoughtcatalog.com/adrienne-west/2014/09/my-best-friends-dad-gave-me-the-weirdest-sex-toy-in-the-world/ https://www.lushstories.com/ http://www.humandigest.com/blog/ http://www.humandigest.com/blog/2015/03/sex_during_the_bangalore_hyderabad_bus_trip_ii.html

http://sexuality.about.com/od/anatomyresponse/ht/masturbatewomen.htm http://jezebel.com/5107639/92-of-women-masturbate-but-how-often-do-they-do-it http://std.about.com/od/sextips/a/How-To-Have-Sex-Safely.htm http://www.projectx.net.au/Sexual-Health-News/to-fuck-or-be-fucked-that-is-the-question-myths-about-anal-sex-that-help-the-spread-of-hiv/menu-id-121.html http://the3bromigos.com/2013/12/09/5-ways-sex-without-condom/ http://www.sexualityandu.ca/sexual-health/pregnancy/when-is-it-safe-to-have-sex http://www.babycentre.co.uk/x536429/is-it-safe-to-have-sex-during-pregnancy http://www.buzzfeed.com/augustafalletta/what-was-the-sex-talk-like-for-you https://eagleman6788.wordpress.com/2012/03/29/how-to-talk-dirty-50-examples-that-will-make-you-blush/ http://www.reddit.com/r/sexstories/ http://www.sexstories.com/ http://www.lovepanky.com/women/how-to-tips-and-guide-for-women/how-to-talk-dirty-to-a-guy http://www.buzzfeed.com/caseygueren/how-women-talk-about-sex	http://sexstories-xxx.com/14485_my-daddy-s-cock http://sexstories-xxx.com/14479_jim-s-teenage-daughter-part-3 http://sexstories-xxx.com/13851_mary-s-marvelous-family-chapter-2 http://sexstories-xxx.com/13849_sarah-and-daddy-part-3 http://sexstories-xxx.com/13841_year-thirteen http://sexstories-xxx.com/13863_the-academy-3 http://sexstories-xxx.com/13858_titcage-chapter-41 http://sexstories-xxx.com/13883_vixens-lust http://sexstories-xxx.com/13880_jim-s-teenage-daughter http://sexstories-xxx.com/14671_cocksucking-ex-girlfriends http://sexstories-xxx.com/14675_hope-s-stratching-day http://sexstories-xxx.com/14668_pizza-or-tits-your-choice http://sexstories-xxx.com/13952_fun-for-the-wife-and-cindy http://www.freechatnow.com/ http://www.12chats.com/

Παράρτημα Α.3: Λέξεις (tokens) κλειδιά σε ιστοσελίδες πορνογραφικού περιεχομένου

asian	porno	japanese	blonde	amateur	milf
huge	fucked	tube	fucks	hairy	forced
russian	babe	horny	latina	hardcore	dick
fucking	bisexual	pussy	fuck	sperm	fingering
sucking	masturbating	tits	ass	xxx	boobs
slut	blowjob	sexy	suck	handjob	lesbian
nude	threesome	cock	anal	masturbation	pornos
pornostars	stripper				