
# ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ


# Κατηγοριοποίηση Κειμένων σε Συνεχή Χώρο με χρήση Gaussian Mixture Models

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του


**ΛΑΖΑΡΟΥ ΔΟΥΓΙΑΚΗ**

**Επιβλέπων :** Βασίλης Διγαλάκης
Καθηγητής Π.Κ.



Χανιά, Οκτώβριος 2013

# Ευχαριστίες

Από τη θέση αυτή, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής μου κ. Βασίλη Διγαλάκη, για τη πολύτιμη καθοδήγηση που μου προσέφερε καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας, καθώς και για την εμπιστοσύνη που μου έδειξε αναθέτοντας μου το έργο. Ακόμα θα ήθελα να ευχαριστήσω θερμά και τον κ. Βασίλη Διακολουκά που με τη βοήθεια και τη συμπαράσταση του συνέβαλε στην ολοκλήρωση της εργασίας αυτής. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου, τους φίλους μου και τους συμφοιτητές μου για τη ψυχολογική στήριξη που μου παρείχαν σε όλη τη διάρκεια των σπουδών μου .

# Περίληψη

Η ταχεία εξάπλωση του διαδικτύου και η συνεχώς αυξανόμενη διάθεση υλικού σε ηλεκτρονική μορφή κάνει τη χρήση τεχνικών Κατηγοριοποίησης Κειμένου, για την εύρεση σχετικών πληροφοριών, ολοένα και πιο απαραίτητη.

Η σπουδαιότητα της Κατηγοριοποίησης Κειμένων μπορεί να γίνει αντιληπτή από την πληθώρα εφαρμογών στις οποίες χρησιμοποιείται για την εξόρυξη πληροφορίας και από τη συνεχή έρευνα που γίνεται πάνω στο τομέα για την ανεύρεση νέων τεχνικών και αλγορίθμων. Στη παρούσα διπλωματική εργασία, εξετάζεται η κατηγοριοποίηση των κειμένων ανάλογα με το θέμα τους χρησιμοποιώντας συνεχή χαρακτηριστικά διανύσματα για την αναπαράσταση των κειμένων, συγκεκριμένα Gaussian Mixture Models.

Η υλοποίηση της εργασίας ξεκινά από την προ-επεξεργασία των κειμένων ώστε να μετατραπούν σε μια πιο συμπαγή μορφή για τον αλγόριθμο κατηγοριοποίησης. Στη συνέχεια ακολουθεί το στάδιο δημιουργίας του λεξικού των κειμένων και η κατανομή των βαρών στους όρους του κειμένου φέρνοντας ένα κείμενο από τη κλασσική αναπαράσταση του ,μέσω λέξεων, σε διανυσματική μορφή (πιο κατανοητή στον υπολογιστή). Επειδή τα διανύσματα σ' αυτή τη μορφή είναι πολύ μεγάλα και αραιά εφαρμόζουμε τεχνικές μείωσης των διαστάσεων τους χρησιμοποιώντας τη μέθοδο Singular Value Decomposition (SVD), για τη διάσπαση πινάκων της αριθμητικής γραμμικής άλγεβρας. Τέλος τα τελικά κείμενα – διανύσματα χρησιμοποιούνται για την εκπαίδευση των GMM (Gaussian Mixture Models).

Στην παρούσα εργασία ο αλγόριθμος εφαρμόζεται σε 4 διαφορετικές συλλογές δεδομένων (Webkb, 20Newsgroups, Reuters, Cade), οι οποίες περιέχουν διαφορετικό αριθμό προ-κατηγοριοποιημένων κειμένων η κάθε μία, όπως επίσης διαφορετικό μέγεθος και περιεχόμενο. Η απόδοση του αλγορίθμου γίνεται σε σύγκριση με άλλες τεχνικές στην αιχμή της τεχνολογίας (state-of-art), όπως οι μέθοδοι Naïve Bayes, LSI, k-NN και Vector Method. Ο αλγόριθμος της εργασίας εμφανίζεται να έχει αποτελέσματα καλύτερα από τη πλειονότητα των μεθόδων κα συγκρίσιμα με τη μέθοδο Naïve Bayes που είναι και η μέθοδος με τα καλυτέρα αποτελέσματα συνολικά.



**Λέξεις Κλειδιά:** <<Κατηγοριοποίηση Κειμένων, Συνεχής Χώρος, Συνεχή Γλωσσικά Μοντέλα, GMM, Gaussian, SVD >>

# Abstract

The rapid spread of the Internet and the ever increasing availability of electronic material makes the usage of text classification techniques for finding the relevant information, even more essential.

The importance of text classification can be clearly seen from the plethora of applications where it is used to extract information from the continuous research done on the field to find new techniques and algorithms. This thesis examines the classification of texts depending on their topic using continuous mixture Gaussian distributions (Gaussian Mixture Models).

The implementation work starts from the pre-processing of texts in order to appear into a more compact form, followed by the creation stage of dictionary and term weighting of bringing a text version of the classical representation through words in vector format (more understandable to the computer).

Next step is to reduce the dimensions of the vectors using the Singular Value Decomposition (SVD), for the division of numerical linear algebra in tables. Finally, the final texts - vectors are being used for the training of GMM.

In this work, the algorithm applied in 4 different datasets (Webkb, 20Newsgroups, Reuters, Cade), which contains different number of pro-processing documents and different topic. The accuracy of the model compared with other state-of-the-art methods like Naïve Bayes, LSI, k-NN και Vector Method. Our algorithm seems to have the best results in the most cases, very close to Niave Bayes method which has totally the best result.



**Keywords:** << Text Categorization, Continuous Space, Continuous Space Language Models, GMM, Gaussian, SVD >>

# Πίνακας περιεχομένων

# Πίνακας σχημάτων

# Λίστα Πινάκων

# 1

# *Εισαγωγή*

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή στη περιοχή της Κατηγοριοποίησης Κειμένων, παρουσιάζεται η συνεισφορά της διπλωματικής εργασίας στο τομέα και περιγράφεται η οργάνωση υλοποίησής της.

## *1.1 Κατηγοριοποίηση Κειμένων*

Κατηγοριοποίηση κειμένων (Text Categorization), ονομάζεται η αυτόματη κατηγοριοποίηση ενός αριθμού κειμένων σε διαφορετικές κατηγορίες χρησιμοποιώντας ένα προκαθορισμένο σύνολο δεδομένων. Αν το κείμενο ανήκει σε μόνο μια κατηγορία, τότε αναφερόμαστε σε μονοσήμαντη κατηγοριοποίηση (single-label), ενώ εάν ανήκει σε περισσότερες από μία αναφερόμαστε σε πολυσήμαντη κατηγοριοποίηση (multi-label).

Η κατηγοριοποίηση κειμένων ερευνητικά σχετίζεται με τους τομείς της Ανάκτησης Πληροφορίας (Information Retrieval - IR) και της Μηχανικής Μάθησης (Machine Learning - ML) έχοντας λάβει ιδιαίτερη προσοχή τα τελευταία χρόνια και από τις δύο αυτές ερευνητικές περιοχές στον ακαδημαϊκό χώρο, καθώς επίσης εφαρμογές της χρησιμοποιούνται και στη βιομηχανία.

Όσον αφορά την Ανάκτηση Πληροφορίας, η κατηγοριοποίηση κειμένων χρησιμοποιεί εργαλεία που αναπτύχθηκαν από ερευνητές του IR αφού και τα δυο έχουν ως στόχο τη διαχείριση κειμένων και την εξαγωγή πληροφορίας. Ένα τέτοιο παράδειγμα είναι η αναζήτηση κειμένου, όπου ο στόχος είναι να βρεθεί το σύνολο των κειμένων (ή αποσπάσματα κειμένων) που είναι πιο σχετικά με μια συγκεκριμένη ερώτηση. Επίσης, τα κείμενα που χρησιμοποιούνται στη κατηγοριοποίηση κειμένων αντικαθιστούνται με διανύσματα ή πίνακες χρησιμοποιώντας όμοιες τεχνικές με το IR. Επιπλέον, τα κείμενα συγκρίνονται, και η μεταξύ τους ομοιότητα μετριέται κάνοντας χρήση τεχνικών που αναπτύχθηκαν αρχικά για Ανάκτηση Πληροφορίας. Τέλος η αξιολόγηση γίνεται συχνά χρησιμοποιώντας τα ίδια μέτρα απόδοσης με το IR.

Η κατηγοριοποίηση κειμένων παρουσιάζει επίσης ενδιαφέρον για τους ερευνητές της μηχανικής μάθησης, αφού χρησιμοποιούνται παρόμοιες τεχνικές και μεθοδολογίες. Πολλοί αλγόριθμοι που αναπτύχθηκαν στη μηχανική μάθηση έχουν χρησιμοποιηθεί για κατηγοριοποίηση κειμένων (π.χ. Support Vector Machines - SVM).

Στο τομέα της βιομηχανίας, η κατηγοριοποίηση κειμένων είναι σημαντική λόγω της μεγάλης ποσότητας των εγγράφων που πρέπει να τύχουν της κατάλληλης επεξεργασίας και να ταξινομηθούν. Οι τεχνικές της κατηγοριοποίησης κειμένων χρησιμοποιούνται σε μια μεγάλη ποικιλία εργασιών, όπως η εύρεση απαντήσεων σε παρόμοιες ερωτήσεις, η ταξινόμηση ειδήσεων με βάση το θέμα, η διαλογή Spam μηνυμάτων από τα νόμιμα μηνύματα ηλεκτρονικού ταχυδρομείου, η οργάνωση των εγγράφων σε διαφορετικούς φακέλους, κλπ.

Η αξιολόγηση των επιδόσεων γίνεται βάσει της απόδοσης (ποσοστό επιτυχών προβλέψεων του ταξινομητή), καθώς και της αποτελεσματικότητας του χρόνου κατηγοριοποίησης (πόσο καιρό παίρνει για να κατηγοριοποιηθεί ένα κείμενο).

Γενικά κάθε μέθοδος κατηγοριοποίησης κειμένων μπορεί να έχει τα πλεονεκτήματα και τα μειονεκτήματα της αντίστοιχα, σε σύγκριση με άλλες μεθόδους ανάλογα με το τι απαιτεί η κάθε εφαρμογή. Κυριότερο κριτήριο όπως αναφέρθηκε είναι η απόδοση, δηλαδή το ποσοστό των επιτυχών προβλέψεων σε άγνωστα κείμενα, παρόλα αυτά και σε αυτό το κομμάτι διαφορετικές τεχνικές μπορεί να είναι καλύτερες ανάλογα με το περιεχόμενο των κειμένων (δεν υπάρχει μέθοδος που να εγγυάται τα καλύτερα αποτελέσματα σε όλες τις περιπτώσεις). Επίσης μπορεί κάποια εφαρμογή να απαιτεί όσο το δυνατών καλύτερα

αποτελέσματα σε ελάχιστο χρόνο. Αυτό κυρίως αναζητείται σε live εφαρμογές, όπως εμφάνιση προτεινόμενων κείμενων (related), για παράδειγμα σε σελίδες αγορών ή ειδήσεων. Σε αυτή τη περίπτωση προτιμάται η μέθοδος που μπορεί να δώσει τα ταχύτερα αποτελέσματα, ακόμα και αν υστερεί (λίγο) σε απόδοση.

## *1.2 Αντικείμενο εργασίας*

Στη παρούσα εργασία σκοπός ήταν να εξεταστεί μια νέα προσέγγιση, σε σχέση με τις κλασσικές προσεγγίσεις, που αφορά τη προσπάθεια κατηγοριοποίησης κειμένων σε συνεχή χώρο με χρήση συνεχών κατανομών. Οι περισσότερες από τις υπάρχουσες εργασίες στον τομέα έχουν γίνει πάνω στο διακριτό χώρο χρησιμοποιώντας είτε τις λέξεις ως όρους (bag of words), μη λαμβάνοντας όμως υπόψη τη δομή του κειμένου, είτε χρησιμοποιώντας τα κείμενα ως διανύσματα με τους όρους να είναι οι λέξεις με κάποιο βάρος. Άλλες μέθοδοι που είναι στο συνεχή χώρο χρησιμοποιούν αποστάσεις αντί για κατανομές (Latent Semantic Indexing). Στην εργασία αυτή γίνεται χρήση Gaussian κατανομών και συγκεκριμένα Gaussian Mixture Models για τη κατηγοριοποίηση των κειμένων.

Αρχικά χρησιμοποιείται η διανυσματική μορφή αναπαράστασης των κειμένων, έπειτα ακολουθεί μείωση των διαστάσεων των χαρακτηριστικών διανυσμάτων των κειμένων και στη συνέχεια γίνεται χρήση της συνεχούς Gaussian κατανομής, πιο συγκεκριμένα χρησιμοποιούνται Gaussian Mixture Models (GMM). Με αυτό τον τρόπο τα κείμενα-διανύσματα χαρακτηρίζονται από τα στατιστικά χαρακτηριστικά τους (π.χ. μέση τιμή ,βάρη, διασπορά).

## *1.3 Οργάνωση κειμένου*

Η παρούσα διπλωματική εργασία οργανώνεται σε έξι κεφάλαια :

- Το κεφάλαιο **2** περιγράφει τεχνολογίες αιχμής του τομέα της κατηγοριοποίησης κειμένων (state-of-the-art) οι οποίες χρησιμοποιούνται για να γίνει σύγκριση της απόδοσης του συστήματος της παρούσας εργασίας.

- Το κεφάλαιο **3** περιγράφει τον τρόπο με τον οποίο γίνεται η μετατροπή του κειμένου από ένα σύνολο λέξεων που το περιγράφουν (μορφή κατανοητή από τον άνθρωπο), σε ένα διάνυσμα όρους που περιέχουν κάποιο βάρος (μορφή κατανοητή από τον υπολογιστή). Παρουσιάζονται δυο ευρέως χρησιμοποιούμενες τεχνικές κατανομής βαρών, η δυαδική και η tf-idf αναπαράσταση. Επιπλέον παρουσιάζεται η τεχνική Singular Value Decomposition (SVD) της αριθμητικής γραμμικής άλγεβρας για τη μείωση των διαστάσεων των χαρακτηριστικών διανυσμάτων των κειμένων.

- Το κεφάλαιο **4** παρουσιάζει τα δεδομένων (dataset) που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας και περιγράφει τον τρόπο που γίνεται η προ-επεξεργασία τους.

- Το κεφάλαιο **5** περιγράφει αναλυτικά όλα τα βήματα υλοποίησης της παρούσας εργασίας.

- Στο κεφάλαιο **6** γίνεται αξιολόγηση των αποτελεσμάτων του συστήματος σε σύγκριση με τα αποτελέσματα σχετικών εργασιών που παρουσιαστήκαν στο κεφάλαιο 2 .

- Τέλος, το κεφάλαιο **7** κάνει μια σύνοψη και παρουσιάζει τα συμπεράσματα για τα αποτελέσματα της εργασίας. Επίσης παρουσιάζει τις μελλοντικές επεκτάσεις που μπορούν να γίνουν σαν συνέχεια της υπάρχουσας εργασίας .

# 2

# *State-of-the-art*

Στο κεφάλαιο αυτό παρουσιάζονται τεχνολογίες αιχμής που έχουν υλοποιηθεί στο τομέα της κατηγοριοποίησης κειμένων, οι οποίες χρησιμοποιούνται σαν μέτρο σύγκρισης της απόδοσης του δικού μας συστήματος. Παρακάτω αναλύονται τέσσερις πολύ γνωστοί μέθοδοι κατηγοριοποίησης κειμένων, οι οποίοι χρησιμοποιούνται ευρέως με πολύ ικανοποιητικά αποτελέσματα. Όλες οι παρακάτω τεχνικές αφορούν μονοσήμαντη (single-label) κατηγοριοποίηση.

## *2.1 Naïve Bayes*

Η Naïve Bayes μέθοδος χρησιμοποιείται ευρέως στο τομέα της κατηγοριοποίησης κειμένων. Χρησιμοποιεί τις από κοινού πιθανότητες των λέξεων και των κλάσεων-κατηγοριών για να εκτιμήσει τη πιθανότητα να ανήκει ένα κείμενο σε μια κατηγορία.

Δεδομένου ενός συνόλου $r$ κειμένων $D = \{\overrightarrow{d_1}, \dots, \overrightarrow{d_r}\}$ , η κατηγοριοποίηση πάνω σε ένα σετ $q$ κλάσεων – κατηγοριών $C = \{c_1, \dots, c_q\}$ γίνεται υπολογίζοντας την παρακάτω πιθανότητα :

$$P\left(c_k \mid \overrightarrow{d_\iota}\right) = \frac{P(c_k)P\left(\overrightarrow{d_\iota} \mid c_k\right)}{P\left(\overrightarrow{d_\iota}\right)} \qquad 2.1.1$$

Σε αυτή την εξίσωση το $P\left(\overrightarrow{d_\iota}\right)$ είναι η πιθανότητα ένα τυχαία επιλεγμένο κείμενο με διανυσματική αναπαράσταση $\overrightarrow{d_\iota}$, και το $P(c_k)$ είναι η πιθανότητα ότι αυτό το τυχαία επιλεγμένο κείμενο-διάνυσμα να ανήκει στην κατηγορία $c_k$. Επειδή ο αριθμός των πιθανών κειμένων $\overrightarrow{d_\iota}$ είναι πολύ μεγάλος ο υπολογισμός του $P\left(\overrightarrow{d_\iota} \mid c_k\right)$ είναι προβληματικός.

Για να απλοποιηθεί ο υπολογισμός της $P\left(\overrightarrow{d_\iota} \mid c_k\right)$, θεωρείται ότι η πιθανότητα μιας δεδομένης λέξης – όρου είναι ανεξάρτητη από τους όρους που εμφανίζονται στο ίδιο κείμενο. Χρησιμοποιώντας τη παραπάνω απλοποίηση, είναι τώρα δυνατό να υπολογιστεί η πιθανότητα $P\left(\overrightarrow{d_\iota} \mid c_k\right)$ σαν γινόμενο των πιθανοτήτων για κάθε όρο που εμφανίζεται στο κείμενο.

Έτσι η πιθανότητα $P\left(\overrightarrow{d_\iota} \mid c_k\right)$ , όπου $\overrightarrow{d_\iota} = \{w_{1i}, \dots, w_{|T|i}\}$ , μπορεί πλέον να υπολογιστεί ως :

$$P\left(\overrightarrow{d_\iota} \mid c_k\right) = \prod_{j=1}^{|T|} P\left(w_{ji} \mid c_k\right) \qquad 2.1.2$$

όπου $|T|$ είναι ο συνολικός αριθμός των όρων που υπάρχουν στον σύνολο των κειμένων $D$ (το μέγεθος του λεξικού) , το $w_{ji}$ είναι το βάρος του όρου $t_j$ στο κείμενο $d_i$ . Το κείμενα ανήκουν στη κατηγορία όπου μεγιστοποιείται η πιθανότητα $argmax_{c_k}\left(P\left(c_k \mid \overrightarrow{d_\iota}\right)\right)$.

Παρόλο που η μέθοδος Naive Bayes φαίνεται απλουστευμένη παρόλα αυτά τα αποτελέσματα της είναι πολύ ανταγωνιστικά σε σχέση με άλλες πιο περιπλοκές μεθόδους.

## *2.2 Vector Method*

Ιστορικά η Vector Method [Salton and Lesk [9]; Salton, [8]] ήταν μια από τις πρώτες μεθόδους που εμφανίστηκαν στο τομέα της ανάκτησης πληροφορίας. Στην Vector Method τα κείμενα αναπαριστώνται ως ένα σύνολο από διανύσματα που έχουν βαρύτητα στους όρους τους, και τα οποία ανήκουν στη $n$ –διάστατο χώρο, όπου $n$ είναι ο συνολικός αριθμός των όρων του λεξικού.

Βασισμένο στο βάρος των όρων του, κάθε κείμενο μπορεί να κατηγοριοποιηθεί σε φθίνουσα σειρά βάσει της ομοιότητας του με τα αντίστοιχα κείμενα όλων των κατηγοριών. Η ομοιότητα ενός κειμένου $d_i$ της συλλογής με ένα άλλο νέο κείμενο $q$ υπολογίζεται ως το συνημίτονο της γωνίας (cosine similarity) που σχηματίζεται από τα διανύσματα που τα αναπαριστούν :

$$sim(\overrightarrow{d_i}, \overrightarrow{q}) = \frac{\overrightarrow{d_i} \cdot \vec{q}}{\|\overrightarrow{d_i}\| \times \|\overrightarrow{q}\|} \qquad 2.2.1$$

Η κατηγορία του κείμενου $q$ είναι η κατηγορία του κειμένου $d_i$ που είναι πιο όμοιο σε αυτό βάσει του παραπάνω τύπου ομοιότητας συνημίτονων.

## *2.3 K-Nearest Neighbors*

Η αρχική εφαρμογή της μεθόδου k-nearest Neighbors (k-NN) στη κατηγοριοποίηση κειμένων αναφέρεται από τον Masand και τους συνεργάτες του [Creecy et al [21] ; Masand et al. [20]]. Η βασική ιδέα της μεθόδου αυτής είναι ότι ένα κείμενο δε προσδιορίζεται μόνο σύμφωνα με τη κατηγορία στη οποία ανήκει το πιο κοντινό σε αυτό κείμενο (π.χ. Vector Method), αλλά συμφώνα με τις κατηγορίες που ανήκουν τα k πλησιεστέρα σε αυτό κείμενα. Έχοντας αυτό κατά νου, η Vector Method μπορεί να θεωρηθεί ως ένα παράδειγμα μεθόδου k-NN, όπου k = 1.

Στη παρούσα εργασία χρησιμοποιείται η k-NN μέθοδος με k=10, υπολογίζοντας την ομοιότητα των κειμένων όπως υπολογίζετε στη Vector-Method. Έτσι για κάθε νέο κείμενο q βρίσκονται τα k=10 πιο κοντινά του σύμφωνα με το cosine similarity (σχέση 2.32.2.1). Η τελική κατηγορία του κειμένου q είναι εκείνη στην οποία ανήκουν τα περισσότερα από τα k=10 κείμενα $d_i$.

## *2.4 Latent Semantic Indexing*

Το LSI αποτελεί μια μέθοδο κατηγοριοποίησης κειμένων που χρησιμοποιεί μεθόδους της γραμμικής άλγεβρας για να μειώσει το μέγεθος του πίνακα συχνοτήτων ώστε να γίνει πιο εύκολη η επεξεργασία του. Το μοντέλο LSI αναπτύχθηκε από τους [Furnas et al [14]; Deerwester et al [15]] .

Προκειμένου να εφαρμόσουμε το μοντέλο του Latent Semantic Indexing δημιουργείται αρχικά ένας πινάκας $A$ μεγέθους $t \times d$ , όπου ο $t$ αριθμός των όρων και ο $d$ αριθμός των κειμένων. Τα στοιχεία του πίνακα $A$ αποτελούν το βάρος κάθε όρου σε ένα συγκεκριμένο κείμενο. Έπειτα χρησιμοποιείται η τεχνική του Singular Value Decomposition (SVD) (εξηγείται αναλυτικά στο κεφάλαιο 3) για τη μείωση των διαστάσεων του πίνακα . Τέλος η κατηγορία ενός νέου κειμένου q βρίσκεται σύμφωνα με το cosine similarity .

### *2.4.1 Παράδειγμα*

Έστω η συλλογή 9 κειμένων τα οποία χαρακτηρίζονται με βάση το τίτλο τους

| Τίτλος | |
|---|---|
| $HCI_1$ | *Human* machine *interface* for *computer* applications |
| $HCI_2$ | A *survey* of *user* opinion of *computer system response time* |
| $HCI_3$ | The *user interface* of *EPS* management *system* |
| $HCI_4$ | *System* and *human system* engineering testing of *EPS* |
| $HCI_5$ | Relation of *user* perceived *response time* to error measurement |
| $GR_1$ | The generation of random,binary,unordered *trees* |
| $GR_2$ | The intersection *graph* of paths in trees |
| $GR_3$ | *Graph minors:* Widths of *trees* and well-quasi ordering |
| $GR_4$ | *Graph minors:*A *survey* |

**Figure 2.1 –** Παράδειγμα κειμένων χωρισμένα σε δυο κατηγορίες

Με πλάγια γράμματα χαρακτηρίζονται οι λέξεις που χρησιμοποιηθήκαν σαν λεξικό.

Στη συνέχεια με βάση αυτές τις λέξεις κατασκευάζεται ο πίνακας όρων-εγγράφων διαστάσεων $12 \times 9$. Παρακάτω παρουσιάζεται ο πίνακας όρων με tf-idf βάρη (τα tf-idf βάρη εξηγούνται αναλυτικά στο επόμενο κεφάλαιο).

**Tf-idf**

| $w_{ij}$ | $HCI_1$ | $HCI_2$ | $HCI_3$ | $HCI_4$ | $HCI_5$ | $GR_1$ | $GR_2$ | $GR_3$ | $GR_4$ |
|---|---|---|---|---|---|---|---|---|---|
| **Human** | 0.577 | 0 | 0 | 0.491 | 0 | 0 | 0 | 0 | 0 |
| **Interface** | 0.577 | 0 | 0.571 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Computer** | 0.577 | 0.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **User** | 0 | 0.324 | 0.417 | 0 | 0.458 | 0 | 0 | 0 | 0 |
| **System** | 0 | 0.324 | 0.417 | 0.718 | 0 | 0 | 0 | 0 | 0 |
| **Response** | 0 | 0.444 | 0 | 0 | 0.628 | 0 | 0 | 0 | 0 |
| **Time** | 0 | 0.444 | 0 | 0 | 0.628 | 0 | 0 | 0 | 0 |
| **EPS** | 0 | 0 | 0.571 | 0.491 | 0 | 0 | 0 | 0 | 0 |
| **Survey** | 0 | 0.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0.628 |
| **Trees** | 0 | 0 | 0 | 0 | 0 | 1.0 | 0.707 | 0.508 | 0 |
| **Graph** | 0 | 0 | 0 | 0 | 0 | 0 | 0.707 | 0.508 | 0.458 |
| **Minors** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.695 | 0.628 |

**Πίνακας 2.1 – Πίνακας όρων-εγγράφων**

Κάθε κείμενο χαρακτηρίζεται από ένα διάνυσμα 12 διαστάσεων με κάθε διάσταση να αντιστοιχεί σε ένα όρο . Επίσης όπως φαίνεται υπάρχουν 2 κατηγορίες κειμένων στη συλλογή (HCI , GR). Έστω τώρα ότι υπάρχει ένα κείμενο-query = «human computer interaction» ως είσοδος για κατηγοριοποίηση. Σε αυτό το σημείο θα περιμέναμε να υπάρχει ομοιότητα με χρήση cosine similarity με τα κείμενα της κατηγορίας HCI, όμως όπως φαίνεται στα πίνακα όρων-εγγράφων τα κείμενα $HCI_3$ και $HCI_5$ δεν περιέχουν τους όρους του query. Στις μεθόδους Vector Method και k-NN δηλαδή, τα κείμενα $HCI_3$ και $HCI_5$ θα λαμβάνονταν ως τελείως ανόμοια με το query παρόλο που ανήκουν στη ίδια κατηγορία.

Εφαρμόζοντας, όμως, τώρα τη μέθοδο LSI και SVD κρατώντας μόνο δυο διαστάσεις (2 πρώτες στήλες του $S$ , $k = 2$) κάθε όρος αναπαρίσταται με τον πολλαπλασιασμό της αντίστοιχης στήλης του πίνακα (όπως προκύπτει από τη μέθοδο SVD – κεφάλαιο 3) με την ιδιοτιμή $\sigma_1$ για τη x-συντεταγμένη και με τη δεύτερη ιδιοτιμή $\sigma_2$ για τη y-συντεταγμένη. Έτσι τοποθετούνται οι όροι στο δισδιάστατο χώρο. Για διαφορετικό $k$ ο νέος χώρος θα ήταν $k$-διάσταστος.

Το query τώρα, θα αντιστοιχισθεί στις διαστάσεις του LSI και στη συνέχεια θα επιστραφούν όλα τα κείμενα τα οποία βρίσκονται πιο κοντά σε αυτό σύμφωνα με τη συνάρτηση συνημιτόνου. Έτσι ακόμα και κείμενα τα οποία σχετίζονται με το θέμα αλλά δε μοιράζονται τους ίδιους όρους θα επιστραφούν. Όπως φαίνεται και στη παρακάτω εικόνα τα κείμενα έχουν σαφώς διαχωριστεί στο χώρο σε δύο ομάδες .

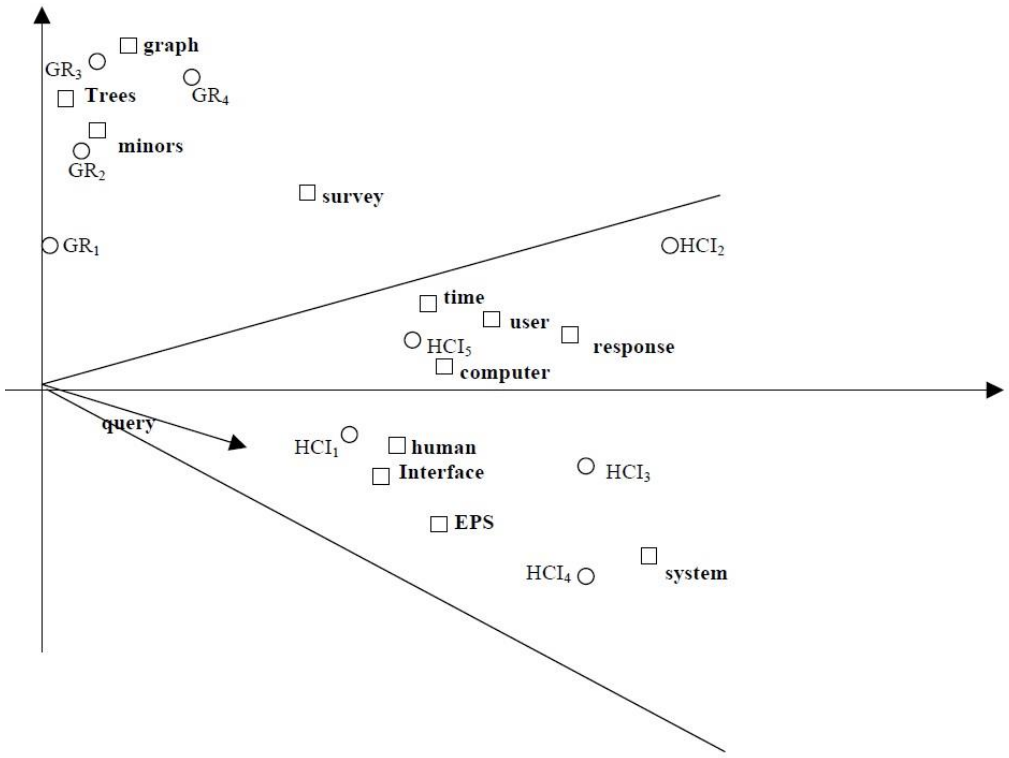


**Figure 2.2 –** Παράδειγμα κατηγοριοποίησης κειμένου με LSI

Η μέθοδος Naïve Bayes από την άλλη δε χρησιμοποίει διανυσματική αναπαράσταση οπότε για να βρεθεί κατηγορία στην οποία ανήκει το query = «human computer interaction» υπολογίζεται που μεγιστοποιείται η πιθανότητα $argmax_{c_k}\big(P(c_k \mid \text{query})\big)$, όπως εξηγήθηκε στην ενότητα 2.1, όπου $c_k$ είναι η κατηγορία HCI ή GR. Πιο αναλυτικά υπολογίζονται οι δυο παρακάτω πιθανότητες

$$P(HCI \mid' human\ computer\ interaction'\ ) = P(\ human|HCI\ ) * P(\ computer|HCI\ )$$
$$P(GR \mid' human\ computer\ interaction'\ ) = P(\ human|GR\ ) * P(\ computer|GR\ )$$

, η λέξη interaction δε λαμβάνεται υπόψη καθώς δεν υπάρχει στο λεξικό.

Έτσι σύμφωνα με τη μέθοδο naïve Bayes η κατηγορία στην οποία ανήκει το query θα είναι η κατηγορία όπου είναι μέγιστη μια από τις δυο παραπάνω πιθανότητες.

# 3

# *Εξαγωγή χαρακτηριστικών διανυσμάτων από κείμενο*

Ένα σημαντικό πρόβλημα στις διαδικασίες κατηγοριοποίησης κειμένων είναι η απεικόνιση των κειμένων. Οι ταξινομητές (classifiers) δεν μπορούν να αναγνωρίσουν τα γραπτά κείμενα στην αρχική τους μορφή. Είναι λοιπόν απαραίτητο να απεικονίσουμε τα κείμενα σε μορφή πιο κατανοητή για τους αλγορίθμους κατηγοριοποίησης, για το λόγο αυτό τα κείμενα αναπαριστώνται συνήθως ως διανύσματα. Επειδή όμως τα διανύσματα αποτελούνται από πάρα πολλούς όρους κάνοντας έτσι δύσκολη και χρονοβόρα τη διαδικασία κατηγοριοποίησης εφαρμόζονται τεχνικές μείωσης διαστάσεων των διανυσμάτων αυτών. Στη παρούσα εργασία χρησιμοποιείται η τεχνική Singular Value Decomposition.

## *3.1 Απεικόνιση ως διάνυσμα*

Ένα κείμενο αναπαρίσταται από όρους οι οποίοι μπορεί να είναι λέξεις, αριθμοί, σημεία στίξης κ.α. Στην κατηγοριοποίηση κειμένων η κύρια πληροφορία βρίσκεται κυρίως στις λέξεις, για το λόγο αυτό απαλείφονται οι υπόλοιποι όροι, όπως εξηγείται στο κεφάλαιο 4 αναλυτικά (4.3). Σε πολλές περιπτώσεις όμως, όπως για παράδειγμα οικονομικά κείμενα, ίσως είναι απαραίτητο να μην απαλειφθούν οι αριθμοί καθώς μπορεί να περιέχουν σημαντική πληροφορία για το κείμενο. Γενικά οι όροι που θα χρησιμοποιηθούν επιλέγονται βάσει του είδους των κειμένων. Στη παρούσα εργασία στα κείμενα της συλλογής (κεφάλαιο 4) οι αριθμοί δε παρέχουν κάποια σημαντική πληροφορία, οπότε απαλείφονται.

Ένα γραπτό κείμενο $d_j$ απεικονίζεται σαν ένα διάνυσμα $\vec{d_j} = (w_{1j}, \dots \dots, , w_{|T|j})$ κάθε στοιχείο του οποίου περιέχει ένα βάρος . Όπου :

- **T:** είναι το λεξικό (dictionary). Ένα σετ που περιέχει τους όρους που εμφανίζονται τουλάχιστον μια φορά σε τουλάχιστον ένα από τα έγγραφα $d_j$ της συλλογής .
- $\boldsymbol{w_{kj}}$ : είναι το βάρος (weight) του $k$ όρου στο j κείμενο (document). Το βάρος μπορούμε να πούμε ότι εκφράζει το βαθμό στον οποίο ο κάθε όρος συμβάλει στο σημασιολογικό περιεχόμενο του κάθε εγγράφου .

Υπάρχουν πολλοί τρόποι διαμόρφωσης των διανυσμάτων $d_j$ ανάλογα με το :

1. Τι θα χρησιμοποιήσουμε ως όρους του διανύσματος
2. Τις μεθόδους υπολογισμού των βαρών

Όσον αφορά το πρώτο (1), οι όροι μπορούν να αναφέρονται σε λέξεις, φράσεις συνδυασμό αυτών ή ακόμα σε εξαγόμενες, μετά από επεξεργασία, τιμές (για παράδειγμα αριθμοί που μπορεί να αναφέρονται σε χρήματα, ημερομηνίες κ.α. που έχουν επεξεργαστεί έτσι ώστε έχουν την ίδια μορφή (π.χ. 1 euro=1€ )). Μια συνηθισμένη επιλογή είναι να λάβουμε ως όρους τις λέξεις. Η τεχνική αυτή ονομάζεται bag of words, και παίρνει τιμές ανάλογα με το αν τα βάρη είναι δυαδικά (παίρνουν τιμές 0 ή 1) ή όχι. Τα μειονεκτήματα αυτής της τεχνικής είναι ότι χρησιμοποιώντας τις λέξεις ως όρους αγνοούμε τη δομή του κειμένου και τη σειρά των λέξεων σε αυτό, χάνοντας έτσι ένα μέρος της πληροφορίας . Παρόλα αυτά, σύμφωνα με μελέτες [17][18], παραμένει αυτός ο καλύτερος τρόπος απεικόνισης κειμένου.

Έχουν γίνει αρκετές προσπάθειες στον τομέα της κατηγοριοποίησης κειμένων να χρησιμοποιηθούν φράσεις ως όροι του διανύσματος, για παράδειγμα ακολουθίες N συνεχόμενων λέξεων (N-grams) [24][25], μέχρι σήμερα όμως τα πειραματικά αποτελέσματα προς αυτή την κατεύθυνση δεν είναι ενθαρρυντικά [18]. Είτε οι φράσεις χωριστούν με βάση τη σύνταξη, είτε με βάση στατιστικές μεθόδους (λέξεις που εμφανίζονται συχνά μαζί να αποτελούν φράση), τα αποτελέσματα εξακολουθούν να υστερούν συγκριτικά με τη χρήση λέξεων. Ο Lewis[18] απέδωσε αυτό το γεγονός στο ότι αν και οι φράσεις διατηρούν κάποια ανώτερα εννοιολογικά χαρακτηριστικά, οι λέξεις έχουν καλύτερα στατιστικά χαρακτηριστικά. Πάντως οι μελέτες πάνω στο phrase indexing συνεχίζονται καθώς είναι πιθανό να έχει να προσφέρει πολλά προς αυτή τη κατεύθυνση [7].

Πιο συγκεκριμένα με χρήση φράσεων :

- Έχουμε πολύ περισσότερους όρους και άρα αυξάνεται η πολυπλοκότητα των υπολογισμών.
- Έχουμε πολλούς συνώνυμους ή σχεδόν συνώνυμους όρους αφού, έστω και μία μόνο λέξη της φράσης να διαφέρει, οι όροι θεωρούνται ξεχωριστοί και ανεξάρτητοι.
- Έχουμε μικρότερη συχνότητα εμφάνισης των όρων στα έγγραφα (μικρότερη document frequency). Είναι αρκετά σπάνιο να επαναλαμβάνονται φράσεις σε μια συλλογή κειμένων από τον να επαναλαμβάνονται λέξεις .

Όσον αφορά τα βάρη των διανυσμάτων, αυτά συνήθως κυμαίνονται :

- μεταξύ του 0 και 1 (π.χ. παρακάτω – tf-idf)
- ή παίρνουν τιμές 0 ή 1 (δυαδικά βάρη) .

### *3.1.1 Δυαδικά βάρη*

Έστω ότι είναι το βάρος $w_{ij}$ ενός όρου $i$ που εμφανίζεται ή όχι στο κείμενο $d_j$ Ο γενικός τύπος κατανομής δυαδικών βάρών είναι :

$$w_{ij} = \begin{cases} 0, & αν\ ο\ i - οστός\ όρος\ δεν\ εμφανίζεται\ στο\ dj \\ 1, & αν\ ο\ i - οστός\ όρος\ εμφανίζεται\ στο\ dj \end{cases} \quad 3.1.1$$

Το πλεονέκτημα του δυαδικού διανυσματικού μοντέλου είναι ότι είναι πολύ εύκολο στην εφαρμογή του καθώς το μόνο που χρειάζεται είναι να βρεθεί αν κάποιος όρος εμφανίζεται σε ένα συγκεκριμένο κείμενο. Το μεγάλο μειονέκτημα του είναι ακριβώς το ότι τα βάρη είναι δυαδικά και δε δίνουν κάποια επιπλέον πληροφορία για τη συχνότητα εμφάνισης του όρου $w_{ij}$ στο κείμενο. Ένας σημαντικός όρος-λέξη για παράδειγμα σε ένα κείμενο μπορεί να εμφανίζεται περισσότερες από μια φορές. Με τη δυαδική αναπαράσταση και ο σημαντικός αυτός όρος και κάποιος που εμφανίζεται μια φόρα (λιγότερο σημαντικός) παίρνουν τη τιμή 1.
.

### *3.1.2 Term Frequency – Inverse Document Frequency*

Μια σημαντική ιδιότητα για τον χαρακτηρισμό των βαρών των όρων-λέξεων ενός συνόλου δεδομένων είναι η συχνότητα εμφάνισης του $k_i$ όρου σε κάθε κείμενο $d_j$. Διαισθητικά όσο πιο συχνά εμφανίζεται ένα όρος $k_i$ σε ένα κείμενο $d_j$, τόσο πιο καλή περιγραφή του $d_j$ αποτελεί ο $k_i$. Η συχνότητα εμφάνισης του όρου, ονομάζεται $tf$ (term frequency). Επίσης ένα μέτρο για το διαχωρισμό των δυο διαφορετικών συνόλων δεδομένων, αποτελεί η αντίστροφη συχνότητα εμφάνισης του $k_i$ στα κείμενα της συλλογής. Αν ο $k_i$ έχει μεγάλη συχνότητα εμφάνισης σε όλα τα κείμενα, δεν είναι πολύ χρήσιμος για να χαρακτηρίσει ένα κείμενο και άρα να διαχωρίσει μια ομάδα κειμένων από μια άλλη μες στη συλλογή . Η αντίστροφη συχνότητα εμφάνισης αναφέρεται ως $idf$ (inverse document frequency). Ο καλύτερος τρόπος υπολογισμού των βαρών προκύπτει από τον κατάλληλο συνδυασμό των δύο αυτών συχνοτήτων .

Έστω ο $N$ συνολικός αριθμός των κειμένων και $n_i$, ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος $k_i$ . Έστω $freq_{ij}$ αριθμός εμφανίσεων του όρου $k_i$ στο κείμενο $d_j$ . Τότε η κανονικοποιημένη συχνότητα $tf_{ij}$ του όρου $k_i$ στο $d_j$ δίνεται από τη σχέση :

$$tf_{ij} = \frac{freq_{ij}}{max_x\{freq_{xj}\}} \quad 3.1.2$$

όπου το $max_x\{freq_{xj}\}$ είναι ο αριθμός εμφανίσεων του όρου με τις περισσότερες εμφανίσεις μέσα στο κείμενο $d_j$. Εναλλακτικά υπάρχουν και άλλοι ορισμοί για το term frequency όπως για παράδειγμα ο λογαριθμικός ο οποίος μπορεί να οριστεί ως :

$$tf_{ij} = \log(freq_{ij} + 1) \quad 3.1.3$$

Στη παρούσα εργασία χρησιμοποιείται η κανονικοποιημένη μορφή, όπως φαίνεται στο τύπο 3.1.2.

Επιπλέον , έστω $idf_i$ η αντίστροφη συχνότητα εμφάνισης για τον $k_i$ που δίνεται από τον τύπο:

$$idf_i = \log\frac{N}{n_i} \quad 3.1.4$$

Τότε το καλύτερο γνωστό σχήμα υπολογισμού βαρών δίνεται από το γινόμενο :

$$w_{ij} = tf_{ij} \times idf_i \quad 3.1.5$$

### *3.1.3 Παράδειγμα Binary και tf-idf*

Έστω μια συλλογή 9 κειμένων τα οποία χαρακτηρίζονται με βάση το τίτλο τους ,όπου με πλάγια γράμματα χαρακτηρίζονται οι λέξεις που χρησιμοποιηθήκαν σαν λεξικό

| Τίτλος | |
|---|---|
| $HCI_1$ | *Human* machine *interface* for *computer* applications |
| $HCI_2$ | A *survey* of *user* opinion of *computer system response time* |
| $HCI_3$ | The *user interface* of *EPS* management *system* |
| $HCI_4$ | *System* and *human system* engineering testing of *EPS* |
| $HCI_5$ | Relation of *user* perceived *response time* to error measurement |
| $GR_1$ | The generation of random,binary,unordered *trees* |
| $GR_2$ | The intersection *graph* of paths in trees |
| $GR_3$ | *Graph minors:* Widths of *trees* and well-quasi ordering |
| $GR_4$ | *Graph minors:*A *survey* |

**Figure 3.1 –** Παράδειγμα κειμένων χωρισμένα σε δυο κατηγορίες

Η αναπαράσταση της παραπάνω συλλογής με δυαδικά βάρη είναι ένας πίνακας που έχει τις τιμές που φαίνονται στο παρακάτω πίνακα :

**Binary term weight**

| $w_{ij}$ | $HCI_1$ | $HCI_2$ | $HCI_3$ | $HCI_4$ | $HCI_5$ | $GR_1$ | $GR_2$ | $GR_3$ | $GR_4$ |
|---|---|---|---|---|---|---|---|---|---|
| **Human** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Interface** | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Computer** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **User** | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **System** | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Response** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **Time** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **EPS** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Survey** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Trees** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| **Graph** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| **Minors** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

**Πίνακας 3.1 –** Δυαδική κατανομή βαρών

Όπως φαίνεται οι τιμές είναι 1 μόνο εκεί που εμφανίζεται ένας όρος-λέξη σε ένα κείμενο.

Η αναπαράσταση της παραπάνω συλλογής με tf-idf βάρη είναι ένας πίνακας που έχει τις τιμές που φαίνονται στο παρακάτω πίνακα :

**Tf-idf**

| $w_{ij}$ | $HCI_1$ | $HCI_2$ | $HCI_3$ | $HCI_4$ | $HCI_5$ | $GR_1$ | $GR_2$ | $GR_3$ | $GR_4$ |
|---|---|---|---|---|---|---|---|---|---|
| **Human** | 0.577 | 0 | 0 | 0.491 | 0 | 0 | 0 | 0 | 0 |
| **Interface** | 0.577 | 0 | 0.571 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Computer** | 0.577 | 0.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **User** | 0 | 0.324 | 0.417 | 0 | 0.458 | 0 | 0 | 0 | 0 |
| **System** | 0 | 0.324 | 0.417 | 0.718 | 0 | 0 | 0 | 0 | 0 |
| **Response** | 0 | 0.444 | 0 | 0 | 0.628 | 0 | 0 | 0 | 0 |
| **Time** | 0 | 0.444 | 0 | 0 | 0.628 | 0 | 0 | 0 | 0 |
| **EPS** | 0 | 0 | 0.571 | 0.491 | 0 | 0 | 0 | 0 | 0 |
| **Survey** | 0 | 0.444 | 0 | 0 | 0 | 0 | 0 | 0 | 0.628 |
| **Trees** | 0 | 0 | 0 | 0 | 0 | 1.0 | 0.707 | 0.508 | 0 |
| **Graph** | 0 | 0 | 0 | 0 | 0 | 0 | 0.707 | 0.508 | 0.458 |
| **Minors** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.695 | 0.628 |

**Πίνακας 3.2 –** Tf-idf κατανομή βαρών

Σε αυτό το πίνακα , όπως φαίνεται , οι όροι έχουν διαφορετικό βάρος σε κάθε κείμενο ανάλογα με το πόσο συχνά εμφανίζονται . Η πληροφορία που δίνει αυτός ο πίνακας για τι μορφή των κειμένων είναι σαφώς περισσότερη σε σχέση με τη δυαδική αναπαράσταση .

## *3.2 Προβολή σε χώρο μικρότερης διάστασης*

Ένα σημαντικό πρόβλημα στις διαδικασίες κατηγοριοποίησης κειμένων είναι ο μεγάλος αριθμός των όρων που εμφανίζονται στα κείμενα και ο οποίος οδηγεί σε χαρακτηριστικά διανύσματα μεγάλων διαστάσεων. Στην πράξη τα χαρακτηριστικά διανύσματα μπορεί να αποτελούνται από δεκάδες ή και εκατοντάδες χιλιάδες όρους ακόμα και για κείμενα σχετικά μικρού μεγέθους, κάνοντας έτσι απαγορευτική της εφαρμογή των αλγορίθμων εκπαίδευσης. Επίσης τα διανύσματα που δημιουργούνται είναι αραιά (sparse), δηλαδή έχουν τους περισσότερους όρους τους ίσους με 0, αφού οι περισσότεροι όροι του λεξικού εμφανίζονται μόνο σε πολύ λίγα κείμενα. Καταλήγουμε λοιπόν να έχουμε πολύ μεγάλους sparse πίνακες και άρα είναι σημαντικό να χρησιμοποιήσουμε αλγορίθμους μείωσης της διάστασης των διανυσμάτων διατηρώντας παράλληλα τη σημαντική πληροφορία που περιλαμβάνεται σ' αυτά. Κυρίαρχη μέθοδος διάσπασης πίνακα, που χρησιμοποιείται και στην παρούσα εργασία, είναι το Singular Value Decomposition (SVD) της αριθμητικής γραμμικής άλγεβρας.

### *3.2.1 Πλεονεκτήματα χρήσης SVD*

Τα κύρια πλεονεκτήματα της χρήσης του SVD για τη μείωση των διαστάσεων, όπως παρουσιαστήκαν από τον Deerwester et al [15] χωρίζονται στις εξής κατηγορίες :

- Συνωνυμία
- Πολυσημία

***<u>Συνωνυμία</u>***

Συνωνυμία αναφέρεται στο γεγονός ότι η ίδια έννοια μπορεί να περιγραφεί χρησιμοποιώντας διαφορετικούς όρους. Οι παραδοσιακές μέθοδοι κατηγοριοποίησης έχουν πρόβλημα να ανακαλύψουν τα έγγραφα του ίδιου θέματος που χρησιμοποιούν διαφορετικό λεξιλόγιο. Τα έγγραφα που σχετίζονται μεταξύ τους είναι πιθανόν να εκπροσωπούνται από ένα παρόμοιο συνδυασμό όρων.

***Πολυσημία***

Η πολυσημία περιγράφει λέξεις που έχουν περισσότερες από μία έννοιες. Μεγάλος αριθμός τέτοιων λέξεων σε ένα έγγραφο μπορεί να μειώσει την ακρίβεια της κατηγοριοποίησης σημαντικά. Με τη χρήση SVD και τη μειωμένη αναπαράσταση, ελπίζει κανείς να αφαιρέσει όρους που χαρακτηρίζονται ως θόρυβος, οι οποίοι θα μπορούσαν να χαρακτηριστούν ως σπάνιοι και λιγότερο σημαντικοί.

### *3.2.2 Πίνακας όρων – κειμένων*

Πριν αναλυθεί η τεχνική SVD χρήσιμο θα είναι να γίνει μια περιγραφή του είναι ο πίνακας όρων-κειμένων (term-document matrix) ο οποίος είναι ο πίνακας πάνω στον οποίο εφαρμόζεται το SVD για να γίνει η μείωση των διαστάσεων των χαρακτηριστικών διανυσμάτων των κειμένων .

Έστω ότι έχουμε ένα σύνολο όρων τα οποία εμφανίζονται σε ένα σύνολο κειμένων τότε ο πίνακας term-document θα περιέχει στις στήλες του τα σχετιζόμενα κείμενα και στις γραμμές του τους αντίστοιχους όρους που εμφανίζονται σε αυτά. Για παράδειγμα έστω τα δύο παρακάτω (μικρού μήκους) κείμενα :

- D1 = “I like databases”
- D2 = “I hate databases”

Τότε ο πίνακας term-document για τα δύο αυτά κείμενα θα είναι της μορφής :

| | **D1** | **D2** |
|---|---|---|
| I | 1 | 1 |
| like | 1 | 0 |
| hate | 0 | 1 |
| databases | 1 | 1 |

**Πίνακας 3.3 –** term-document matrix

*σε αυτό το παράδειγμα χρησιμοποιήθηκαν δυαδικά βάρη.*

### *3.2.3 Singular Value Decomposition (SVD)*

Η Singular Value Decomposition (SVD) είναι μια από τις πιο σημαντικές μεθόδους διάσπασης πίνακα της αριθμητικής γραμμικής άλγεβρας. Στην παρούσα εργασία χρησιμοποιούμε SVD για να μειώσουμε τις διαστάσεις του διανυσματικού χώρου και να αφαιρέσουμε τις αρνητικές επιδράσεις της συνωνυμίας και της πολυσημίας, αποτυπώνοντας ταυτόχρονα τις πολύτιμες σημασιολογικές σχέσεις μεταξύ των όρων. Μια σημαντική ιδιότητα του νέου χώρου είναι ότι δύο λέξεις των οποίων οι απεικονίσεις είναι παρόμοιες έχουν την τάση να εμφανίζονται σε έγγραφα με παραπλήσιο περιεχόμενο, ανεξάρτητα από το αν πλαισιώνονται από τις ίδιες λέξεις στα έγγραφα αυτά. Αντιστρόφως δυο έγγραφα με κοντινές απεικονίσεις έχουν παρόμοιο σημασιολογικό περιεχόμενο ακόμα και χωρίς να έχουν ίδιες λέξεις.

Από τα training documents κατασκευάζεται ο πίνακας όρων – κειμένων (term document matrix) $A(m \times n)$, όπου $n$ ο αριθμός των εγγράφων της συλλογής και $m$ ο αριθμός των διαφορετικών όρων που εμφανίζονται στα έγγραφα αυτά. Υποθέτουμε ότι $m \geq n$ και ότι ο βαθμός (rank) του πίνακα $A$ είναι $r$. Η Singular Value Decomposition (SVD) ενός πινάκα είναι η διάσπαση του σε τρεις νέους πίνακες

$$A = USV^T \qquad 3.2.1$$

Όπου :

- Ο πίνακας $U$ είναι ένας ορθογώνιος πίνακας $(m \times n)$ , οι στήλες του οποίου είναι τα αριστερά ιδιοδιανύσματα του πίνακα $A$ . Ισχύει ότι $U^T U = I_n$
- Ο πίνακας $V$ είναι ένας ορθογώνιος πίνακας $(m \times n)$ , οι στήλες του οποίου είναι τα δεξιά ιδιοδιανύσματα του πίνακα $A$ . Ισχύει ότι $V^T V = I_n$
- Ο πίνακας $S$ είναι ένας διαγώνιος πίνακας $(n \times n)$, του οποίου τα διαγώνια στοιχεία είναι οι ιδιοτιμές του πίνακα $A$. Ισχύει ότι $S = diag(\sigma_1, \ldots., \sigma_n)$ όπου $\sigma_i > 0$ όταν $1 < i < r$ και $\sigma_i = 0$ όταν $i > r$

Για να μειώσουμε τις διαστάσεις χρησιμοποιούμε μια ειδική μορφή της SVD, την truncated SVD. Επιλέγουμε τις $k$ μεγαλύτερες ιδιοτιμές και τα αντίστοιχα αριστερά και δεξιά ιδιοδιανύσματα και παράγουμε τον πίνακα $A_k(m \times n)$. Αυτός είναι ο πίνακας βαθμού $k$ που προσεγγίζει καλύτερα τον αρχικό πίνακα $A$ και δίνεται από τη σχέση :

$$A_k = U_k S_k V_k^T \qquad 3.2.2$$

όπου:

- Ο $U_k$ αποτελείται από τις $k$ πρώτες στήλες του πίνακα $U$ και έχει διαστάσεις $(m \times k)$ .
- Ο $V_k^T$ αποτελείται από τις $k$ πρώτες σειρές του πίνακα $V^T$και έχει διαστάσεις $(k \times n)$.
- Ο $S$ είναι ο διαγώνιος πινάκας $S_k = diag(\sigma_1, \ldots., \sigma_k)$ που αποτελείται από τις $k$ πρώτες διαγώνιες τιμές του πίνακα $S$ και έχει διαστάσεις $(k \times k)$.

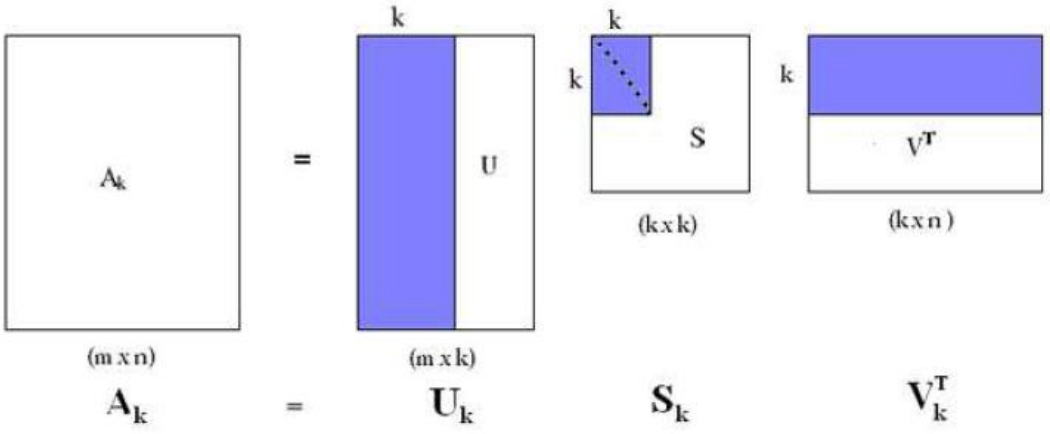


**Figure 3.2** – Διάσπαση πινάκα SVD

Ο πίνακας $A_k$ συλλαμβάνει τις σημαντικότερες κρυμμένες δομές των συσχετίσεων όρων-εγγράφων, αγνοώντας τον θόρυβο που οφείλεται στην επιλογή των λέξεων. Χρησιμοποιώντας αυτούς τους πίνακες, μπορούμε να κάνουμε το μετασχηματισμό της απεικόνισης των εγγραφών στο νέο διανυσματικό χώρο. Συγκεκριμένα, το αρχικό διάνυσμα $\vec{d}$ ενός εγγράφου μετασχηματίζεται με βάση τη σχέση :

$$\overrightarrow{d_k} = \overrightarrow{d^T} U_k S_k^{-1} \qquad 3.2.3$$

Όπου $\overrightarrow{d_k}$ είναι η νέα απεικόνιση του εγγράφου, διαστάσεων $(k \times 1)$ .

# 4

# *Δεδομένα και Προ-επεξεργασία*

Τα σύνολα δεδομένων (datasets) είναι μια συλλογή από προ-κατηγοριοποιημένα κείμενα διαφόρων κατηγοριών τα οποία είναι απαραίτητα για να αναπτυχθεί και να αξιολογηθεί το σύστημα .

## *4.1 Λεπτομέρειες Datasets*

Κάθε σύνολο δεδομένων (Dataset) περιλαμβάνει ένα σύνολο από κείμενα – έγγραφα , μαζί με την κατηγορία στην οποία ανήκει το κάθε κείμενο – έγγραφο. Ένα ποσοστό από αυτά χρησιμοποιείται για την εκπαίδευση (training) του συστήματος το υπόλοιπο ποσοστό χρησιμοποιείται για να αξιολογηθεί (testing) πώς συμπεριφέρεται το σύστημα όταν δίνεται ένα νέο κείμενο για αξιολόγηση.

Στο πρώτο βήμα (training phase) , τα κείμενα που είναι για εκπαίδευση (training documents) χρησιμοποιούνται για την εκπαίδευση, επιτρέποντας στο μοντέλο να "μάθει" από αυτά . Έπειτα ακολουθεί το βήμα της αξιολόγησης (testing phase) , το οποίο χρησιμοποιεί τα υπόλοιπα κείμενα (testing documents) για την αξιολόγηση τους συστήματος με σκοπό να φανεί πόσο καλά συμπεριφέρεται το σύστημα όταν κατηγοριοποιούνται άγνωστα κείμενα .

Προκειμένου να είναι δίκαιη η σύγκριση με άλλες μεθόδους κατηγοριοποίησης κειμένων είναι επιθυμητό τα datasets να έχουν δοκιμαστεί σε ισοδύναμες ρυθμίσεις. Με βάση αυτό το σκοπό τα dataset που χρησιμοποιούνται έχουν δημιουργηθεί και δημοσιοποιούνται σε μια τυπική διάσπαση train / test documents έτσι ώστε τα αποτελέσματα που προκύπτουν από τα διαφορετικά συστήματα κατηγοριοποίησης να είναι άμεσα και σωστά συγκρίσιμα μεταξύ τους.

### *4.1.1 The 20-Newsgroups Collection*

Η συλλογή δεδομένων 20-Newsgroups είναι ένα σύνολο περίπου 20.000 εγγράφων, κατανεμημένων (σχεδόν) ομοιόμορφα σε 20 διαφορετικές κατηγορίες. Τα δεδομένα είναι τυπικά mail και έτσι έχουν συμπεριλαμβανομένων γραμμές θέματος, αρχεία, και ανέφερα τμημάτων άλλων αρχείων.

Η συλλογή 20-Newsgroups κατεβάστηκε (downloaded) από τη σελίδα του Jason Rennie και χρησιμοποιήθηκε η έκδοση όπου τα κείμενα είναι χωρισμένα "ανά ημερομηνία" , επειδή τα κείμενα είναι ήδη χωρισμένα σε train και test . Όπως αναφέρθηκε η συλλογή περιλαμβάνει 20.000 κειμένων –mail τα οποία χωρίζονται σχεδόν ομοιόμορφα σε 20 διαφορετικές κατηγορίες όπως παρουσιάζεται στο παρακάτω πίνακα .

**20 Newsgroups**

| Class | # train docs | # test docs | Total # docs |
|---|---|---|---|
| alt.atheism | 480 | 319 | 799 |
| comp.graphics | 584 | 389 | 973 |
| comp.os.ms-windows.misc | 572 | 394 | 966 |
| comp.sys.ibm.pc.hardware | 590 | 392 | 982 |
| comp.sys.mac.hardware | 578 | 385 | 963 |
| comp.windows.x | 593 | 392 | 985 |
| misc.forsale | 585 | 390 | 975 |
| rec.autos | 594 | 395 | 989 |
| rec.motorcycles | 598 | 398 | 996 |
| rec.sport.baseball | 597 | 397 | 994 |
| rec.sport.hockey | 600 | 399 | 999 |
| sci.crypt | 595 | 396 | 991 |
| sci.electronics | 591 | 393 | 984 |
| sci.med | 594 | 396 | 990 |
| sci.space | 593 | 394 | 987 |
| soc.religion.christian | 598 | 398 | 996 |
| talk.politics.guns | 545 | 364 | 909 |
| talk.politics.mideast | 564 | 376 | 940 |
| talk.politics.misc | 465 | 310 | 775 |
| talk.religion.misc | 377 | 251 | 628 |
| **Total** | **11293** | **7528** | **18821** |

**Πίνακας 4.1** – Αναλυτικός πίνακας για το 20Newsgroups Dataset

### *4.1.2 The Reuters-21578 Collection*

Η συλλογή δεδομένων Reuters-21578 περιλαμβάνει κείμενα τα οποία εμφανίστηκαν στο ειδησεογραφικό πρακτορείο Reuters και έχει κατηγοριοποιηθεί χειροκίνητα από το προσωπικό του Reuters Ltd. Η συλλογή αυτή είναι πολύ ασύμμετρη με την κατανομή των κειμένων να είναι πολύ άνιση σε αυτή τη συλλογή .

Η συλλογή Reuters-21578 κατεβάστηκε (downloaded) από τη σελίδα του David Lewis. Τα κείμενα όπως προαναφέρθηκε είναι κείμενα του ειδησεογραφικού πρακτορείου Reuters του 1987 . Εξαιτίας του γεγονότος ότι η κατανομή των κειμένων είναι πολύ ασύμμετρη , δυο υποκατηγορίες χρησιμοποιούνται συνήθως για την κατηγοριοποίηση κειμένων

- R10 – Το σύνολο των κειμένων ανήκει στις 10 κλάσεις - κατηγορίες με τον υψηλότερο αριθμό παραδειγμάτων εκπαίδευσης (training).

- R90 – Το σύνολο των δεδομένων ανήκει στις 90 κλάσεις – κατηγορίες η οποίες περιέχουν τουλάχιστον από ένα παράδειγμα για εκπαίδευση και αξιολόγηση (training – testing).

Εκτός του ότι η κατανομή των κειμένων είναι ασύμμετρη , πολλά από τα κείμενα αυτής της συλλογής έχουν χωριστεί – κατηγοριοποιηθεί έτσι ώστε πολλά από αυτά δεν έχουν καν κατηγορία (topic) ενώ αντίθετα πολλά από αυτά ανήκουν σε παραπάνω από μια κατηγορία. Ο πίνακας 4.2 δείχνει τη κατανομή των κειμένων της συλλογής βάσει του αριθμού των κατηγοριών (topics).

| Reuters 21578 | | | | |
|---|---|---|---|---|
| **# Topics** | **# train docs** | **# test docs** | **# other** | **Total # docs** |
| 0 | 1828 | 280 | 8103 | 10211 |
| 1 | 6552 | 2581 | 361 | 9494 |
| 2 | 890 | 309 | 135 | 1334 |
| 3 | 191 | 64 | 55 | 310 |
| 4 | 62 | 32 | 10 | 104 |
| 5 | 39 | 14 | 8 | 61 |
| 6 | 21 | 6 | 3 | 30 |
| 7 | 7 | 4 | 0 | 11 |
| 8 | 4 | 2 | 0 | 6 |
| 9 | 4 | 2 | 0 | 6 |
| 10 | 3 | 1 | 0 | 4 |
| 11 | 0 | 1 | 1 | 2 |
| 12 | 1 | 1 | 0 | 2 |
| 13 | 0 | 0 | 0 | 0 |
| 14 | 0 | 2 | 0 | 2 |
| 15 | 0 | 0 | 0 | 0 |
| 16 | 1 | 0 | 0 | 1 |

**Πίνακας 4.2 –** Αναλυτικός πίνακας κειμένων – κατηγοριών για το Reuters Dataset

Επειδή ο σκοπός της συγκεκριμένης εργασίας είναι η μοναδική κατηγοριοποίηση κειμένων (single-label text categorization) δηλαδή κάθε κείμενο κατηγοριοποιείται σε μία και μόνο κατηγορία, όλα τα κείμενα που ανήκουν σε λιγότερη από μια ή αντίστοιχα σε περισσότερες από μια κατηγορίες έχουν εξαλειφθεί. Μέτα από αυτή τη διαγραφή ορισμένων κειμένων οι μερικές από τις συλλογές των R10 και R90 έχουν μείνει χωρίς training και testing documents. Λαμβάνοντας υπόψη μόνο τα κείμενα τα οποία έχουν μόνο μία κατηγορία και τις κλάσεις που έχουν τουλάχιστον ένα training και ένα testing κείμενο, η συλλογή R10 περιέχει πλέον 8 αντί για 10 κλάσεις , ενώ αντίστοιχα η συλλογή R90 περιέχει πλέον 52 αντί για 90 κλάσεις. Οι δυο νέες συλλογές ονομάζονται R8 και R52 αντίστοιχα. Στην παρούσα εργασία χρησιμοποιείται η συλλογή R8 καθώς η R52 περιέχει πολλές κλάσεις με πολύ λίγα κείμενα για εκπαίδευση και αξιολόγηση και εξακολουθεί να είναι αρκετά ασύμμετρη. Από τη μετάβαση της R10 στη R8 (που χρησιμοποιείται στην εργασία) οι κλάσεις “com” και “wear” , οι οποίες συνδέονται στενά με τη κλάση “grain”, με τη τελευταία να χάνει πολλά από τα κείμενα της. Η κατανομή των κειμένων εκπαίδευσης και αξιολόγησης για κάθε κλάση της R8 παρουσιάζεται στον πίνακα 4.3.

| R8 | | | |
|---|---|---|---|
| **Class** | **# train docs** | **# test docs** | **Total # docs** |
| acq | 1596 | 696 | 2292 |
| crude | 253 | 121 | 374 |
| earn | 2840 | 1083 | 3923 |
| grain | 41 | 10 | 51 |
| interest | 190 | 81 | 271 |
| money-fx | 206 | 87 | 293 |
| ship | 108 | 36 | 144 |
| trade | 251 | 75 | 326 |
| **Total** | **5485** | **2189** | **7674** |

**Πίνακας 4.3** – Αναλυτικός πίνακας για το Reuters Dataset

### *4.1.3 The Webkb Collection*

Η συλλογή δεδομένων Webkb περιλαμβάνει ιστοσελίδες από τμήματα επιστήμης υπολογιστών οι οποίες συλλέχτηκαν από το theWorldWide Knowledge Base (Web->Kb) project της ομάδας CMU text learning το 1997. Για κάθε διαφορετική κατηγορία , η συλλογή περιλαμβάνει σελίδες από τέσσερα πανεπιστήμια (Cornell, Texas, Washington and Wisconsin) , και διάφορες άλλες σελίδες που συλλέχτηκαν από διάφορα άλλα πανεπιστήμια .

Η συλλογή Webkb (η οποία επίσης αναφέρεται ως 4 University Data Set ) κατεβάστηκε (downloaded) τη σελίδα του World Wide Knowledge Base (webkb). Οι ιστοσελίδες που περιλαμβάνει η συλλογή έχουν κατηγοριοποιηθεί χειροκίνητα (manually) σε εφτά κατηγορίες και περιλαμβάνουν σελίδες από τα τέσσερα πανεπιστήμια , συν κάποιες άλλες σελίδες που συλλέχθηκαν από διάφορα άλλα πανεπιστήμια . Σύμφωνα με το Nigam et Al [Nigam et al.[22] ] , οι κλάσεις "Department" και "Staff" της συλλογής εξαλείφονται επειδή υπάρχουν πολύ λίγες σελίδες για κάθε πανεπιστήμιο. Επίσης η κλάση που περιέχει τις σελίδες από άλλα πανεπιστήμια επίσης εξαλείφεται καθώς οι σελίδες είναι πολύ διαφορετικές μεταξύ των παραδειγμάτων της κλάσεις. Η κατανομή των κειμένων για τη συλλογή Webkb παρουσιάζεται στον πίνακα 4.4.

| WebKB | | | |
|---|---|---|---|
| **Class** | **# train docs** | **# test docs** | **Total # docs** |
| project | 336 | 168 | 504 |
| course | 620 | 310 | 930 |
| faculty | 750 | 374 | 1124 |
| student | 1097 | 544 | 1641 |
| **Total** | **2803** | **1396** | **4199** |

**Πίνακας 4.4** – Αναλυτικός πίνακας για το Webkb Dataset

### *4.1.4 The Cade Collection*

Η συλλογή δεδομένων Cade περιλαμβάνει κείμενα τα οποία αναφέρονται σε ιστοσελίδες που προέρχονται από το CADE Web Directory, η οποία αφορά σε ανθρώπινες ιστοσελίδες οι οποίες έχουν κατηγοριοποιηθεί από εδικούς.

Μία προ-επεξεργασμένη έκδοση αυτής της συλλογής είναι διαθέσιμη από τον Marco Cristo του Universidade Federal de Minas Gerais της Βραζιλίας. Τα κείμενα είναι χωρισμένα σε 12 κατηγορίες. Τα δυο τρίτα των κειμένων έχουν επιλεγεί για εκπαίδευση (training) και τα υπόλοιπα για αξιολόγηση (testing). Η κατανομή των κειμένων για τη συλλογή Cade παρουσιάζεται στον πίνακα 4.5.

| Cade12 | | | |
|---|---|---|---|
| **Class** | **# train docs** | **# test docs** | **Total # docs** |
| 01--servicos | 5627 | 2846 | 8473 |
| 02--sociedade | 4935 | 2428 | 7363 |
| 03--lazer | 3698 | 1892 | 5590 |
| 04--informatica | 2983 | 1536 | 4519 |
| 05--saude | 2118 | 1053 | 3171 |
| 06--educacao | 1912 | 944 | 2856 |
| 07--internet | 1585 | 796 | 2381 |
| 08--cultura | 1494 | 643 | 2137 |
| 09--esportes | 1277 | 630 | 1907 |
| 10--noticias | 701 | 381 | 1082 |
| 11--ciencias | 569 | 310 | 879 |
| 12--compras-online | 423 | 202 | 625 |
| **Total** | **27322** | **13661** | **40983** |

**Πίνακας 4.5** – Αναλυτικός πίνακας για το Cade Dataset

## *4.2 Στατιστικά Συνόλων Δεδομένων (dataset)*

Στην υπάρχουσα εργασία χρησιμοποιούνται τα 4 σύνολα δεδομένων (datasets) με διαφορετικά περιεχόμενα που εξηγήθηκαν λεπτομερώς παραπάνω ( 20Newsgroups , Webkb , Reuters8, Cade).

Τα datasets έχουν διαφορετικά μεγέθη , με το μικρότερο (Webkb) να έχει 4.199 κείμενα και το μεγαλύτερο (Cade) 40.983 κείμενα. Επίσης το 20newsgroup έχει ομαλή κατανομή στον αριθμό των κειμένων για κάθε κατηγορία σε αντίθεση με τα άλλα τρία στα οποία η κατανομή των κειμένων είναι ασύμμετρη .

Για παράδειγμα το R8 έχει τη μικρότερη κλάση να περιέχει μόλις 51 κείμενα ενώ αντίθετα η μεγαλύτερή του κλάση περιέχει 3.923 κείμενα . Στο Πίνακας 4.6 παρουσιάζονται ο αριθμός των κειμένων και για τις 4 συλλογές , ο αριθμός των κειμένων εκπαίδευσης (training documents) , ο αριθμός των κειμένων αξιολόγησης (testing documents) ,ο συνολικός αριθμός των κειμένων κάθε συλλογής ,ο αριθμός των κειμένων της κλάσης με τα λιγότερα κείμενα για κάθε συλλογή και αντίστοιχα ο αριθμός των κειμένων της κλάσης με τα περισσότερα κείμενα για κάθε συλλογή. Επιπλέων οι συλλογές 20Newsgroup , Reuters και Webkb περιέχουν αγγλικά κείμενα σε αντίθεση με τη Cade που περιέχει κείμενα στα πορτογαλικά .

| *Dataset* | *Train Docs* | *Testn Docs* | *Total Docs* | *Smallest Class* | *Largest Class* |
|---|---|---|---|---|---|
| **20Ng** | 11293 | 7528 | 18821 | 628 | 999 |
| **R8** | 5485 | 2189 | 7674 | 51 | 3923 |
| **Webkb** | 2803 | 1396 | 4199 | 504 | 1641 |
| **Cade12** | 27322 | 13661 | 40983 | 625 | 8473 |

**Πίνακας 4.6** – Στατιστικά δεδομένα Dataset

*Στο Appendix στο τέλος του κειμένου υπάρχουν παραδείγματα κειμένων από κάθε κατηγορία αφού έχουν υποστεί προ-επεξεργασία.*

## 4.3 *Προ-επεξεργασία κειμένων*

Σε κάθε εργασία κατηγοριοποίησης πριν από τη δεικτοδότηση των κειμένων προηγούνται μερικές βασικές διαδικασίες οι οποίες χρησιμοποιούνται για την απλοποίηση των κειμένων . Οι συνήθεις διαδικασίες που ακολουθούνται φαίνονται στο παρακάτω σχήμα. Μπορεί να γίνει καμία, μία, όλες ή μερικές από αυτές τις διαδικασίες. Στην παρούσα εργασία πραγματοποιούνται όλες όπως φαίνεται στο παρακάτω σχήμα.

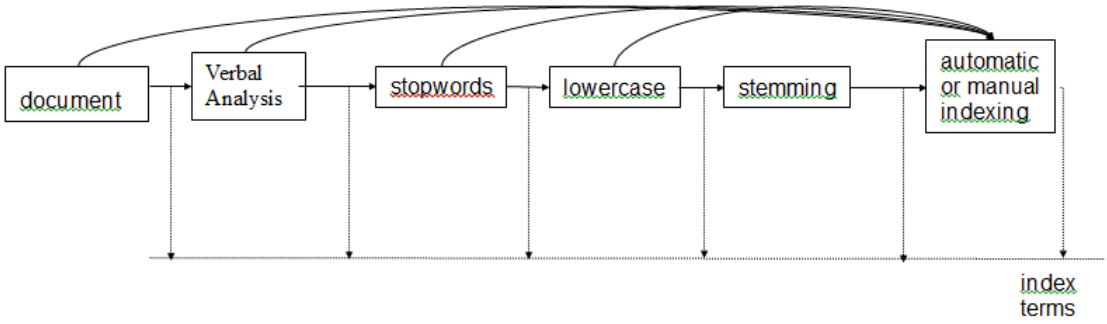


**Figure 4.1 –** Διαδικασία προ-επεξεργασίας των κειμένων

### 4.3.1 ***Verbal Analysis.***

Διαδικασία μετατροπής ενός κειμένου από σειρά χαρακτήρων (character stream) σε σειρά λέξεων (word stream). Κάθε λέξη χωρίζεται από την επόμενη με κενούς χαρακτήρες (space) Πέρα από τις λέξεις όμως υπάρχουν και αριθμοί , σημεία στίξης κ.α. τα οποία πρέπει να ληφθούν υπόψη (και να απαλειφθούν) , όπως εξηγείται παρακάτω :

- Οι αριθμοί δεν θεωρούνται καλές περιπτώσεις index terms διότι χωρίς τα συμφραζόμενα το νόημά τους είναι αρκετά ασαφές
- Συνήθως η απαλοιφή του συμβόλου συλλαβισμού (‘-’) δε δημιουργεί προβλήματα στην ανάκτηση πληροφορίας (π.χ. State-of-the-art -> state of the art).
- Συνήθως τα σύμβολα στίξης αφαιρούνται εντελώς κατά τη φάση της λεκτικής ανάλυσης κειμένων και ερωτήσεων (I.K.A -> IKA, D.N.A. -> DNA)

### *4.3.2 **Stopwords***

Λέξεις οι οποίες εμφανίζονται στην πλειοψηφία των κειμένων δεν είναι χρήσιμες.

- Αυτές οι λέξεις καλούνται stopwords.
- Τέτοιες λέξεις τα Άρθρα, οι προθέσεις, οι σύνδεσμοι κ.α.
- Η απαλοιφή των stopwords μειώνει σημαντικά το μέγεθος ενός κειμένου.

### *4.3.3 **Lowercase***

Όλα τα γράμματα μετατρέπονται σε πεζά ή σε κεφαλαία. (HORSE, Horse, horse). Αντίστοιχα μπορούν όλα τα γράμματα να μετατραπούν σε κεφαλαία.

### *4.3.4 **Stemming***

Η ίδιες σημασιολογικά λέξεις μπορεί να υπάρχουν με διαφορετική την μορφή στο κείμενο (π.χ. connect, connecting). Stem είναι το τμήμα της λέξης που απομένει μετά την απομάκρυνση prefix και suffix. Μετά το stemming μειώνεται σημαντικά ο αριθμός των διακριτών λέξεων του κειμένου. Το πιο σημαντικό μέρος είναι η απομάκρυνση του suffix, διότι οι διαφορετικές εκδοχές μίας λέξης προσδιορίζονται με διαφορετικές καταλήξεις.

- Αλγόριθμος Porter, για την απομάκρυνση των καταλήξεων από τις λέξεις. Χρησιμοποιούνται μερικοί κανόνες (π.χ. s->null).

### *4.3.5 Συνοπτικά:*

Για τις συλλογές Webkb , Reuters – R8 και 20Newsgroups ακολουθείται η παρακάτω προ-επεξεργασία ώστε τα κείμενα να έρθουν στη τελική τους μορφή για εκπαίδευση και αξιολόγηση :

1. Αντικατάσταση των TAB , NEWLINE , RETURN και των σημείων στίξης με SPACE .
2. Αντικατάσταση των πολλαπλών SPACE με ένα SPACE .
3. Μετατροπή όλων των γραμμάτων σε πεζά .
4. Προσθήκη του τίτλου (κατηγορία) κάθε εγγράφου στην αρχή του κειμένου .
5. Απαλοιφή λέξεων με λιγότερα από 3 γράμματα .
6. Απαλοιφή 524 SMART-words πχ. "they","with" τα οποία δε δίνουν κάποια πληροφορία για τη κατηγορία του κειμένου . Πολλές από αυτές τις λέξεις έχουν ήδη αναληφθεί αφού έχουν λιγότερα από 3 γράμματα πχ "and" .
7. Εφαρμογή του αλγορίθμου Porter Stemmer .

*Η συλλογή Cade έχει κατεβαστεί ήδη προ-επεξεργασμένη .*

### *4.3.6 Ο αλγόριθμος του Porter*

Ο αλγόριθμος του Porter [23] παρουσιάστηκε το 1980. Βασίζεται στην ιδέα ότι οι καταλήξεις στην αγγλική γλώσσα (περίπου 1200) δημιουργούνται από συνδυασμούς μικρότερων και απλούστερων καταλήξεων. Αποτελείται από 5 ή 6 βήματα (ανάλογα με τον ορισμό του βήματος) κάθε ένα από τα οποία εκτελείται γραμμικά. Στον αλγόριθμο γίνονται ορισμένες παραδοχές:

- Ένα σύμφωνο είναι ένα γράμμα εκτός από τα A, E, I, O, U και Y. Ακόμα ως σύμφωνο θεωρείται ένα φωνήεν του οποίου προηγείται ένα φωνήεν. Για παράδειγμα στη λέξη "boy" τα σύμφωνα είναι τα B και Y ενώ στη λέξη "try" είναι τα T και R.
- Ένα φωνήεν είναι ένα γράμμα που δεν είναι σύμφωνο. Μία ακολουθία συμφώνων με μέγεθος μεγαλύτερο ή ίσο με ένα αποτυπώνεται σαν **C** ενώ η αντίστοιχη ακολουθία από φωνήεντα αναπαρίσταται από το γράμμα **V**. Έτσι μια λέξη μπορεί να αναπαρασταθεί σαν **[C] (VC) m [V]** όπου ο δείκτης m δείχνει m επαναλήψεις του **VC** και οι αγκύλες **[ ]** ορίζουν την προαιρετική εμφάνιση των περιεχομένων τους. Η τιμή **m** ονομάζεται μέτρο μιας λέξης και μπορεί να πάρει οποιαδήποτε τιμή μεγαλύτερη ή ίση με το μηδέν. Χρησιμοποιείται για να αποφασιστεί εάν μια κατάληξη θα πρέπει να αφαιρεθεί.

### *4.3.7 Η λειτουργία του αλγορίθμου του Porter*

Παρουσιάζονται τα βήματα που ακολουθεί αλγόριθμος Porter Stemmer :

- Στο πρώτο βήμα του αλγορίθμου γίνεται χειρισμός των πληθυντικών και των αορίστων . Το βήμα αυτό λόγω πολυπλοκότητας χωρίζεται σε τρία υπο-βήματα. Το πρώτο χειρίζεται τους πληθυντικούς (π.χ. kisses -> kiss και αφαίρεση του es).
- Το δεύτερο αφαιρεί τις καταλήξεις ed και ing ή μετατρέπει την κατάληξη eed σε ee όπου αυτό απαιτείται. Η διαδικασία συνεχίζεται και αν έχει αφαιρεθεί η κατάληξη ed ή η ing η ρίζα που απομένει μετασχηματίζεται ακολουθώντας συγκεκριμένους κανόνες.
- Το τρίτο κομμάτι απλώς μετατρέπει το τελικό y σε i. Τα βήματα 2-5 ασχολούνται κυρίως με τη διαφορετική σειρά στις ομάδες καταλήξεων. Για το λόγο αυτό μετατρέπουν τις διπλές καταλήξεις σε μια κατάληξη ενώ αφαιρούν και καταλήξεις που πληρούν ορισμένα κριτήρια.

#### *4.3.7.1 Σχηματική απεικόνιση της λειτουργίας του αλγορίθμου του Porter*

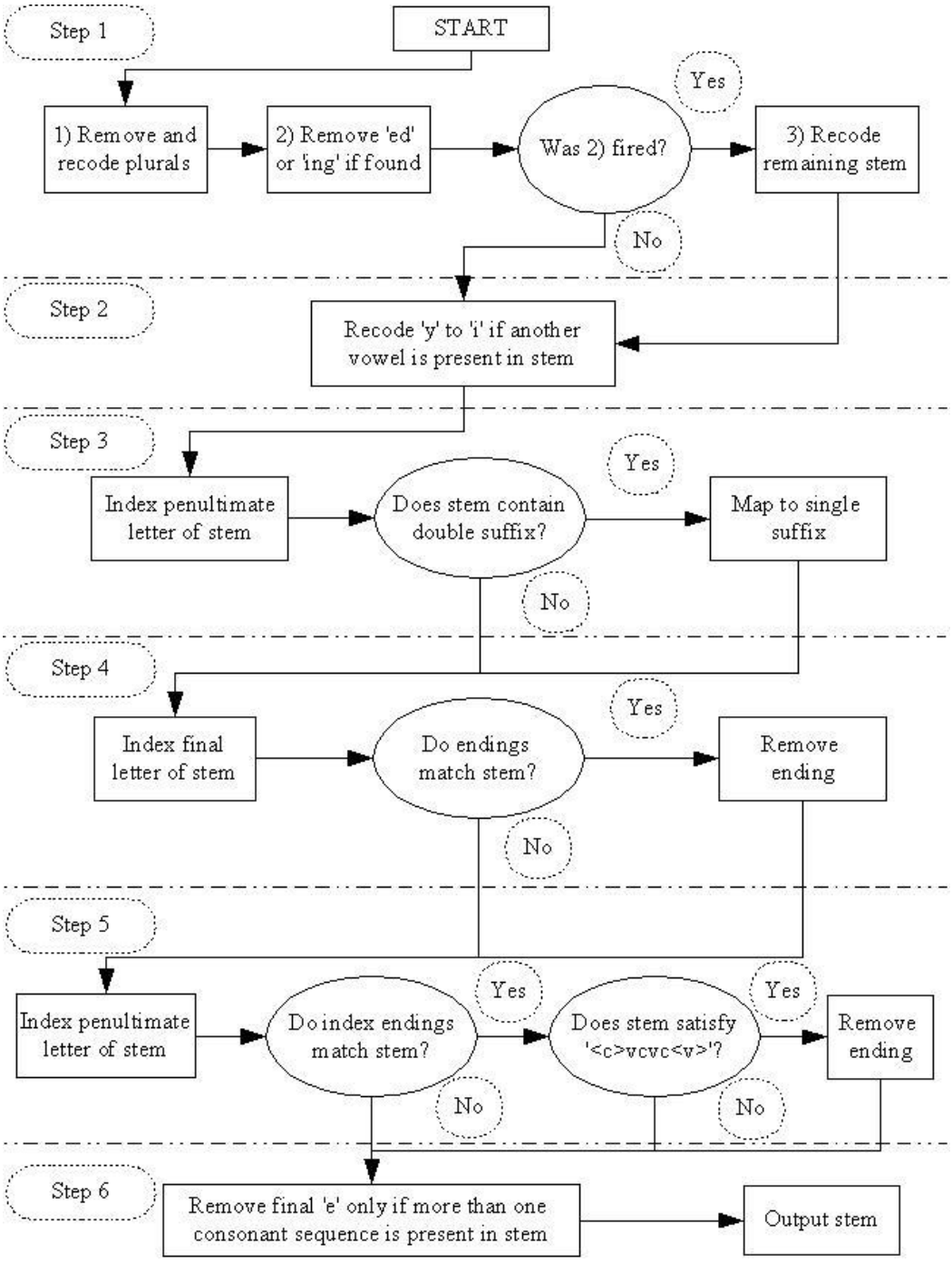


**Figure 4.2 –** Σχηματική περιγραφή του αλγορίθμου του Porter

# 5

# *Κατηγοριοποίηση Κειμένων με χρήση GMM*

Στο κεφάλαιο αυτό θα παρουσιαστεί το κύριο κομμάτι αυτής της εργασίας , το οποίο είναι η κατηγοριοποίηση κειμένων με βάση το θέμα, χρησιμοποιώντας συνεχείς κατανομές και πιο συγκεκριμένα τα Gaussian Mixture Models.

## *5.1 Εισαγωγή*

Όλες οι σχετικές εργασίες (state-of-the-art), που παρουσιάστηκαν στο κεφάλαιο 4, αφορούν τη κατηγοριοποίηση κείμενων με βάση το θέμα έχοντας υλοποιηθεί είτε στο διακριτό χώρο (Naïve Bayes, Vector Method, k-NN), είτε στο συνεχή χώρο χρησιμοποιώντας τις αποστάσεις των χαρακτηριστικών διανυσμάτων μεταξύ των κειμένων. Το ενδιαφέρων αυτής της εργασίας είναι η αναπαράσταση των κειμένων στο συνεχή χώρο, χρησιμοποιώντας συνεχείς κατανομές, και μετέπειτα η κατηγοριοποίηση τους. Ο λόγος είναι ότι με αυτό τον τρόπο τα κείμενα-διανύσματα μπορούν να χαρακτηριστούν από τα στατιστικά χαρακτηριστικά τους (π.χ. μέση τιμή ,βάρη, διασπορά) και όχι από τους όρους τους αυτούς καθαυτούς.

Το πρόβλημα είναι πως μπορεί να αναπαρασταθεί ένα κείμενο σε συνεχή χώρο. Για την αναπαράσταση αυτή χρησιμοποιούνται συνεχείς κατανομές, συγκεκριμένα Gaussian Mixture Models (GMM), έτσι ώστε τα κείμενα πλέον αναπαρίσταται από Gaussian κατανομές με μέση τιμή, διασπορά και βάρος. Τα κείμενα αρχικά αναπαριστώνται ως διανύσματα στο διακριτό χώρο (όπως στις μεθόδους Vector Method – k-NN), μετέπειτα χρησιμοποιείται η τεχνική SVD (όπως στο LSI) για τη μείωση των διαστάσεων των χαρακτηριστικών διανύσματων και για να μη χαθεί η σημασιολογική εξάρτηση των λέξεων μεταξύ των κειμένων. Κατόπιν τα διανύσματα των κειμένων αναπαριστώνται με Gaussian κατανομές.

## *5.2 Gaussian Mixture Models*

Στη κατηγοριοποίηση κειμένων με χρήση GMM θεωρούμε ότι κάθε κατηγορία – θέμα αντιπροσωπεύεται από μια μίξη Gaussian μοντέλων, Gaussian Mixture Models, τα οποία για συντομία αναφέρονται ως GMM. Αρχικά θα γίνει μια εισαγωγή-υπενθύμιση των GMM .

### *5.2.1 Μείγμα Gaussian κατανομών (GMM)*

Τα GMM αποτελούνται από έναν σταθμισμένο γραμμικό συνδυασμό (μίξη) Gaussian κατανομών. Από τη θεωρία του Bayes είναι γνωστό ότι :

$$P(\omega_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_j) * P(\omega_j)}{p(\boldsymbol{x})} \qquad 5.2.1$$

όπου $p(\boldsymbol{x}) = \sum_{J=1}^{n} p(\boldsymbol{x}|\omega_j)P(\omega_j)$.

Οι όροι $p(\boldsymbol{x})$ και $P(\omega_j)$ μπορούν να μην ληφθούν υπόψη, διότι ο παράγοντας $p(\boldsymbol{x})$ δε παίζει κάποιο ρόλο στην λήψη της απόφασης, ενώ ο παράγοντας $P(\omega_j) = 1/c$, για κάθε $j = 1,2, \ldots, c$ που σημαίνει ότι οι πιθανότητες για την πραγματική κλάση ενός προτύπου να είναι $\omega_1, \omega_2, \ldots. \omega_c$ είναι ίδιες (ισοπίθανες).

Η ποσότητα της παραπάνω σχέσης $p(\boldsymbol{x}|\omega_j)$ είναι η Gaussian συνάρτηση πυκνότητας πιθανότητας, η οποία εκφράζει τη πυκνότητα πιθανότητας του $\boldsymbol{x}$ δεδομένου ότι η πραγματική κλάση είναι $\omega_j$, και $p(\omega_j|\boldsymbol{x})$ οι εκ των υστέρων πιθανότητα η κλάση $\omega_j$ να είναι δεδομένου ότι έχει παρατηρηθεί ένα διάνυσμα $\boldsymbol{x}$. Έτσι η Gaussian κατανομή μπορεί να γράφει ως :

$$b_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_j) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\boldsymbol{\Sigma}_i|}} e^{\left(-\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)]\right)} \quad 5.2.2$$

και θα αναφέρεται ως $b_i(\boldsymbol{x})$.

Συνάρτηση πυκνότητας πιθανότητας του μίγματος Gaussian κατανομών είναι ένα σταθμισμένο άθροισμα $M$ συνιστωσών Gaussian κατανομών και δίνεται από τη σχέση :

$$P(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} p_i \, b_i(\boldsymbol{x}) \quad 5.2.3$$

,όπου $p_i$ είναι τα βάρη που πολλαπλασιάζονται με τις συναρτήσεις πυκνότητας πιθανότητας και πρέπει να ισχύει ότι : $\sum_{i=1}^{M} p_i = 1$ και το $\lambda$ είναι το σύνολο των παραμέτρων του μοντέλου (GMM), που καθορίζεται πλήρως από τα διανύσματα μέσης τιμής, από τους πίνακες συνδιασποράς και τα βάρη από όλες τις συνιστώσες Gaussian κατανομών.

Έτσι το $\lambda$ είναι μια συντομογραφία για την ισχύει ότι :

$$\lambda = \{p_i \, , \, \mu_i \, , \, \Sigma_i\} \, , \text{όπου } i = 1,2, \dots \dots , M$$

Στη ταξινόμηση κειμένων, κάθε κατηγορία – θέμα από τα θέματα της συλλογής δεδομένων αντιπροσωπεύεται από ένα GMM και αντιστοιχίζεται σε αυτό ένα μοντέλο $\lambda$ .

Παρακάτω φαίνεται πως σχηματίζεται μία συνάρτηση πυκνότητας πιθανότητας GMM με συνιστώσες , με δοσμένο ένα διάνυσμα .

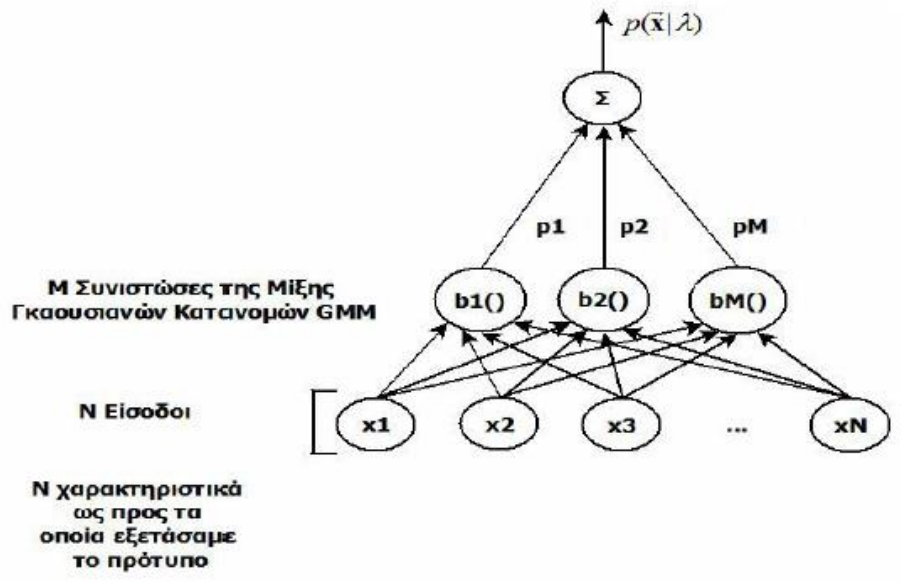


**Figure 5.1 –** Δημιουργία GMM pdf δοσμένου ενός διανύσματος x

Με δοσμένα κάποια δεδομένα εκπαίδευσης, ο στόχος της εκπαίδευσης είναι να γίνει εκτίμηση των παραμέτρων του GMM, δηλαδή του μοντέλου $\lambda$, το οποίο θα ταιριάζει καλυτέρα στην κατανομή των διανυσμάτων των χαρακτηριστικών, που έχουν χρησιμοποιηθεί στην εκπαίδευση. Υπάρχει πληθώρα τεχνικών για την εκτίμηση των παραμέτρων του GMM και μια από τις δημοφιλής είναι η μέθοδος της Εκτίμησης της Μέγιστης Πιθανοφάνειας (Maximum Likehood Estimation) ή σε συντομογραφία ML εκπαίδευση.

Τέλος θα πρέπει να επιλεγεί και η κατάλληλη μορφή πίνακα συνδιασπορών για τα GMM. Στη παρούσα εργασία εξετάζονται οι παρακάτω τύποι covariance :

- Diagonal : ο Σ είναι διαγώνιος πίνακας
- Spherical : ο Σ είναι ίσος με $\sigma_k^2 I$
- Full : ο Σ είναι τυχαίος για κάθε κλάση
- Tied : ο Σ είναι κοινός για όλα τα components

,όπου Σ ο πίνακας συνδιασποράς.

### *5.2.2 Expectation Maximization (EM)*

Ο αλγόριθμος ΕΜ είναι η ιδανική περίπτωση αλγορίθμου που χρησιμοποιείται προκειμένου να λύσει προβλήματα εκτίμησης παραμέτρων για τα GMM. Ο στόχος της εκπαίδευσης είναι να βρει τις παραμέτρους του μοντέλου , οι οποίες θα μεγιστοποιούν την πιθανοφάνεια του GMM για δοσμένα δεδομένα εκπαίδευσης. Για παράδειγμα έστω ότι δίνονται Κ πρότυπα (διανύσματα χαρακτηριστικών) για την εκπαίδευση, $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_K}\}$. Η πιθανοφάνεια του GMM δίνεται από τη σχέση :

$$P(\boldsymbol{X}|\lambda) = \prod_{k=1}^{K} p(\boldsymbol{x}_k|\lambda) \qquad 5.2.4$$

έχοντας υποθέσει ότι τα δεδομένα εκπαίδευσης είναι ανεξάρτητα μεταξύ τους. Αυτή η συνάρτηση είναι μια μη γραμμική συνάρτηση ως προς τις παραμέτρους $\lambda$, με συνέπεια να μην είναι πιθανό να γίνει η μεγιστοποίηση της άμεσα και άρα και ο υπολογισμός των παραμέτρων. Παρόλα αυτά η Εκτίμηση Μεγίστης Πιθανοφάνειας για τον υπολογισμό των παραμέτρων μπορεί να γίνει χρησιμοποιώντας αλγόριθμο τη Μέγιστης Προσδοκίας (Expectation-Maximization, EM). Η βασική ιδέα του ΕΜ αλγορίθμου είναι να ξεκινήσει με ένα αρχικό μοντέλο $\lambda$ για να υπολογίσει (εκτιμήσει) ένα καινούργιο $\lambda'$ τέτοιο ώστε να ισχύει ότι :

$$P(\boldsymbol{X}|\lambda') \geq P(\boldsymbol{X}|\lambda) \qquad 5.2.5$$

Κάθε μια επανάληψη του αλγορίθμου ΕΜ αποτελείται από δύο βήματα : το Ε-βήμα (Expectation-step) και το Μ-βήμα (Maximization-step). Ας υποθέσουμε ότι έχουμε $I$ διανύσματα – κειμένων $x_1, x_2, \ldots, x_I$ για μια κατηγορία από τη συλλογή (dataset) των δεδομένων εκπαίδευσης. Οι παράμετροι του GMM για μια κατηγορία αρχικοποιούνται με τη μέση του αλγορίθμου k-means. Κατά το Μ-βήμα μεγιστοποιείται η συνάρτηση πιθανοφάνειας η οποία και υπολογίζεται σε κάθε επανάληψη του αλγορίθμου κατά το βήμα Ε-βήμα.

Σκοπός της εκτίμησης πιθανοφάνειας (ML), όπως αναφέρθηκε στη προηγούμενη ενότητα, είναι να βρεθούν οι παράμετροι του μοντέλου που μεγιστοποιούν τη πιθανοφάνεια του GMM, δοθέντος των διανυσμάτων-κειμένων εκπαίδευσης. Η βασική ιδέα του ΕΜ αλγορίθμου είναι ότι αρχίζοντας από ένα αρχικό μοντέλο $\lambda$ να υπολογιστεί (εκτιμηθεί) ένα καινούργιο $\lambda'$ τέτοιο ώστε να ισχύει ότι : $P(\boldsymbol{X}|\lambda') \geq P(\boldsymbol{X}|\lambda)$. Το καινούργιο αυτό μοντέλο γίνεται αρχικό για την επόμενη επανάληψη και η διαδικασία αυτή επαναλαμβάνεται μέχρι να επιτευχθεί σύγκλιση.

Η αρχικοποίηση του αλγορίθμου γίνεται όπως αναφέρθηκε με τη χρήση του αλγορίθμου k-means. Σε κάθε επανάληψη του αλγορίθμου βρίσκονται τα κέντρα των K clusters. Κάθε ένα από τα οποία έχει και κάποιο ποσοστό των διανυσμάτων-κειμένων των κατηγοριών της συλλογής. Συνεπώς στο τέλος του αλγορίθμου k-means θα έχουμε τις αρχικές τιμές για τις μέσες τιμές, για τους πίνακες συνδιασποράς καθώς και για τα αρχικά βάρη των K Gaussian συναρτήσεων πυκνότητας πιθανότητας του GMM.

Έχοντας υπολογιστεί οι αρχικές τιμές των παραμέτρων μια επανάληψη του αλγορίθμου το ΕΜ υπολογίζεται ως εξής:

*Κατά το Ε-βήμα* : Οι εκ των υστέρων πιθανότητες (για κάθε μια από τις K Gaussian συναρτήσεις πυκνότητας πιθανότητας) υπολογίζονται για όλα τα διανύσματα-κείμενα εκπαίδευσης της κάθε κατηγορίας με χρήση της ακόλουθης σχέσης :

$$p(i|\boldsymbol{x}(n), \lambda) = \frac{p_i b_i(\boldsymbol{x}(n))}{\sum_{k=1}^{K} p_k b_k(\boldsymbol{x}(n))} \qquad 5.2.6$$

*Κατά τα Μ-βήμα* : Χρησιμοποιούνται οι εκ των υστέρων πιθανότητες από το Ε-βήμα για να εκτιμηθούν οι νέες παράμετροι του μοντέλου με τη χρήση των παρακάτω εξισώσεων :

$$\grave{p_i} = \frac{1}{I}\sum_{n=1}^{I} p(i|\boldsymbol{x}(n), \lambda) \qquad 5.2.7$$

$$\grave{\mu_i} = \frac{\sum_{n=1}^{I} p(i|\boldsymbol{x}(n), \lambda)\, \boldsymbol{x}(n)}{\sum_{n=1}^{I} p(i|\boldsymbol{x}(n), \lambda)} \qquad 5.2.8$$

$$\grave{\Sigma_i} = \frac{\sum_{n=1}^{I} p(i|\boldsymbol{x}(n), \lambda)\, (\boldsymbol{x}(n) - \mu_i)(\boldsymbol{x}(n) - \mu_i)^T}{\sum_{k=1}^{K} p_k b_k(\boldsymbol{x})} \qquad 5.2.9$$

Στη συνέχεια θέτοντας $p_i = \grave{p_i}\,, \mu_i = \grave{\mu_i}$ και $\Sigma_i = \grave{\Sigma_i}$ με επανάληψη του Ε-βήματος και κατόπιν του Μ-βήματος, γίνονται τόσες επαναλήψεις όσες χρειάζονται ώστε να επιτευχθεί η συνθήκη της σύγκλισης.

Συνοψίζοντας, τα αναγκαία βήματα που εκτελούνται κατά την υλοποίηση του αλγορίθμου ΕΜ προκειμένου να εκτιμηθούν οι παράμετροι $\lambda$, για το GMM, είναι τα ακόλουθα :

1. Αποφασίζονται αρχικές τιμές των παραμέτρων .Οι παράμετροι που λαμβάνουν μέρος είναι οι μέσες τιμές, οι πίνακες συνδιασποράς και οι συντελεστές βαρύτητας. Χρησιμοποιείτε ο αλγόριθμος k-means για την αρχικοποίηση των παραμέτρων.
2. Για κάθε διάνυσμα υπολογίζεται η εκ των υστέρων πιθανότητα.
3. Γίνεται επανεκτίμηση των παραμέτρων λ για κάθε μια από τις K Gaussian συναρτήσεις πυκνότητας πιθανότητας.
4. Επαναλαμβάνονται τα Βήματα 2 και 3 μέχρι οι εκτιμήσεις των παραμέτρων να συγκλίνουν.

Το βήμα 4 εφαρμόζεται με συνεχόμενες επαναλήψεις μέχρι η διαφορά των εκτιμήσεων δυο διαδοχικών επαναλήψεων να είναι μικρότερη από κάποιο προκαθορισμένο όριο. Τα παραπάνω βήματα του αλγορίθμου διαγραμματικά φαίνονται παρακάτω :

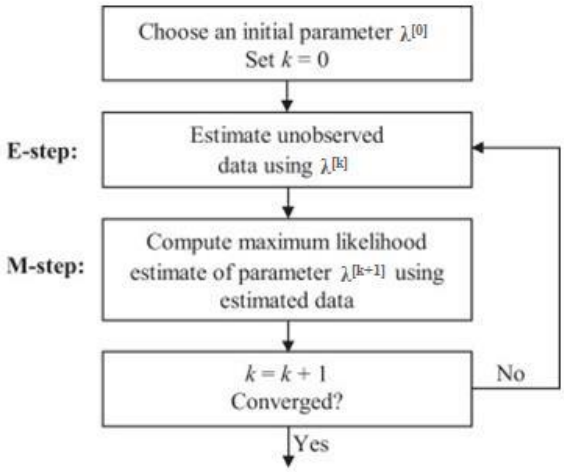


**Figure 5.2 –** Λειτουργία του αλγορίθμου EM

### *5.2.2.1 Ο αλγόριθμος k-means*

Στην παράγραφο αυτή γίνεται μια συνοπτική παρουσίαση των βημάτων που ακολουθεί ο αλγόριθμος k-means ο οποίος χρησιμοποιείται για την αρχικοποίηση των τιμών (μέσες τιμές, πίνακες συνδιασποράς, βάρη ) του GMM :

1. Όρισε το πλήθος των clusters
2. Αρχικοποίησε clusters με τυχαία επιλογή κέντρων clusters
3. Υπολόγισε το μέσο κάθε cluster
4. Απέδωσε κάθε δείγμα στο πλησιέστερο μέσο
5. Αν η κατανομή των δειγμάτων δεν άλλαξε τερμάτισε, αλλιώς βήμα 3

## *5.3 Αλγόριθμος Αναγνώρισης Κειμένου*

Το τελικό στάδιο της εργασίας είναι να γίνει η αναγνώριση ενός αγνώστου κειμένου και η κατηγοριοποίηση του στη κατηγορία που ανήκει. Τα άγνωστα κείμενα βρίσκονται στα dataset και είναι τα κείμενα εκπαίδευσης (testing documents), για τα οποία υπάρχει η πληροφορία της κατηγορία στην οποία ανήκουν έτσι ώστε να μπορεί να γίνει η σύγκριση του αποτελέσματος και να βγει η απόδοση του συστήματος.

Το πρώτο στάδιο της αναγνώρισης ενός νέου κειμένου από τον κατηγοριοποιητή είναι η προ-επεξεργασία του, όπως ακριβώς γίνεται και στο στάδιο της εκπαίδευσης. Τα αρχικά κείμενα περιέχουν όρους οι οποίοι δεν περιέχουν χρήσιμη πληροφορία για το περιεχόμενο του κειμένου όπως αριθμοί , σημεία στίξης , άρθρα κ.α τα οποία αν αφαιρεθούν το κείμενο έρχεται σε μια πιο συμπαγή μορφή για τους αλγορίθμους κατηγοριοποίησης. Τα βήματα της προ-επεξεργασία των κειμένων περιγράφηκαν αναλυτικά στο κεφάλαιο 4.

Έπειτα ακολουθεί το στάδιο της αναπαράστασης του κειμένου ως διάνυσμα για να δοθούν τα βάρη στους όρους-λέξεις του. Στην εργασία έχουν υλοποιηθεί δύο τρόποι υπολογισμού βαρών, τα δυαδικά βάρη και τα βάρη σε μορφή tf-idf. Στη δυαδική μορφή τα βάρη $w_{\kappa j}$ παίρνουν τιμές 0 ή 1 ανάλογα με το αν υπάρχει ο όρος στο κείμενο , ενώ στη μορφή tf-idf παίρνουν τιμές στο συνεχές διάστημα [0,1] ανάλογα με το βάρος ο συγκεκριμένος όρος στο κείμενο. Αναλυτικά οι δύο αυτές αναπαραστάσεις έχουν εξηγηθεί στο κεφάλαιο 3. Εδώ θα πρέπει να χρησιμοποιηθεί το ίδιο μοντέλο με αυτό που χρησιμοποιήθηκε στην εκπαίδευση. Έτσι πχ αν η εκπαίδευση του ταξινομητή έχει γίνει με το μοντέλο binary ή tf-idf τότε αντίστοιχα και το νέο κείμενο θα πρέπει να έχει την αντίστοιχη μορφή κατανομής βαρών στους όρους.

Επόμενο βήμα είναι η μείωση των διαστάσεων του διανύσματος του νέου "αγνώστου" κειμένου. Και σε αυτό το στάδιο ο αλγόριθμος μείωσης διαστάσεων είναι ο ίδιος με αυτών που χρησιμοποιείται στην εκπαίδευση. Στην παρούσα εργασία όπως έχει αναφερθεί χρησιμοποιείται η μέθοδος Singular Value Decomposition (SVD). Έτσι το διάνυσμα του κειμένου για κατηγοριοποίηση έχει τις ίδιες διαστάσεις με τα κείμενα εκπαίδευσης. Το πρόβλημα της συγκεκριμένης μεθόδου είναι ότι δεν υπάρχει ένα σταθερός αριθμός $k$, μείωσης διαστάσεων, ώστε να είναι αποδοτική η μέθοδος για διαφορετικές συλλογές κειμένων. Αυτό συμβαίνει επειδή κάθε συλλογή έχει διαφορετικό μέγεθος και διαφορετικά είδη κειμένων. Αν για παράδειγμα έχουμε μόνο 2 κατηγορίες κειμένων με κατηγορίες αρκετά διαφορετικές η μια από την άλλη τότε ένας πολύ μικρός αριθμός $k$ όρων είναι αρκετός για να τα περιγράψει επαρκώς χωρίς απώλεια πληροφορίας. Αντίθετα για πολλά κείμενα, πολλών διαφορετικών (και σε μερικές περιπτώσεις όμοιων κατηγοριών) χρειάζονται αρκετά περισσότεροι όροι ώστε να μη χαθεί πληροφορία. Επίσης η επιλογή παραπάνω όρων έχει επίσης αντίθετα αποτελέσματα, καθώς χρησιμοποιούνται και όροι που δε εμφανίζονται σπάνια και δε περιέχουν χρήσιμη πληροφορία. Έχει παρατηρηθεί πάντως ότι γενικά για $k > 500$ η απόδοση της SVD είναι πολύ χαμηλή.

Αφού το διάνυσμα είναι στην κατάλληλη μορφή, όμοια με τα διανύσματα εκπαίδευσης, σχηματίζεται η συνάρτηση πυκνότητας πιθανότητας GMM του διανύσματος.

Στο τελικό στάδιο υπολογίζεται σε ποια από τις GMM που αντιπροσωπεύουν τις κατηγορίες για ταξινόμηση, έχει μεγαλύτερη πιθανότητα να ανήκει το νέο "άγνωστο" κείμενο. Η τελική απόφαση παίρνεται βάσει της τιμής της μέγιστης πιθανοφάνειας (Maximum Likehood –ML). Στο παρακάτω σχήμα φαίνονται διαγραμματικά τα βήματα για την κατηγοριοποίηση ενός αγνώστου κειμένου.

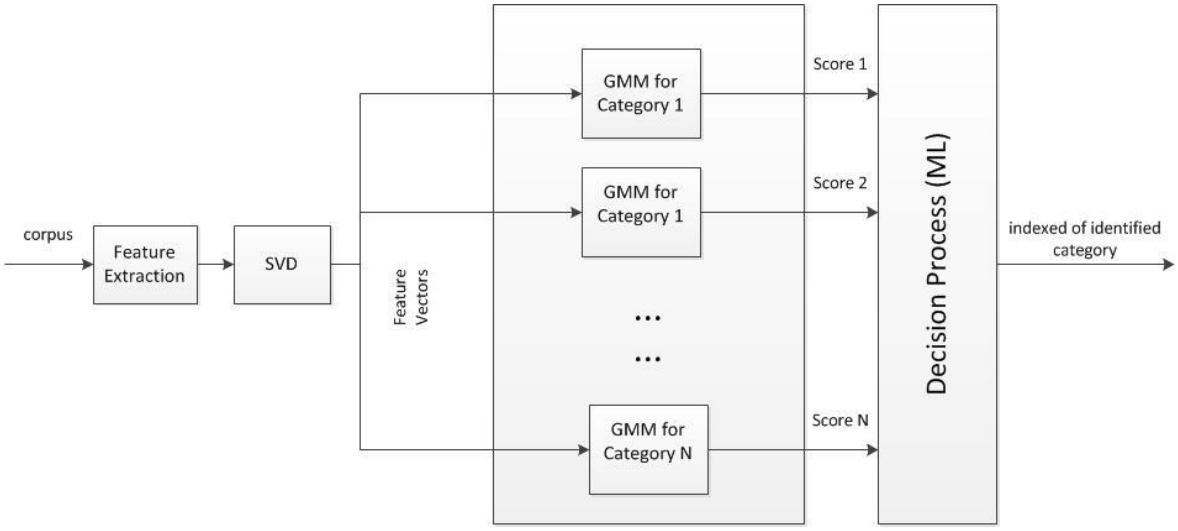


**Figure 5.3 –** Διαδικασία κατηγοριοποίησης ενός αγνώστου κειμένου

# 6

# *Αξιολόγηση*

Στο κεφάλαιο αυτό γίνεται η αξιολόγηση του συστήματος κατηγοριοποίησης κειμένων και παρουσιάζονται τα αποτελέσματα όλων των τεχνικών που χρησιμοποιήθηκαν και αναλύθηκαν στα παραπάνω κεφάλαια. Η προσομοίωση των πειραμάτων έγινε σε γλώσσα Python.

## *6.1 Μετρική ακρίβειας*

Η απόδοση του συστήματος μετριέται χρησιμοποιώντας την ευστοχία (accuracy) του συστήματος στα testing documents. Εφόσον είναι γνωστές οι κατηγορίες των testing documents το accuracy είναι η καλύτερη μέθοδος υπολογισμού της απόδοσης. Η ακρίβεια (accuracy) ενός κατηγοριοποιητή είναι το ποσοστό των κειμένων που έχουν κατηγοριοποιηθεί σωστά προς το συνολικό αριθμό των κειμένων για κατηγοριοποίηση . Περιγράφεται από τον τύπο :

$$Accuracy = \frac{\#Correctly\ classified\ documents}{\#Total\ documents}$$

## *6.2 Baseline συστήματος*

Για να γίνει η αξιολόγηση του συστήματος χρησιμοποιούνται τα αποτελέσματα σχετικών εργασιών (κεφάλαιο 2), που έχουν γίνει στο τομέα της κατηγοριοποίησης κειμένων, και συγκρίνεται η απόδοση τους με της απόδοση της παρούσας εργασίας. Στο κεφάλαιο 2 (state-of-the-art) παρουσιάστηκαν τέσσερις διαφορετικές μέθοδοι με τις οποίες θα συγκριθεί η απόδοση χρησιμοποιώντας όλα τα dataset (κεφάλαιο 4). Η προ-επεξεργασία των κειμένων είναι κοινή για όλες τις μεθόδους ώστε να αποτελέσματα να είναι άμεσα συγκρίσιμα . Επίσης υπολογίζεται και ο μέσος όρος απόδοσης κάθε μεθόδου για όλα τα dataset. Παρακάτω παρουσιάζονται οι αποδόσεις των τεσσάρων μεθόδων (Naive Bayes, Vector Method , K-nn , LSI) για κάθε dataset καθώς και ο μέσος όρος απόδοσης αυτών :

| | ***Naïve Bayes*** | ***Vector Mecthod*** | ***k-NN*** | **LSI** |
|---|---|---|---|---|
| **Webkb** | 0.835 | 0.644 | 0.725 | 0.735 |
| **20Ng** | 0.810 | 0.724 | 0.759 | 0.749 |
| **R8** | 0.960 | 0.788 | 0.852 | 0.941 |
| **Cade** | 0.572 | 0.414 | 0.512 | 0.432 |
| **Average** | 0.794 | 0.643 | 0.712 | 0.715 |

**Πίνακας 6.1 -** Πίνακας αποτελεσμάτων για το baseline της εργασίας

## *6.3 Οργάνωση πειραμάτων*

Για να βρεθεί η βέλτιστη απόδοση του συστήματος είναι σημαντικό να δοκιμαστούν διαφορετικές παράμετροι. Τα στάδια για την υλοποίηση έχουν περιγραφεί στα προηγούμενα κεφάλαια όμως σε πολλά σημεία υπάρχει δυνατότητα διαφορετικής επιλογής παραμέτρων και τιμών. Τα στάδια αυτά συνοπτικά είναι ο (1) τρόπος κατανομής των βαρών στην αναπαράσταση των κειμένων ως διάνυσμα, (2) ο αριθμός των διαστάσεων μετά τη μείωση τους με τη τεχνική SVD, (3) ο αριθμός των Gaussian components σε κάθε GMM καθώς επίσης και (4) το πώς συμπεριφέρεται το σύστημα με τους διαφορετικούς τύπους covariance (tied, full, diagonal, spherical). Επίσης ένα σημαντικό μέτρο για την απόδοση ενός συστήματος είναι και ο χρόνος που χρειάζεται ένα σύστημα κατηγοριοποίησης κειμένων να ταξινομήσει ένα άγνωστο κείμενο.

Αρχικά θα πρέπει να επιλεγεί ο κατάλληλος τύπος covariance για τις GMM υπολογίζοντας την απόδοση (accuracy) για κάθε έναν χρησιμοποιώντας όλα τα dataset (ανάλογα με τη μορφή τους τα κείμενα μπορεί να συμπεριφέρονται καλύτερα με διαφορετικούς τύπους covariance). Επίσης δοκιμάζεται διαφορετικό μέγεθος διαστάσεων των διανυσμάτων καθώς, ανάλογα με το περιερχόμενο τους, τα κείμενα κάθε dataset συμπεριφέρονται καλύτερα σε διαφορετικό αριθμό διαστάσεων.

Άλλη παράμετρος που επηρεάζει την απόδοση του συστήματος είναι η μορφή αναπαράστασης των βαρών των όρων των διανυσμάτων. Όπως έχει αναφερθεί δοκιμάζονται δυο τρόποι κατανομής βαρών , η δυαδική μορφή και η μορφή tf-idf. Σε αυτό το σημείο η tf-df μορφή αναμένεται να δώσει καλύτερα αποτελέσματα, αλλά η δυαδική μορφή εξακολουθεί να χρησιμοποιείται σε πολλές εργασίες διανυσματικής απεικόνισης κειμένων, όποτε είναι σημαντικό να ελεγχθεί η απόδοση της.

Πολύ σημαντικό σε κάθε εργασία που χρησιμοποιούνται GMM είναι να βρεθεί ο σωστό αριθμός Gaussian για τα GMM. Μικρότερος αριθμός μπορεί να μπορεί να συγχωνεύσει δεδομένα διαφορετικών κατηγοριών λαμβάνοντας τα ως ίδια, ενώ αντίθετα μεγαλύτερος αριθμός μπορεί να ξεχωρίσει δεδομένα ίδιων χαρακτηριστικών .

Όπως περιγράφηκε, η μείωση διαστάσεων με χρήση της τεχνικής singular value decomposition (SVD), αναφέρθηκε ότι δεν υπάρχει ένας αριθμός k διαστάσεων που να εγγυάται τα καλύτερα αποτελέσματα. Έτσι πρέπει να μετρηθεί η απόδοση του συστήματος για διάφορες τιμές k για κάθε dataset. Το dataset που επιλέγεται παίζει μεγάλο ρόλο στη βέλτιστη τιμή διαστάσεων του SVD καθώς κείμενα με πιο διακριτές κατηγορίες θα συμπεριφέρονται καλύτερα σε λιγότερες διαστάσεις σε αντίθεση με κείμενα με περισσότερες κοινές κατηγορίες.

Τέλος, ο χρόνος αναγνώρισης και κατηγοριοποίησης νέων κειμένων είναι σχεδόν το ίδιο σημαντικός με την απόδοση (accuracy) του κατηγοριοποιητή. Σε πολλές περιπτώσεις (κυρίως live εφαρμογές) επιλέγεται ο κατηγοριοποιητής που δίνει καλά αποτελέσματα σε ελάχιστο χρόνο σε σχέση με έναν άλλο που θα έχει ελαφρώς καλύτερη απόδοση αλλά χρειάζεται μεγάλο χρονικό διάστημα για να κάνει την αναγνώριση. Έτσι συγκρίνεται η απόδοση βάσει χρόνου για το σύστημα της εργασίας με τους υπόλοιπους τρόπους ταξινόμησης (state-of-the-art).

## 6.4 *Πειράματα*

Στην ενότητα αυτή γίνεται παρουσίαση όλων των πειραμάτων που πραγματοποιήθηκαν ώστε να βρεθούν οι κατάλληλες παράμετροι που μεγιστοποιούν την απόδοση του συστήματος. Η υλοποίηση των πειραμάτων έγινε σε γλώσσα Python.

### *6.4.1 Binary και tf-idf μορφή*

Στο πρώτο πείραμα υπολογίζεται ποια κατανομή βαρών για τους όρους του διανύσματος δίνει τα καλύτερα αποτελέσματα. Τα βάρη έχουν κατανεμηθεί με δυο τρόπους, τη δυαδική και τη tf-idf αναπαράσταση.

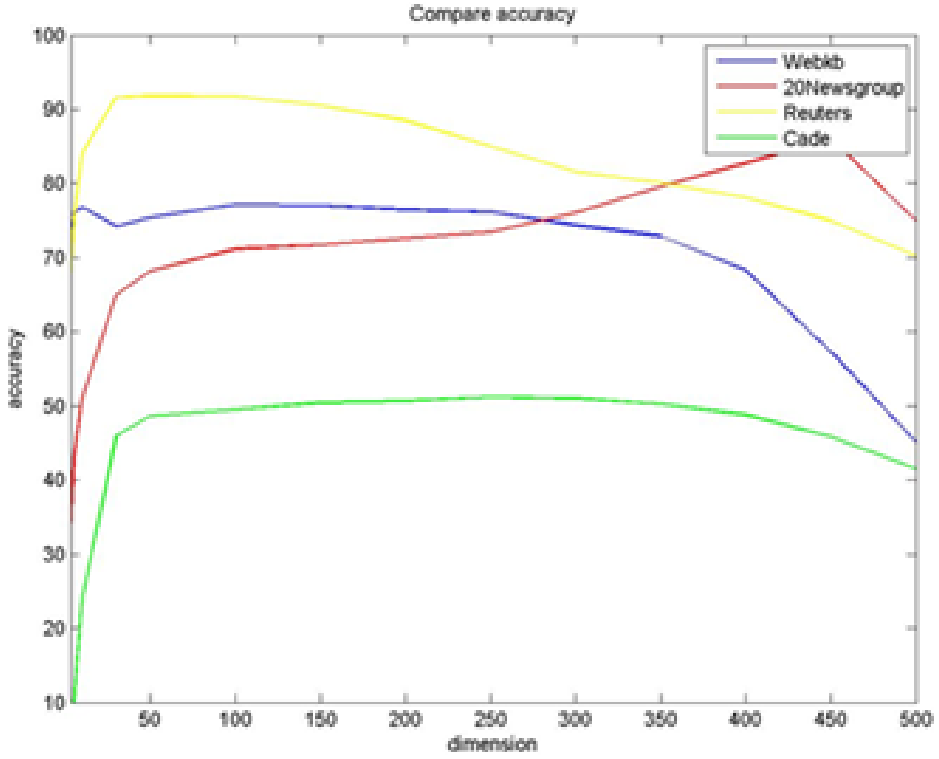


**Figure 6.1 –** Απόδοση με χρήση tf-idf

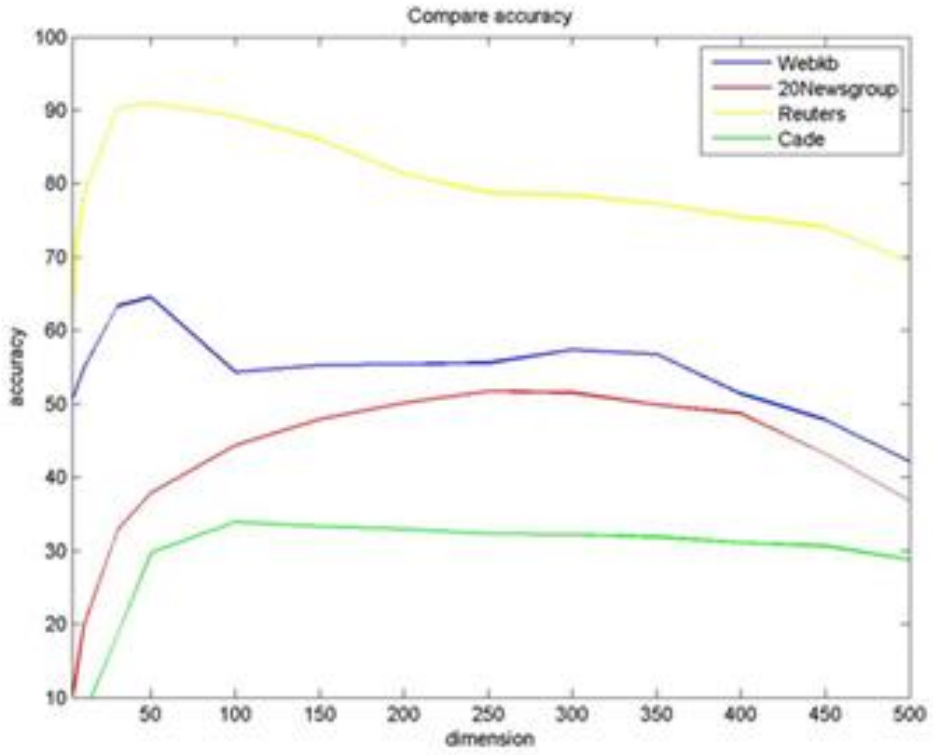


**Figure 6.2 –** Απόδοση με χρήση δυαδικών βαρών

Όπως φαίνεται και από τα παραπάνω διαγράμματα και στα 4 dataset η μορφή tf-df δίνει αισθητά καλύτερα αποτελέσματα όπως ήταν και αναμενόμενο, καθώς κατανέμει καλύτερα το βάρους στους όρους, δίνοντας μεγαλύτερη βαρύτητα στους όρους που έχουν μεγαλύτερη σημασία για ένα κείμενο και μειώνει τους όρους που είτε εμφανίζονται σπάνια, είτε πολύ συχνά σε όλα τα κείμενα και δε παρέχουν χρήσιμη πληροφορία. Τα δυαδικά βάρη συμπεριφέρονται σχετικά καλά μόνο σε ένα dataset Reuters-R8 και πάλι όμως τα αποτελέσματα είναι χειρότερα σε σύγκριση με τα tf-df βάρη .

### *6.4.2 Covariance Type*

Στο πείραμα αυτό επιλέγεται ο τύπος covariance για τα GMM που έχει τα αποδοτικότερα αποτελέσματα για τα κείμενα των συλλογών. Οι τέσσερεις τύποι covariance που δοκιμάζονται είναι οι spherical , diagonal , full και tied covariance .

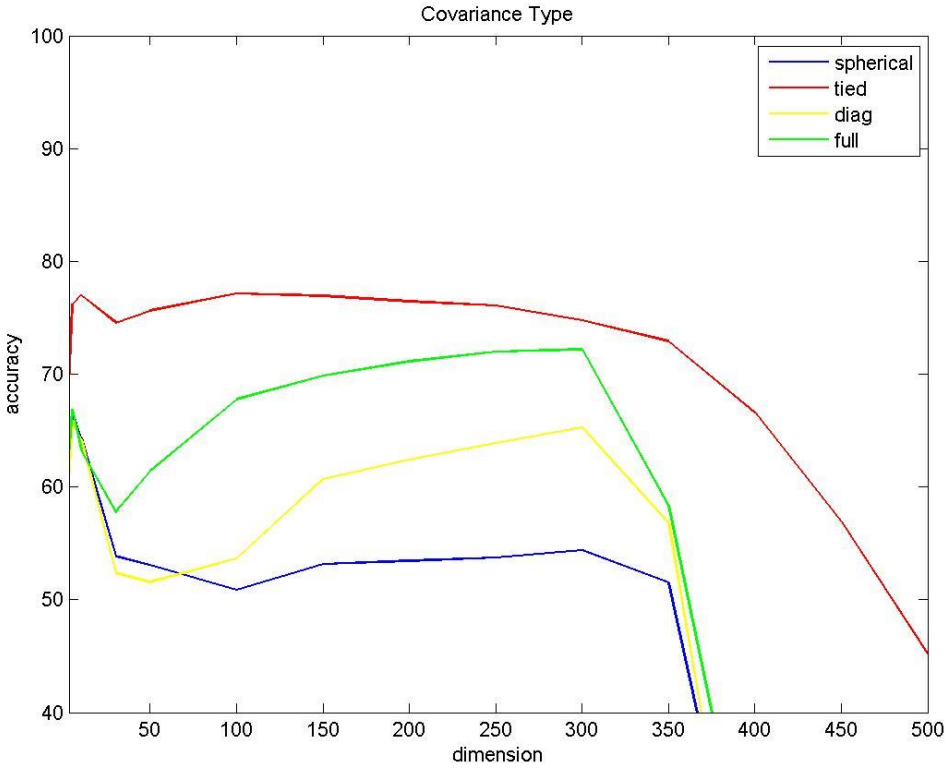


**Figure 6.3 –** Απόδοση βάσει τύπου covariance για το 20Newsgroups

Τα αποτελέσματα και σε αυτό το πείραμα είναι ιδιαίτερα σαφή. Και στις τέσσερις περιπτώσεις το tied covariance εμφανίζεται με διαφορά τα καλύτερα αποτελέσματα. Το παραπάνω σχήμα αναφέρεται στο dataset 20 Newsgroups όμως και στις υπόλοιπες συλλογές τα αποτελέσματα είναι αντίστοιχα.

### *6.4.3 Αριθμός Gaussian στα GMM*

Πολύ σημαντικό για την απόδοση του μοντέλου είναι ο αριθμός των Gaussian σε κάθε GMM. Παρακάτω παρουσιάζεται ένα πίνακας με τις τιμές του accuracy επιλέγοντας διαφορετικούς αριθμούς Gaussian. Έχει επιλεγεί η διάσταση εκείνη που δίνει τα καλύτερα αποτελέσματα για κάθε dataset.

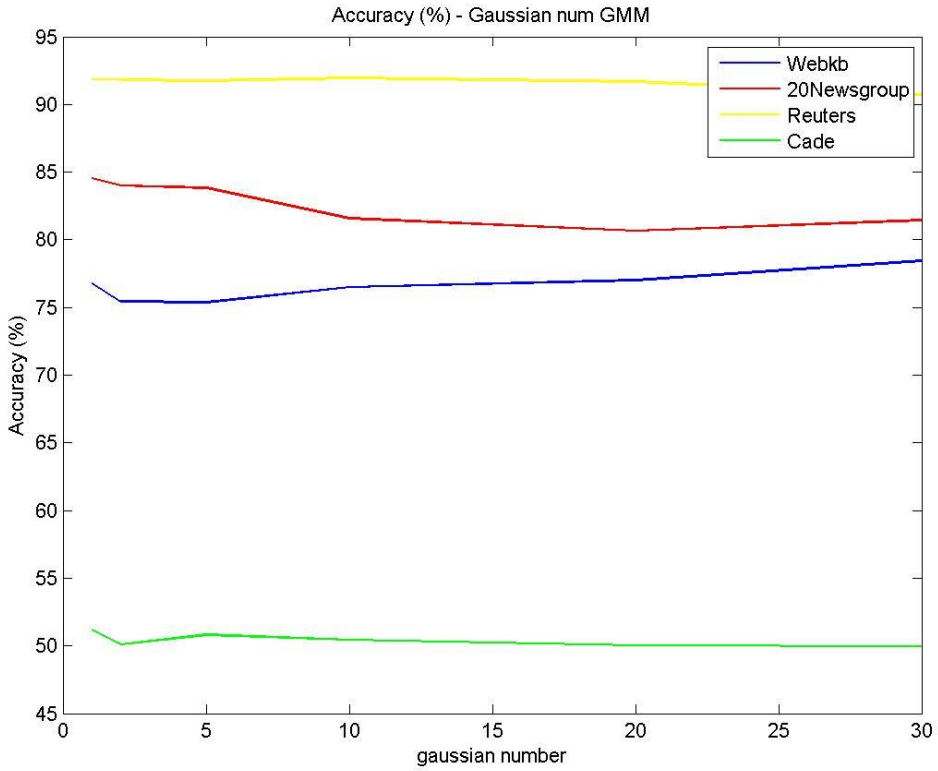


**Figure 6.4** – Απόδοση για κάθε Dataset για διαφορετικούς αριθμούς Gaussian στα GMM

Τα αποτελέσματα αρχικά δεν ήταν τα αναμενόμενα, καθώς όπως φαίνεται ο αριθμός των Gaussian ουσιαστικά δε παίζει ρόλο στην απόδοση του μοντέλου. Για να εξηγηθεί το αποτέλεσμα παρακάτω φαίνεται μια 3D αναπαράσταση των τριών πρώτων διαστάσεων του πίνακα term-document μετά την εφαρμογή του SVD για διαφορετικές κατηγορίες .

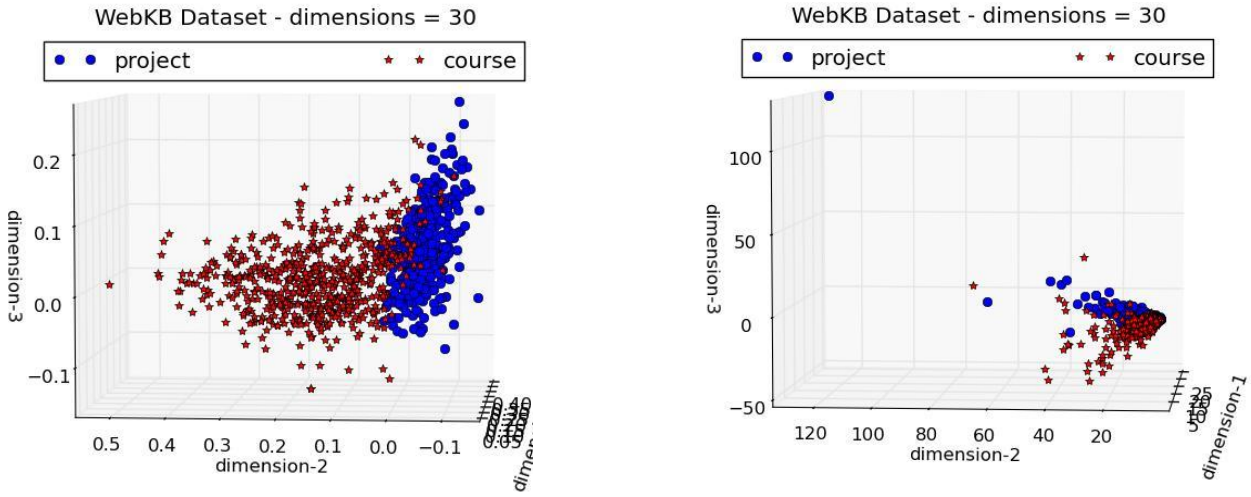


**Figure 6.5 –** Διασπορά στοιχείων για δυο κατηγορία του Webkb dataset με tf-idf (αριστερά) και δυαδικά (δεξιά) βάρη

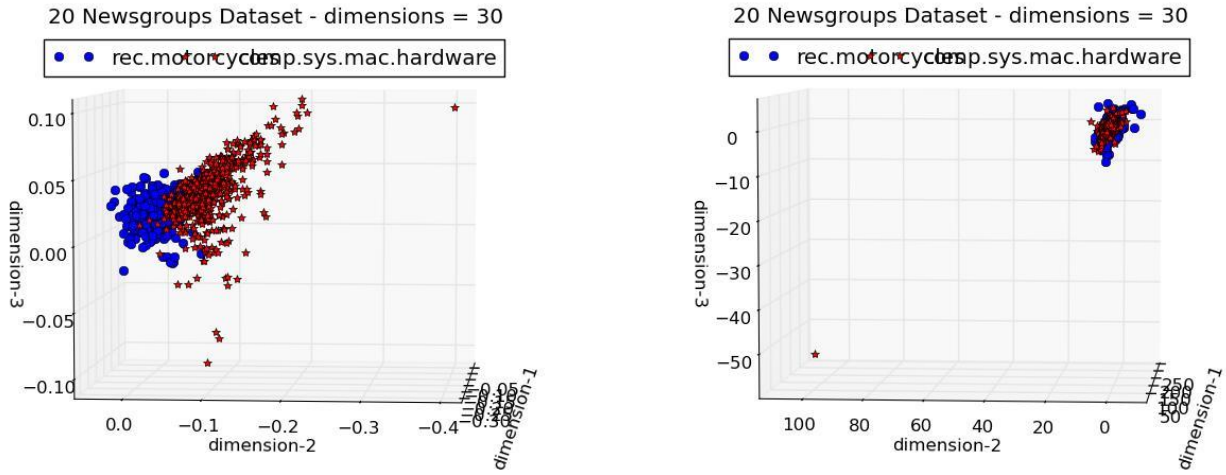


**Figure 6.6 –** Διασπορά στοιχείων για δυο κατηγορία του 20-Newsgroup dataset με tf-idf (αριστερά) και δυαδικά (δεξιά) βάρη

Όπως φαίνεται έστω και από την αναπαράσταση σε τρισδιάστατο χώρο οι διαφορετικές κατηγορίες είναι «μαζεμένες» στο χώρο και ξεχωρίζει η μία από την άλλη. Οπότε αφού τα σημεία δεν είναι διασκορπισμένα στο χώρο και είναι γύρω από μια περιοχή και μία Gaussian είναι αρκετή για να τα χαρακτηρίσει. Ο κυρίαρχος λόγος που συμβαίνει αυτό καταλήξαμε ότι είναι η μορφή tf-df των βαρών η οποία εκτός του ότι δίνει μεγαλύτερο βάρος σε όρους που είναι περισσότερο χρήσιμοι, ταυτόχρονα δίνει πολύ μικρό βάρους σε όρους που εμφανίζονται πολλές φορές σε περισσότερες από μία κατηγορίες ή εμφανίζονται πολύ σπάνια αντίστοιχα. Έτσι τα στοιχεία κάθε κατηγορίας στο ν-διάστατο χώρο είναι περισσότερο μαζεμένα γύρω από ένα σημείο. Παράδειγμα για παράδειγμα μια λέξη που εμφανίζεται ελάχιστες φορές σε κάποια κείμενα με τη tf-idf μορφή θα πάρει τιμή πολύ μικρή, σε αντίθεση με τη δυαδική μορφή που θα πάρει τη τιμή ένα (1) όπως δηλαδή και η λέξη που εμφανίζεται πολύ συχνά και παρέχει μεγαλύτερη πληροφορία.

Παρακάτω φαίνονται ενδεικτικά και τα βάρη των Gaussian στο GMM για 2 dataset , έχοντας επιλεγεί αριθμός 2 Gaussian .

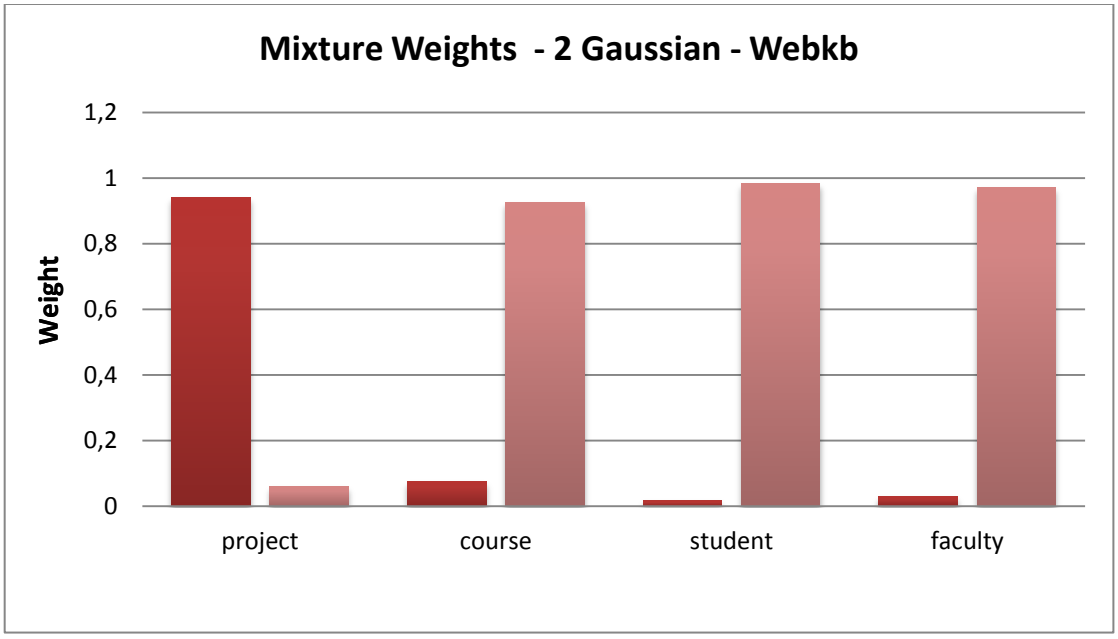


**Figure 6.7 –** Βάρη Gaussian για το Webkb Dataset

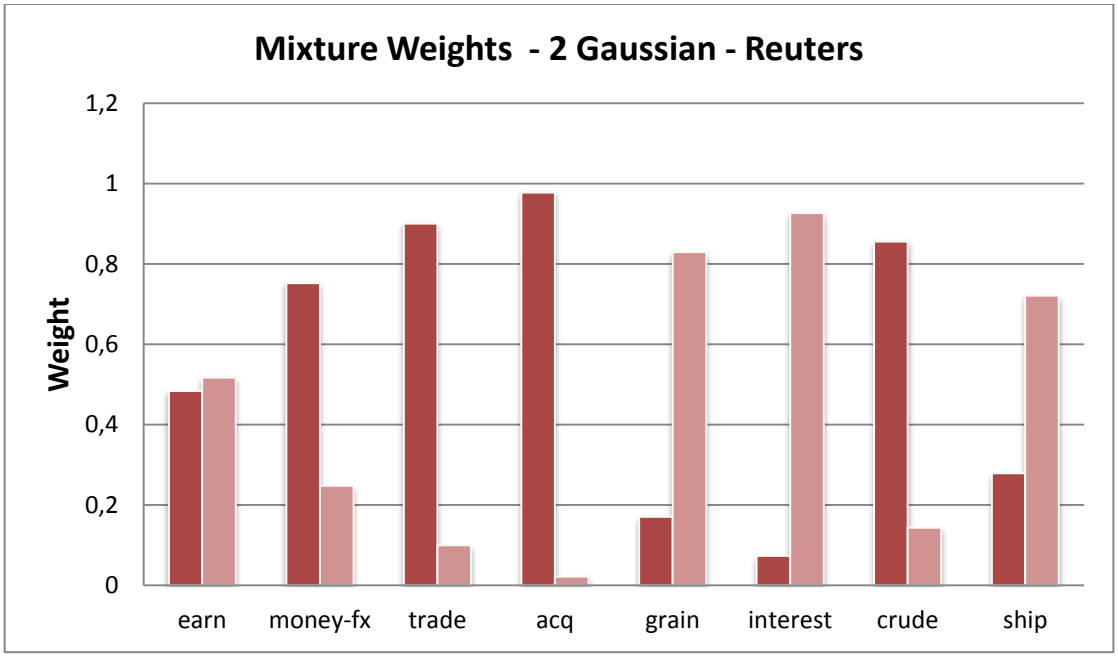


**Figure 6.8 –** Βάρη Gaussian για το Reuters Dataset

Όπως φαίνεται μία Gaussian έχει το μεγαλύτερο βάρος επικαλύπτοντας έτσι τις υπόλοιπες. Παρατηρείται ότι σε όλες σχεδόν τις περιπτώσεις μια Gaussian επικαλύπτει την άλλη έχοντας πολύ μεγαλύτερο βάρος. Το ίδιο συμβαίνει και για μεγαλύτερο αριθμό Gaussian. Έτσι επιβεβαιώνονται τα πειραματικά αποτελέσματα ότι η απόδοση δεν επηρεάζεται σε μεγάλο βαθμό από τον αριθμό των Gaussian στα GMM.

### *6.4.4 Διάσταση SVD*

Ένα από τα σημαντικότερα κομμάτια της υλοποίησης είναι να βρεθεί ο κατάλληλος αριθμός διαστάσεων που είναι αποδοτικότερος μετά τη μείωση τους με SVD. Επειδή δεν υπάρχει κάποιος αριθμός διαστάσεων ο οποίος είναι πάντα αποδοτικός για όλες τις περιπτώσεις είναι απαραίτητο να βρεθεί ο κατάλληλος αριθμός πειραματικά.

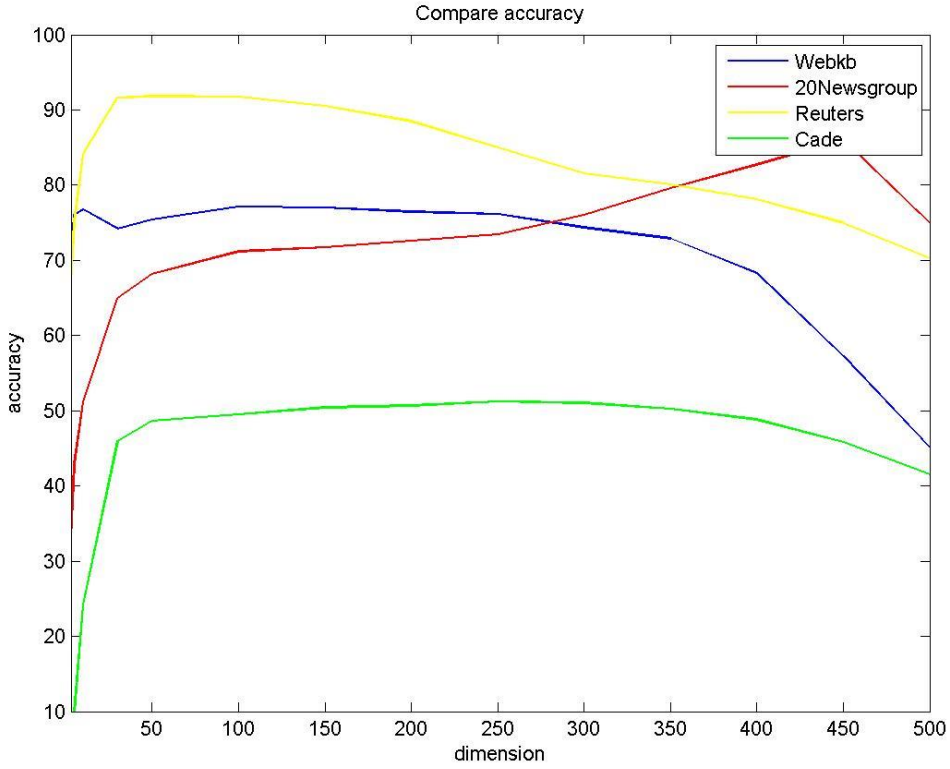


**Figure 6.9 –** Απόδοση και των τεσσάρων dataset σε διαφορετικές διαστάσεις

Παρατηρείται ότι κάθε συλλογή δεδομένων εμφανίζει τα αποδοτικότερα αποτελέσματα σε διαφορετικές διαστάσεις. Το αποτέλεσμα είναι αναμενόμενο καθώς κάθε dataset περιέχει κείμενα διαφορετικού περιεχομένου με διαφορετικό πλήθος κατηγοριών. Για όλες τις συλλογές παρατηρείτε ότι στις μικρές διστάσεις τα η απόδοση είναι μικρή (πολύ λίγες λέξεις - υπάρχει απώλεια πληροφορίας), έπειτα για μεγαλύτερες διαστάσεις η απόδοση ανεβαίνει μέχρι να φτάσει στο μέγιστο επίπεδο της και στη συνέχεια αρχίζει πάλι να μειώνεται (πολλές λέξεις οι οποίες δε είναι χρήσιμες για να χαρακτηρίσουν τα κείμενα - θόρυβος) .

### *6.4.5 Χρόνος κατηγοριοποίησης*

Τέλος πολύ σημαντικό για ένα σύστημα κατηγοριοποίησης κειμένων είναι ο χρόνος που χρειάζεται για να κατηγοριοποιήσει ένα νέο άγνωστο κείμενο. Τα αποτελέσματα παρουσιάζουν το χρόνο που χρειάστηκε για να κατηγοριοποιηθούν όλα τα testing documents για κάθε dataset, είτε σωστά είτε λάθος, στο dataset 20Newsgroups σε μέγεθος διαστάσεων 450 καθώς εκεί παρουσιάζει τα καλύτερα αποτελέσματα βάσει accuracy.

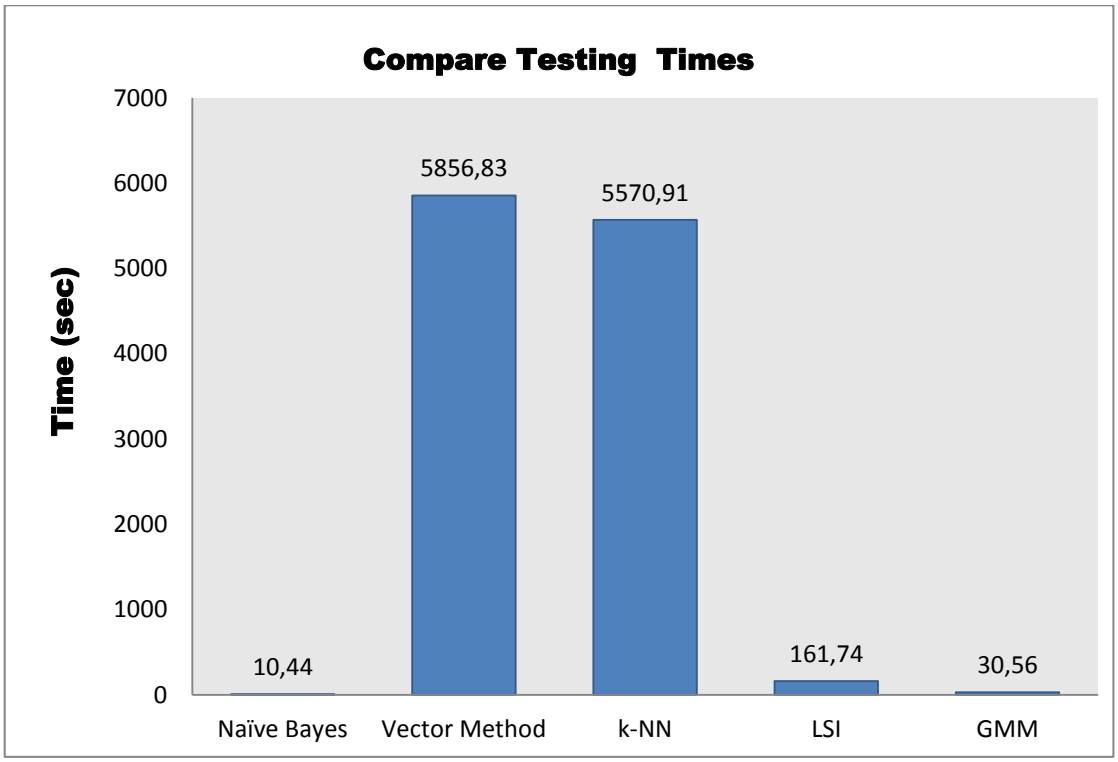


**Figure 6.10 –** Χρόνος κατηγοριοποίησης νέων κειμένων των πέντε μεθόδων για το dataset 20-Newsgroup

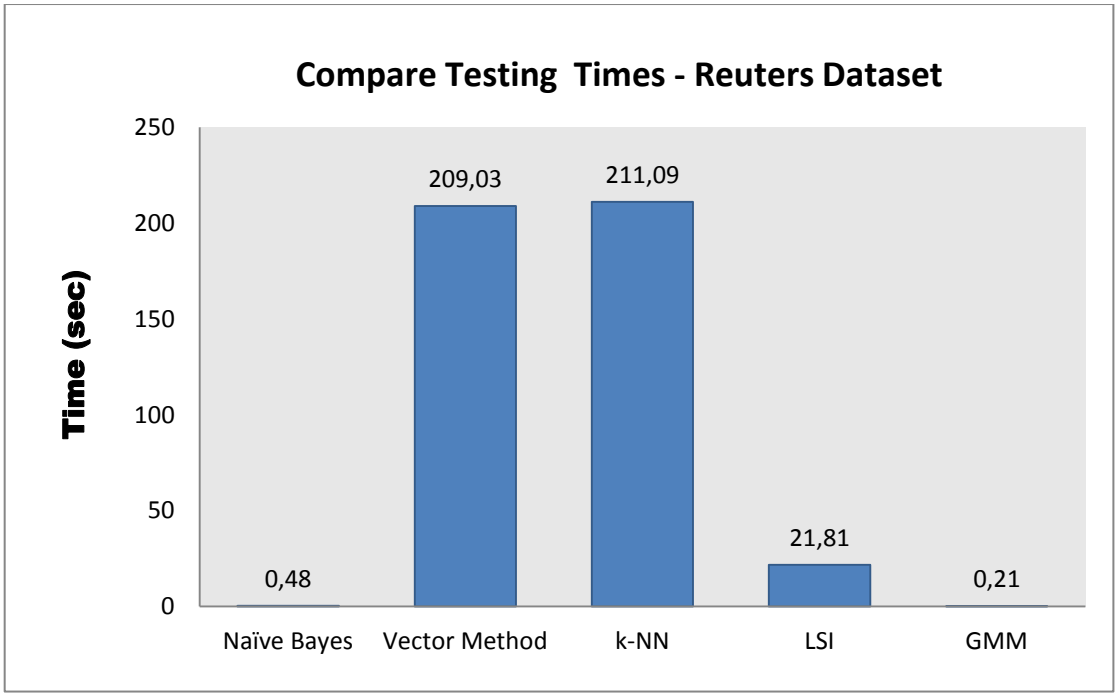


**Figure 6.11 –** Χρόνος κατηγοριοποίησης νέων κειμένων των πέντε μεθόδων για το dataset Reuters

*Έχουν επιλεγεί οι χρόνοι των LSI (100) και GMM (450) στις διαστάσεις εκείνες που δίνουν τα καλύτερα αποτελέσματα .*

Με μεγάλη διαφορά οι μέθοδοι Naïve Bayes και GMM είναι αποδοτικότεροι. Στο Dataset 20Newsgroups επειδή η βέλτιστη απόδοση εμφανίζεται σε διάσταση 450 τα GMM είναι ελαφρώς πιο αργά από τη μέθοδο Naïve Bayes (Figure 6.10). Αντίθετα στο Dataset Reuters που η βέλτιστη απόδοση εμφανίζεται σε διάσταση 30 τα GMM είναι η ταχύτερη μέθοδος (έστω και με μικρή διαφορά) (Figure 6.11). Οι μέθοδοι Vector και K-NN έχουν χαρακτηριστικά τη χειρότερη απόδοση βάσει χρόνου κατηγοριοποίησης και αυτό είναι λογικό αφού τα διανύσματα δεν έχουν υποστεί μείωση διαστάσεων και το cosine similarity μεταξύ των κειμένων υπολογίζεται πάνω σε διανύσματα χιλιάδων όρων. Αντίστοιχα και η μέθοδος LSI έχει αρκετά υψηλοτέρα από άποψη χρόνο αποτελέσματα από τα GMM. Αυτό οφείλεται στο γεγονός ότι η LSI συγκρίνει με cosine similarity κάθε νέο κείμενο με όλα τα κείμενα της εκπαίδευσης κάνοντας έτσι χιλιάδες υπολογισμούς για κάθε νέο κείμενο προς κατηγοριοποίηση. Αντίθετα η GMM υπολογίζει τη μέγιστη πιθανοφάνεια μεταξύ όλως των GMM, δηλαδή κατηγοριών, κάνοντας έτσι πολύ λιγότερους υπολογισμούς. Στη παραπάνω παράδειγμα (20Newsgroup) υπολογίζεται η πιθανοφάνεια μεταξύ 20 των GMM που αντιστοιχούν στις 20 κατηγορίες, ενώ για παράδειγμα το LSI θα υπολόγιζε 11293 φορές το cosine similarity, όσα δηλαδή και τα κείμενα εκπαίδευσης.

Παρακάτω φαίνεται πόσο επηρεάζει ο αριθμός των διαστάσεων το χρόνο κατηγοριοποίησης τόσο για τις μεθόδους LSI και GMM (οι μόνες που χρησιμοποιούν μείωση διαστάσεων).

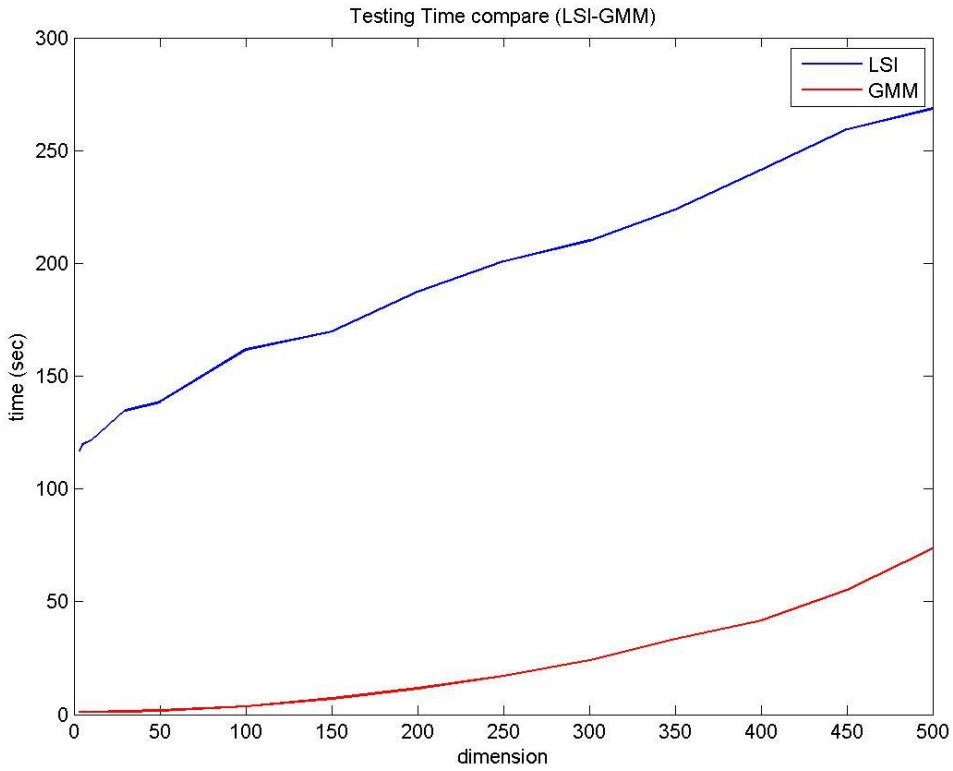


**Figure 6.12 –** Σύγκριση χρόνου κατηγοριοποίησης LSI – GMM σε όσο αυξάνονται διαστάσεις

Όπως είναι λογικό όσο μεγαλύτερες οι διαστάσεις τόσο αυξάνεται και ο χρόνος κατηγοριοποίησης .

### *6.4.6 Συνολική πίνακες αποτελεσμάτων*

Εδώ παρουσιάζονται οι πίνακες των αποτελεσμάτων για όλα τα dataset και για tf-idf βάρη. Φαίνεται πόσο επηρεάζει ο αριθμός των διαστάσεων καθώς και ο αριθμός των Gaussian την απόδοση του συστήματος.

**Webkb Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 69.05 | 76.07 | 76.79 | 74.21 | 75.43 | 77.15 | 77.01 | 76.43 | 76.14 | 74.35 | 72.92 | 68.27 | 57.23 | 45.13 |
| **2** | 68.48 | 76.07 | 75.43 | 74.00 | 75.43 | 76.29 | **77.44** | 76.07 | 76.29 | 75.00 | 72.64 | 67.69 | 57.23 | 45.13 |
| **5** | 69.05 | 77.00 | 75.36 | 74.21 | 75.14 | 76.43 | 76.65 | 76.07 | 75.79 | 74.93 | 72.49 | 66.90 | 56.80 | 45.06 |
| **10** | 69.48 | 76.58 | 76.50 | 72.92 | 74.50 | 76.43 | 75.86 | 76.22 | 75.72 | 74.07 | 72.49 | 65.69 | 55.23 | 43.34 |
| **20** | 69.98 | 77.50 | 77.00 | 70.41 | 72.13 | 75.93 | 76.00 | 75.07 | 75.29 | 74.50 | 71.35 | 64.90 | 51.86 | 42.62 |
| **30** | 69.41 | 77.65 | 78.43 | 71.27 | 72.21 | 75.29 | 75.93 | 75.21 | 75.43 | 73.78 | 70.84 | 61.17 | 50.36 | 41.55 |

**Πίνακας 6.2 –** Πίνακας αποτελεσμάτων για το Webkb Dataset με tf-idf βάρη

**20Newsgroup Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 32.99 | 43.09 | 51.63 | 65.03 | 67.90 | 70.89 | 71.75 | 72.86 | 73.59 | 75.69 | 78.97 | 82.51 | 84.55 | 77.04 |
| **2** | 34.03 | 44.85 | 52.21 | 64.57 | 67.20 | 71.05 | 72.06 | 72.59 | 73.49 | 75.86 | 79.94 | 83.61 | **86.09** | 75.04 |
| **5** | 34.72 | 45.09 | 53.60 | 64.51 | 68.17 | 71.24 | 72.41 | 72.96 | 73.56 | 76.07 | 79.14 | 80.88 | 83.83 | 72.71 |
| **10** | 34.44 | 45.16 | 53.06 | 65.06 | 67.57 | 70.90 | 72.15 | 73.24 | 73.64 | 76.20 | 79.70 | 82.86 | 81.57 | 72.42 |
| **20** | 35.02 | 45.19 | 53.32 | 65.95 | 67.64 | 70.86 | 71.87 | 72.67 | 73.31 | 76.48 | 79.11 | 81.53 | 80.64 | 58.78 |
| **30** | 34.91 | 45.74 | 53.41 | 66.00 | 67.86 | 70.28 | 71.66 | 71.91 | 73.07 | 76.30 | 78.62 | 82.45 | 81.42 | 53.68 |

**Πίνακας 6.3 –** Πίνακας αποτελεσμάτων για το 20-Newsgroup Dataset με tf-idf βάρη

**Reuters R8 Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 67.01 | 75.28 | 84.14 | 91.59 | 91.82 | 91.73 | 90.54 | 88.48 | 85.01 | 81.54 | 80.08 | 78.11 | 74.96 | 70.30 |
| **2** | 65.55 | 75.78 | 83.18 | 91.41 | 91.82 | 91.89 | 91.00 | 88.80 | 85.19 | 81.40 | 80.08 | 78.20 | 75.05 | 70.39 |
| **5** | 67.70 | 74.78 | 80.81 | 91.36 | 91.73 | 91.50 | 90.45 | 88.25 | 84.78 | 80.90 | 79.89 | 78.30 | 75.14 | 70.07 |
| **10** | 74.64 | 75.83 | 82.73 | 91.41 | **91.96** | 91.73 | 89.95 | 87.52 | 84.19 | 80.76 | 80.17 | 78.02 | 74.78 | 69.89 |
| **20** | 73.27 | 75.19 | 83.18 | 90.72 | 91.68 | 90.77 | 89.03 | 86.34 | 83.05 | 80.53 | 79.89 | 77.25 | 74.32 | 69.71 |
| **30** | 74.14 | 73.50 | 83.69 | 89.81 | 90.72 | 90.31 | 88.16 | 85.70 | 81.91 | 80.67 | 79.44 | 76.65 | 73.68 | 69.02 |

**Πίνακας 6.4** – Πίνακας αποτελεσμάτων για το Reuters Dataset με tf-idf βάρη

**Cade Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 7.43 | 9.91 | 24.23 | 45.97 | 48.62 | 49.51 | 50.42 | 50.66 | **51.21** | 51.05 | 50.26 | 48.83 | 45.83 | 41.54 |
| **2** | 6.25 | 8.05 | 23.44 | 45.26 | 48.63 | 49.25 | 49.85 | 50.64 | 50.09 | 51.02 | 49.39 | 48.85 | 45.91 | 41.57 |
| **5** | 6.90 | 9.75 | 22.94 | 45.35 | 48.37 | 49.15 | 50.01 | 50.52 | 50.80 | 50.37 | 50.20 | 48.59 | 45.75 | 41.25 |
| **10** | 9.04 | 9.91 | 23.79 | 43.71 | 48.21 | 48.43 | 49.91 | 49.23 | 50.45 | 50.62 | 49.75 | 47.95 | 45.18 | 40.97 |
| **20** | 8.94 | 12.91 | 24.66 | 42.94 | 46.32 | 46.74 | 47.87 | 50.11 | 50.04 | 49.61 | 49.33 | 47.55 | 44.77 | 40.49 |
| **30** | 10.32 | 12.44 | 25.09 | 44.06 | 45.50 | 47.82 | 47.88 | 49.28 | 49.96 | 49.03 | 48.33 | 47.09 | 44.44 | 40.15 |

**Πίνακας 6.5** – Πίνακας αποτελεσμάτων για το Cade Dataset με tf-idf βάρη

Έπειτα παρουσιάζονται οι πίνακες των αποτελεσμάτων για όλα τα dataset για δυαδικά βάρη.

**Webkb Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 50.57 | 51.64 | 54.87 | 63.32 | **64.54** | 54.29 | 55.30 | 55.37 | 55.58 | 57.37 | 56.73 | 51.43 | 47.85 | 42.19 |
| **2** | 53.58 | 56.37 | 59.45 | 61.74 | 63.46 | 55.22 | 55.58 | 54.87 | 55.87 | 56.30 | 55.22 | 51.50 | 47.35 | 42.12 |
| **5** | 54.37 | 57.66 | 61.31 | 61.39 | 59.95 | 54.22 | 55.37 | 54.94 | 55.65 | 57.23 | 55.01 | 51.64 | 47.63 | 42.26 |
| **10** | 55.87 | 60.17 | 62.17 | 62.75 | 50.86 | 54.15 | 53.79 | 54.87 | 55.58 | 56.37 | 54.79 | 50.86 | 46.34 | 41.90 |
| **20** | 54.65 | 59.67 | 63.61 | 56.30 | 51.43 | 52.43 | 54.08 | 54.08 | 55.01 | 56.66 | 54.87 | 50.35 | 45.91 | 41.40 |
| **30** | 53.58 | 59.74 | 63.25 | 50.57 | 50.71 | 53.29 | 53.94 | 54.51 | 55.30 | 56.73 | 53.94 | 50.14 | 45.20 | 40.54 |

**Πίνακας 6.6 –** Πίνακας αποτελεσμάτων για το Webkb Dataset με δυαδικά βάρη

**20Newsgroup Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 9.95 | 12.15 | 19.77 | 32.74 | 37.91 | 44.32 | 47.91 | 50.10 | **51.70** | 51.54 | 49.85 | 48.68 | 43.21 | 36.82 |
| **2** | 10.74 | 12.80 | 21.37 | 32.69 | 38.04 | 44.31 | 47.95 | 50.06 | 51.35 | 50.93 | 49.62 | 47.52 | 44.02 | 34.88 |
| **5** | 10.62 | 12.58 | 22.54 | 30.88 | 38.32 | 43.96 | 47.76 | 49.48 | 50.58 | 49.90 | 50.45 | 47.37 | 43.10 | 37.72 |
| **10** | 11.78 | 13.93 | 23.71 | 31.05 | 38.61 | 44.15 | 47.86 | 49.28 | 50.90 | 50.45 | 50.19 | 46.38 | 41.57 | 35.18 |
| **20** | 11.33 | 14.69 | 22.86 | 32.47 | 39.36 | 44.02 | 47.45 | 49.16 | 49.78 | 49.30 | 49.45 | 44.62 | 41.07 | 36.05 |
| **30** | 11.42 | 13.45 | 22.58 | 34.03 | 40.40 | 44.15 | 47.47 | 49.25 | 50.01 | 49.85 | 48.12 | 45.59 | 42.33 | 32.22 |

**Πίνακας 6.7 –** Πίνακας αποτελεσμάτων για το 20-Newsgroup Dataset με δυαδικά βάρη

**Reuters R8 Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 60.25 | 70.26 | 78.57 | 90.27 | 91.00 | 89.17 | 86.02 | 81.31 | 78.80 | 78.39 | 77.25 | 75.46 | 74.09 | 69.43 |
| **2** | 61.48 | 70.21 | 80.58 | 90.31 | 91.04 | 89.26 | 85.74 | 81.27 | 78.84 | 78.48 | 77.34 | 75.46 | 74.00 | 69.84 |
| **5** | 65.46 | 74.37 | 86.02 | 90.81 | 91.36 | 88.99 | 85.65 | 81.27 | 78.75 | 78.39 | 77.34 | 75.51 | 73.96 | 69.11 |
| **10** | 68.38 | 79.44 | 88.76 | 91.54 | 91.32 | 88.85 | 85.01 | 80.81 | 78.75 | 78.39 | 77.29 | 75.42 | 73.86 | 69.21 |
| **20** | 72.68 | 83.46 | 89.72 | **91.91** | 90.95 | 87.94 | 84.01 | 79.80 | 78.71 | 78.25 | 76.97 | 75.46 | 73.36 | 68.98 |
| **30** | 75.10 | 84.69 | 89.76 | 91.86 | 90.49 | 87.30 | 83.05 | 79.03 | 78.66 | 77.98 | 76.83 | 75.10 | 73.18 | 67.83 |

**Πίνακας 6.8 –** Πίνακας αποτελεσμάτων για το Reuters Dataset με δυαδικά βάρη

**Cade Dataset – Accuracy (%)**

| Dimensions/ Num_Gaussian | 3 | 5 | 10 | 30 | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 3.42 | 4.01 | 7.46 | 18.54 | 29.74 | **33.92** | 33.27 | 32.90 | 32.29 | 32.20 | 31.88 | 31.11 | 30.67 | 28.73 |
| **2** | 3.38 | 4.14 | 8.87 | 21.65 | 29.14 | 33.75 | 33.26 | 32.84 | 32.26 | 32.28 | 31.76 | 31.15 | 30.78 | 29.06 |
| **5** | 3.96 | 5.43 | 11.71 | 28.04 | 29.47 | 33.33 | 33.32 | 32.75 | 31.96 | 32.17 | 31.64 | 30.89 | 30.42 | 28.29 |
| **10** | 6.77 | 10.94 | 21.92 | 27.61 | 29.19 | 32.85 | 32.81 | 32.32 | 31.68 | 31.68 | 31.30 | 30.50 | 30.07 | 28.53 |
| **20** | 10.96 | 17.72 | 20.61 | 28.61 | 29.22 | 31.89 | 31.80 | 31.43 | 30.84 | 31.00 | 30.76 | 29.92 | 29.58 | 27.55 |
| **30** | 15.02 | 19.64 | 22.58 | 28.22 | 29.20 | 30.91 | 30.93 | 30.71 | 30.26 | 30.30 | 30.1 | 29.44 | 29.43 | 27.53 |

**Πίνακας 6.9 –** Πίνακας αποτελεσμάτων για το Cade Dataset με δυαδικά βάρη

Από τους παραπάνω συγκεντρωτικούς πίνακες φαίνεται και ποσοτικά ότι έδειξαν και τα διαγράμματα στις προηγούμενες ενότητες. Αρχικά φαίνεται ξεκάθαρα σε όλες τις περιπτώσεις ότι τα tf-idf βάρη παρουσιάζουν εμφανώς καλύτερα αποτελέσματα από τα δυαδικά σε όλες τις περιπτώσεις. Έπειτα μπορούμε να παρατηρήσουμε ότι το μέγεθος των διαστάσεων των διανυσμάτων παίζει πολύ μεγάλο ρόλο στην απόδοση του κατηγοριοποιητή. Σε όλες τις περιπτώσεις αρχικά στις πολύ λίγες διαστάσεις όπου υπάρχει απώλεια πληροφορίας η απόδοση κυμαίνεται σε πολύ χαμηλά επίπεδα. Και στις μεγάλες διαστάσει επίσης τα αποτελέσματα παρουσιάζουν μείωση στην απόδοση τους (μη σημαντικοί όροι - θόρυβος). Τέλος, όπως φαίνεται ειδικά στα tf-idf βάρη το πλήθος των Gaussian επιδρά σημαντικά στην απόδοση του κατηγοριοποιητή. Με **bold** είναι οι βέλτιστες αποδόσεις σε κάθε περίπτωση.

## *6.5 Σύγκριση αποτελεσμάτων*

Τα συγκριτικά αποτελέσματα για κάθε κατηγορία παρουσιάζονται παρακάτω.

| | ***Naïve Bayes*** | **Vector Mecthod** | ***k-NN*** | **LSI** | **GMM** |
|---|---|---|---|---|---|
| **Webkb** | 0.835 | 0.644 | 0.725 | 0.735 | 0.784 |
| **20Ng** | 0.810 | 0.724 | 0.759 | 0.749 | 0.860 |
| **R8** | 0.960 | 0.788 | 0.852 | 0.941 | 0.919 |
| **Cade** | 0.572 | 0.414 | 0.512 | 0.432 | 0.512 |
| **Average** | 0.794 | 0.643 | 0.712 | 0.715 | 0.769 |

**Πίνακας 6.10 -** Συγκεντρωτικά αποτελέσματα για κάθε μέθοδο

Τα αποτελέσματα για μια πρώτη προσπάθεια κατηγοριοποίησης κειμένων με χρήση GMM είναι ιδιαίτερα ενθαρρυντικά. Η μέθοδος Naïve Bayes εξακολουθεί να έχει τα καλύτερα αποτελέσματα σχεδόν σε όλες τις περιπτώσεις (εκτός του dataset 20Newsgroups) αλλά παρόλα αυτά η κατηγοριοποίηση με χρήση GMM δίνει πολύ κοντά σε αυτή αποτελέσματα, έχοντας παράλληλα πολύ καλύτερη απόδοση σε σύγκριση με τις άλλες κλασσικές διανυσματικές μεθόδους.

Εδώ θα πρέπει να αναφερθεί ότι η συλλογή 20Newsgroups είναι η δυσκολότερη για κατηγοριοποίηση, αφού έχει τις περισσότερες κατηγορίες (20 διαφορετικές κατηγορίες) καθώς επίσης και οι κατηγορίες της σε πολλές περιπτώσεις έχουν κοινό θέμα. Συγκεκριμένα η συλλογή 20Newsgroups αποτελείται από 7 κύριες κατηγορίες οι οποίες χωρίζονται επιμέρους σε υποκατηγορίες δημιουργώντας έτσι τις τελικές 20. Εξαιτίας αυτού η συλλογή αυτή είναι πιο κοντά σε πρακτικές εφαρμογές, όπου χρειάζεται να διαχωριστούν κείμενα τα οποία αναφέρονται σε ένα γενικό κοινό θέμα σε πολλές διαφορετικές κατηγορίες . Επίσης όπως φαίνεται τα GMM δίνουν διαφορά της τάξης τους 5% καλύτερα αποτελέσματα από τη δεύτερη καλύτερη μέθοδο (Naïve Bayes) και 10% καλύτερα από τις υπόλοιπες τρεις μεθόδους.

Τέλος, τα περιθώρια βελτίωσης της μεθόδου είναι πάρα πολλά όπως περιγράφονται και στις μελλοντικές επεκτάσεις παρακάτω.

# 7

# *Επίλογος*

## *7.1 Σύνοψη και συμπεράσματα*

Τα αποτελέσματα όπως περιγράφηκαν παραπάνω είναι ιδιαίτερα ικανοποιητικά για μια πρώτη προσπάθεια κατηγοριοποίησης κειμένων με χρήση GMM. Παρόλο που η μέθοδος Naïve Bayes παρουσιάζει τα καλυτέρα (όχι όμως με μεγάλη διαφορά) στις περισσότερες περιπτώσεις αποτελέσματα σε σύγκριση με τις άλλες κλασσικές διανυσματικές μεθόδους η χρήση GMM δίνει εμφανώς καλυτέρα αποτελέσματα. Στην πιο κοντινή σε πραγματικά δεδομένα συλλογή (20Newsgroups) τα GMM παρουσιάζουν εμφανώς τα καλύτερα αποτελέσματα σε σύγκριση με τις υπόλοιπες μεθόδους.

Επίσης και από άποψη χρόνου κατηγοριοποίησης τα GMM συμπεριφέρονται εξαιρετικά καλά. Σε αυτό το κομμάτι σημαντικό ρόλο παίζει και οι διαστάσεις μετά τη χρήση του SVD . Όπως φάνηκε και παραπάνω για το dataset 20-Newsgroups που για τα βέλτιστα αποτελέσματα χρειάστηκαν 450 διαστάσεις τα GMM έχουν χρόνο πολύ κοντά στη μέθοδο naïve Bayes. Αντίθετα για το dataset Reuters που αρκούν 30 διαστάσεις τα GMM είναι ταχύτερα από όλες τις μεθόδους.

Οι δυνατότητες βελτιώσεις της μεθόδου, τέλος, είναι πάρα πολλές και είναι πολύ πιθανό να προκύψουν αρκετά καλύτερα αποτελέσματα εφαρμόζοντάς τες. Παρακάτω παρουσιάζονται οι μελλοντικές επεκτάσεις που είναι πιο πιθανό να βελτιώσουν την απόδοση της μεθόδου.

## *7.2 Μελλοντικές επεκτάσεις*

Το έργο που περιγράφεται στην παρούσα εργασία ανοίγει μια σειρά από ενδιαφέρουσες γραμμές της έρευνας για το μέλλον. Μερικές από τις πιο ενδιαφέρουσες πιθανές εξελίξεις για το έργο αυτό περιλαμβάνουν, αλλά δεν περιορίζονται σε αυτά, τα ακόλουθα:

- Πρώτο βήμα με σειρά εκτέλεσης τις εργασίας είναι το κομμάτι της προ-επεξεργασίας των δεδομένων (dataset). Σκοπός αυτού του βήματος είναι να "καθαριστούν" όσο το δυνατό περισσότερο τα κείμενα ώστε να απαλειφθούν οι όροι οι οποίοι δε δίνουν καμία πληροφορία για τη κατηγορία του κειμένου . Στην εργασία χρησιμοποιήθηκαν αρκετά βήματα προ-επεξεργασίας μέχρι να έρθουν τα κείμενα στη τελική τους μορφή . Θα μπορούσαν να γίνει όμως χρήση επιπλέον βημάτων, όπως πχ χρήση του λεξικού WordNet ( http://en.wikipedia.org/wiki/WordNet ) το οποίο περιέχει συνώνυμα λέξεων, συντακτικά στοιχεία των λέξεων κ.α. Επίσης για το κομμάτι του stemming χρησιμοποιείται ο αλγόριθμος του Porter, υπάρχουν όμως ακόμα πολύ επιπλέον τέτοιοι αλγόριθμοι που μπορεί να "κόβουν" λιγότερες λέξεις, ή να "αφήνουν" περισσότερες, για παράδειγμα οι αλγόριθμοι Lovins , Paice/Husk , Dawnson κ.α.

- Έπειτα είναι το κομμάτι στο οποίο μπορεί να γίνει επιπλέον έρευνα είναι αυτό της απεικόνισης των κειμένων σαν διανύσματα. Αρχικά ένα σημείο το οποίο μπορεί να δώσει επιπλέον λύσεις είναι αυτό της δημιουργίας του λεξικού. Στην παρούσα εργασία το λεξικό δημιουργήθηκε βάσει των ξεχωριστών λέξεων που υπάρχουν σε κάθε συλλογή δεδομένων. Μια άλλη λύση θα ήταν να χρησιμοποιηθούν συνεχόμενες σειρές από λέξεις που εμφανίζονται (N-grams), ή ακόμα και ολόκληρες φράσεις.

- Επόμενο κομμάτι πάλι για την απεικόνιση των κειμένων ως διάνυσμα είναι ο τρόπος με των οποίων δίνονται τα βάρη στους όρους. Στην εργασία χρησιμοποιήθηκαν 2 τρόποι κατανομής των βαρών, τα δυαδικά βάρη (binary) και η μορφή tf-idf. Θα μπορούσε σε αυτό το σημείο να αναζητηθούν επιπλέον τρόποι για καλύτερη κατανομή των βαρών των όρων στα διανύσματα.

- Στο κομμάτι της μείωσης των διαστάσεων του πίνακα στην παρούσα εργασία χρησιμοποιείται η τεχνική Singular Value Decomposition (SVD) της αριθμητικής γραμμικής άλγεβρας γραμμικής. Υπάρχουν ωστόσο πολύ ακόμα αλγόριθμοι μείωσης διαστάσεων πίνακα που θα μπορούσαν να δοκιμαστούν ,όπως π.χ. ο NMF , MRMR κ.α., αν και μέχρι στιγμής ο SVD έχει φανεί να δίνει καλυτέρα αποτελέσματα.

- Επίσης θα μπορούσε να χρησιμοποιηθεί και συνδυασμός πινάκων εκτός του SVD ώστε να χρησιμοποιείται περισσότερη πληροφορία. Για παράδειγμα θα μπορούσε να χρησιμοποιηθεί SVD σε συνδυασμό με NMF ή ακόμα και χρήση ενός SVD για τις λέξεις (uni-grams) σε συνδυασμό με έναν ακόμα για το συνδυασμό λέξεων (bi-grams , tri-grams κ.τ.λ.)

- Τέλος αντί για Gaussian κατανομές θα μπορούσαν να δοκιμαστούν και άλλες συνεχείς κατανομές ή συνδυασμός με τη Gaussian κατανομή, όπως για παράδειγμα η Dirichlet κατανομή η οποία έχει επίσης εφαρμογές στη κατηγοριοποίηση κειμένων (Latent Dirichlet Allocation - LDA), και να χρησιμοποιηθεί για παράδειγμα το μοντέλο Dirichlet Process Gaussian Mixture Model (DPGMM) .

- Τέλος, σε όλη την εργασία θεωρούμε ότι κάθε κείμενο αντιστοιχεί σε μία και μόνο κατηγορία (single-label). Στη πραγματικότητα όμως πολλά κείμενα είναι ένα μίγμα πολλών θεμάτων και μπορεί να θεωρηθούν ότι εμπεριέχουν παραπάνω από ένα θέματα. Θα μπορούσε να μελετηθεί επιπλέον αν το μοντέλο που παρουσιάζεται στη παρούσα εργασία μπορεί να βοηθήσει στον εντοπισμό όλων των διαφορετικών θεμάτων που πραγματεύεται ένα κείμενο.

# *Βιβλιογραφία*


**1** M.Afify, O. Siohan and R. Sarikaya, 2007. *Gaussian Mixture Language Models for Speech Recognition,* ICASSP, Honolulu, Hawaii

**2** Wei Chen, 2007. *Building Language models on continuous space using gaussian mixture models.* Technical report, Carnegie Mellon University.

**3** R. Duda, P.Hart, and D. Stork, *Pattern Classification (Second Edition).* Wiley-Interscience, October 2000

**4** D. Oikonomidis, *2002. Language Models for Speech Recognition*. Technical University of Crete, MSc Thesis

**5** R. Sarikaya, M.Afify, Brian Kingsbury, *2007. Tied–Mixture Language Modeling in Continuous Space.* ICASSP, Honolulu, Hawaii.

**6** S. Theodoridis, K. Koutroumbas, 2009. *Pattern Recognition, 4th edition*. Elsevier Inc. AP.

**7** J. L. Fagan. *Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, US, 1987

**8** Salton, G. and Buckley, C. *Term-weighting approaches in automatic text retrieval*. Information Processing and Management, volume 24(5):pages 513–523, 1988

**9** Salton, G. and Lesk, M. *Computer evaluation of indexing and text processing.* Journal of the ACM, volume 15(1):pages 8–36, 1968

**10** Salton, G. *The SMART Retrieval System*. Prentice-Hall, Inc., New Jersey, USA,1971

**11** Yang, Y. Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. In Proceedings of the 17th Annual Inter- national ACM SIGIR Conference on Research and Development in Information Retrieval, pages 13–22

**12** Yang, Y. and Liu, X. A re-examination of text categorization methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, pages 42–49. ACM Press, New York, US, Berkeley, US, 1999.

**13** Lewis, D. D. Naive (Bayes) at forty: *The independence assumption in information retrieval*. In Proceedings of ECML-98, 10th European Conference on MachineLearning, pages 4–15. Springer Verlag, Heidelberg, DE, Chemnitz, DE, 1998.Published in the "Lecture Notes in Computer Science" series, number 1398.

**14** Furnas, G. W., Deerwester, S. C., Dumais, S. T., Landauer, T. K., Harshman, R. A.,Streeter, L., and Lochbaum, K. *Information retrieval using a singular value decomposition model of latent semantic structure*. In Proceedings of the 11thAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 465–480. Grassau, France, 1988.

**15** Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman,R.A. *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, volume 41(6):pages 391–407, 1990.

**16** Yang Y. *An evaluation of statistical approaches to text categorization*, Journal Information Retrieval, volume 1, number 1-2, pages 69-90, 1997

**17** S. Dumais, J. Platt, D. Heckerman, and M. Sahami. *Inductive learning algorithms and representations for text categorization*. In CIKM '98: Proceedings of the 7th international conference on Information and knowledge management, pages 148–155, 1998

**18** D.D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–50, 1992

**19** Cardoso Cachopo A: *Improving Methods for Single-label Text Categorization*. PhD Thesis, Instituto Superior Técnico, Portugal (2007) 6. Li, X., Roth,

**20** Masand, B., Linoff, G., and Waltz, D. *Classifying news stories using memory based reasoning*. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 59–65. ACM Press, Copenhagen, Denmark, 1992

**21** Creecy, R. M., Masand, B. M., Smith, S. J., and Waltz, D. L*. Trading MIPS and memory for knowledge engineering: classifying census returns on the Connection Machine*. Communications of the ACM, volume 39(1):pages 48–63, 1996.

**22** Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. M. *Text classification from labeled and unlabeled documents using em*. Machine Learning, volume 39(2/3):pages 103–134, 2000.

**23** M. F. Porter. *An algorithm for suffix stripping*. Program 14(3):130–137. 33, 529, 1980

**24** William B. Cavnar, and John M. Trenkle. *N-Gram-Based Text Categorization.* Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, page 161--175. (1994)

**25** J.Furnkranz. A study Using n-gram Features for Text Categorization. Technical report, Austrian Research Institute for Artificial Intilligence, 1998

**26** J Furnkran, T. Mitchelll, E.Riloff. *A Case Study in Using Linguistic Phrases for Text Categorization on the WWW*. In AAAI-98 workshop on Learning for Text Categorization, 1998

# *Appendix*

Στο παράρτημα αυτό παρουσιάζονται ενδεικτικά κείμενα από κάθε συλλογή αφότου έχει γίνει η προ-επεξεργασία τους. Είναι δηλαδή στη μορφή που χρησιμοποιούνται προκειμένου να αρχίσει η διαδικασία κατηγοριοποίησης. Πως γίνεται η προ-επεξεργασία των κειμένων έχει αναλυθεί πλήρως στο κεφάλαιο 4.

Η πρώτη λέξη κάθε κειμένου είναι η κατηγορία στην οποία ανήκει, η οποία χρησιμοποιείται ώστε να γίνει ο διαχωρισμός των κατηγοριών για την εκπαίδευση των GMM (κάθε GMM ανήκει σε μία κατηγορία) καθώς επίσης και για να υπολογιστεί η απόδοση. Προφανώς η λέξη-κατηγορία δε λαμβάνεται υπόψη στην αξιολόγηση αφού τα κείμενα θεωρούνται άγνωστα και χρησιμοποιείται μόνο για να γίνει το testing.

## *Webkb Dataset*

***Course*** ec advanc comput architectur credit parallel algorithm principl parallel detect vector compil interconnect network simd mimd machin processor synchron data coher multi dataflow machin special purpos processor prerequisit ec consent instructor inform info fall

***Student*** lin resum java current address mapl avenu apt ithaca tel email cornel perman address ali taipei Taiwan

***Faculty*** associ dean professor univers egypt polytechn univers york north carolina state univers director vlsi design laboratori digit system laboratori research interest fault detect diagnosi design testabl vlsi design back faculti page mail

***Project*** mail mcgill offic mceng geometri lab phone page construct doesn cute construct icon faculti sue student octob

# *Reuters Dataset*

***earn*** champion product approv stock split champion product inc board director approv two for stock split common share for sharehold record april compani board vote recommend sharehold annual meet april increas author capit stock mln mln share reuter

***ship*** agenc report ship wait panama canal panama canal commiss govern agenc daili oper report that backlog ship wait enter canal earli todai two dai expect due schedul transit end dai backlog averag wait time tomorrow super tanker regular vessel north end hr hr south end hr hr reuter

***acq*** cofab inc bui gulfex for undisclos amount cofab inc acquir gulfex inc houston base fabric custom high pressur process vessel for energi and petrochem industri cofab group compani manufactur special cool and lubric system for oil and ga petrochem util pulp and paper and marin industri reuter

***grain*** grain carload fall week grain carload total car week end februari pct previou week pct week year ago associ american railroad report grain mill product load week total car pct previou week pct week year earlier associ reuter

***crude*** diamond shamrock dia cut crude price diamond shamrock corp that effect todai had cut contract price for crude oil dlr barrel reduct bring post price for west texa intermedi dlr barrel copani price reduct todai made light fall oil product price and weak crude oil market compani spokeswoman diamond latest line oil compani that cut contract post price last two dai cite weak oil market reuter

***interest*** fed expect add temporari reserv feder reserv expect enter govern secur market add temporari reserv economist expect suppli reserv indirectli arrang two billion dlr custom repurchas agreement fed add reserv directli system repurchas feder fund averag pct yesterdai open pct and stai earli trade analyst rate upward pressur partli settlement billion dlr year treasuri note reuter

***money-fx*** monei market mln stg late bank england had provid monei market late assist mln stg thi bring bank total todai mln stg and compar latest forecast mln stg shortag system todai reuter

***trade*** appoint trade chief european commun commiss todai appoint chief spokesman hugo paemen top offici charg multilater trade negoti commiss spokesman paemen belgian offici had previous chief aid extern affair commission etienn davignon post sinc januari spokesman paemen will continu chief spokesman retir paul luyten charg depart handl negoti world trade bodi gatt oecd and forum reuter

# *20Newsgroups Dataset*

***alt.atheism*** bibl quiz answer articl healta saturn wwc edu healta saturn wwc edu tammi heali write cheribum ark coven god make graven imag refer idol creat worship ark coven wasn wrodhip and high priest enter holi holi onc year dai aton not familiar knowledg origin languag word for idol and that translat us word idol graven imag had origin idol you wrong suggest determin whether interpret you offer correct dean kaflowitz

***comp.graphics*** patch for sungk due number bug gk suno instal patch and patch appear work fine and fix number problem patch requir fix number annoi bug break applic recent revis patch idea scott sloan email cesw newcastl edu univers newcastl fax nsw Australia

***comp.os.ms-windows.misc*** mous jumpi mous articl apr cti rlister cti russel lister write rlister cti russel lister subject mous jumpi mous date fri apr gmt eckton uc byu edu sean

eckton write microsoft serial mous and mous switch vertic motion nice and smooth horizont motion bad sometim can click mous jump can move mous rel uniform motion and mous will move smoothli for bit jump move smoothli for bit jump thi time left inch thi crazi never had troubl mous solut microsoft vent steam thi problem time result roller insid mous becom dirti good collect grime solut simpl remov ball reveal two roller carefulli clean and ball denni

***comp.sys.ibm.pc.hardware*** port standard port address for port standard sort mail don read thi group mark tomlinson mark garden equinox gen

***comp.sys.mac.hardware*** centri font problem recent centri desk vast improv previou machin iisi encount problem font entri filemak databas look fine print previou mac system wierd space charact increas greatli caus line truncat plain and bold helvetica size increas charact space occur for size and style mixtur truetyp and fix size font exactli iisi thing work perfectli manag similar behaviour word appl adopt usual friendli approach and told call local dealer god idea pete edward depart comput scienc king colleg univers aberdeen tel aberdeen fax scotland email pedward csd abdn artifici flower piec plastic and metal crude fashion bear limit superfici resembl real flower credibl attempt match intern complex term form function behavior artifici intellig smart comput

***comp.windows.x*** server scanlin pad question almost port xfree piec displai hardwar run into snag somewhat commonplac send net feeler displai that interlac memori map bit displai server view world obtain xwd xwud exactli displai version framebuff impress that server scanlin that long bit experiment that problem that server pad line word boundri scanlin size buffer byte isn exactli divis chang defin mit server includ servermd defin bitmap scanlin pad defin log bitmap pad defin log byte per scanlin pad defin bitmap scanlin pad defin log bitmap pad defin log byte per scanlin pad not exactli solut server don pad scan line thi server built run thi displai pad byte boundri custom version xfree mach brian

***misc.forsale*** maxtor info need unix softwar for sale articl cup portal thad cup portal thad floryan write articl colinm cunew colinm max carleton colin mcfadyen write jumper set maxtor that that us info maxtor drive delet sinc you jumper betwwwn select drive address your second drive jumper thad you note that that us not unix note strang cross post not not exactli sort machin intend mount clone jumper correct choic left cross post effect sinc not

newsgroup read thi don email dnichol and uunet ceilidh dnichol dnichol ceilidh beartrack donald nichol don voic dai ev black hole god divid

***rec.autos*** auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write work ga solid adsorpt air con system for auto applic thi kind system energi for regener adsorb exhaust ga interest thi mail email follow thi thread discuss prospect thi technolog bite thi suppos work tim year ago demonstr cold air system us air call rovax unit work short come seal technolog today

***rec.motorcycles*** want polic offic answer post vike iastat edu dan sorenson write and cop heh attitud stop speak gui reciev verbal warn for mph laugh hei dan potenti cool stori stuff share detail never break don enlighten and enliven and live vicari waitin for that stori erc grandrapid usa vfr dod

***rec.sport.baseball*** rickei henderson articl apr msstate edu isi msstate edu jiann ming write bui henderson contract and bag groceri season you sign for noth that for bitch ball player doubt henderson clear waiver and instantli sign for major leagu minimum oakland pick remain million tab gm field perform valentine

***rec.sport.hockey*** jet fan hrivnak tabaracci hrivnak and tabarraci plai you prefer and tyler larter happen you answer will hrivnak choic obviou you tabaracci plai two start and relief effort for beaupr look mighti sharp don forget shutout goal period plai hrivnak give credit david poil for chang thi trade hopefulli tabaracci start isl tonight haven jinx frank salvator fmsalvat eo ncsu edu

***sci.crypt*** kei regist bodi naglec netcom nagl netcom john nagl write sinc law requir that wiretap request execut branch and approv judici branch clear that kei regist bodi control judici branch suggest suprem court region court appeal specif offic clerk that make sens half govern escrow eff admin secur not test arthur rubin rubin dsg dse beckman work beckman instrument brea mcimail compuserv arthur pnet ct person opinion and not repres employ

***sci.electronics*** voltag regul and current limit andi collin acollin uclink berkelei edu wrote not quick question throw for you gui for class project design and build power suppli spec voltatg adjust current limit voltag stai design for rippl rm decid kind circuit you initialpreciss limit current and allow temp drift case you can us ub transistor voltag refer vdc temp drift second case you us bandgap ref and opamp circuit detect maximum current output thi opamp control outputstag limit current bye thoma

***sci.med*** calcium deposit heart valv friend year man calcium deposit heart valv thi happen and can john greze execnet

***sci.space*** solid state tube analog articl cmu edu msu edu tom write electr guitar enthusiast type amp prefer and you tube type sinc tube lower distort and nois transistor your electr guitar type tube sound dude turn reverb gain add analog delai line and fuzz box wouldn notic distort forgot phase shifter transistor advantag wast heat and energi us heater cathod tube compar mechan system pat

***soc.religion.christian*** homosexu issu christiani articl atho rutger edu todd nickel laurentian write question interpret vers establish reason for not believ thi true base interpret scriptur come grip interpret homosexuy wrong vers can interpret read and surround text christ love bryan

***talk.politics.guns*** waco clinton press confer part notic question feder law violat brush law violat and evid origni batf warrant base

***talk.politics.mideast*** uva wow sad that univers virginia begun produc such virul breed jew hater and hate jew roar lion roar

***talk.politics.misc*** rnitedac and violenc articl ovg magpi linknet neal magpi linknet neal view experi polic offic larg metropolitan area and citizen peopl account for behavior and for behavior commun noth will improv wait minut agre you that peopl respons for behavior assum that you meant word account for behavior commun you talk and secondli lot troubl theori social behavior justic charg duti take respons for account for action person william decemb starr wdstarr athena mit edu

***talk.politics.misc*** clinton cave reduc job bill articl apr desir wright edu demon desir wright edu not boomer write clinton back billion job bill word pare core jobless benefit monei for creat full time job summer job monei chalk for hold line spend radio report overli optimist clinton cut billion for commun block grant keep summer job hmmm pressur congressperson brett noth passion vest interest disguis intellectu convict sean casei white plagu frank Herbert

# *Cade Dataset*

**01_servicos** br brasil minas fax cep av belo horizonte gerais joao assis chateaubriand bosco leopoldino mauriciol jbleopoldino

**02_sociedade** webmaster mail online queda direitos reservados visita noticias imagem ultimas pode contate aviao fonte codigo neves john pegar virus delegacia mata teofilo teofilo teofilo pimenta assassino australia elson otoni otoni otoni ataca lennon psiquiatra spawn libertado visitande

**03_lazer** home page novo endereco obrigado click entrar fernando preferencia anuncio soft publicado

**04_informatica** contato informatica informatica treinamento ltda bem vindo visitante site servicos servicos endereco informacoes empresa perguntas produtos rua sl fone fax clientes solucoes consultoria consultoria objetivo comercial curitiba principais parana fundada integrada integrada integrada incluindo oferecendo desembargador estrategico retorno gestao gestao gestao educacional aplicativos aplicativos amaral atuar parceiro otavio construtoras kosmos kosmos

**05_saude** site site site site site site ar favor favor favor administrador administrador administrador temporariamente temporariamente informe informe informe website arpor temporarily inform administrator temporalmente aire

**06_educacao** sao br paulo ribeiro nacional rua fax cep sp cx postal sede criancas pro tenente gomes alianca vl apec apec evangelizacao clementino

**07_internet** pagina web frames browser qualidade acesso usa navegador net suporta atualize provedor merece host tagline

**08_cultura** servicos mundo mundo estao estao veja informacoes cores cores estara gerais memorial memorial entrar curso oferecendo imigrante imigrante

**09_esportes** sao br br br site paulo novo www rua fax cep xx melhor homepage color visualizado desenvolvido brooklin ativa ativa cria cria cria netpoint easy bits dive cancioneiro easydive

**10_noticias** index index br modified size description parent directory www server apache port classiminas

**11_ciencias** tem visite novos menu facil seres consideramos inteligentes professor exercicios exercicios amar aprender fosse vladir vladir demorariamos

**12_compras_on_line** br web site novo informacoes xx melhor fale maiores atende maison vin maisonduvin maisonduvin