Οικονομικό Πανεπιστήμιο Αθηνών

Τμήμα Πληροφορικής

Πτυχιακή Εργασία

«Ανάπτυξη Συστήματος Χειρισμού Ερωτήσεων Ορισμού Προσώπων για Αρχεία Εφημερίδων»

Καρακατσιώτης Γιώργος

Α.Μ.: 3010058

Επιβλέπων: Ίων Ανδρουτσόπουλος

Αθήνα 2005

Περιεχόμενα

Περιεχόμενα **2**

Περίληψη **4**

1 Εισαγωγή **5**

1.1 Αντικείμενο και στόχοι της εργασίας 5

1.2 Διάρθρωση της εργασίας 8

1.3 Ευχαριστίες 8

2 Θεωρητικό Υπόβαθρο **9**

2.1 Σύστημα ερωταποκρίσεων 9

2.1.1 Κατηγορίες ερωτήσεων 9

2.1.2 Οι απαντήσεις σε ένα σύστημα ερωταποκρίσεων 10

2.1.3 Ο τρόπος που λειτουργεί ένα σύστημα ερωταποκρίσεων 11

2.2 Μηχανική μάθηση 14

2.2.1 Κατηγοριοποίηση με επιβλεπόμενη μηχανική μάθηση 14

2.2.2 Διανυσματική αναπαράσταση 14

2.2.3 Μηχανές Διανυσμάτων Υποστήριξης (SVM) 16

3 Το Σύστημα της Εργασίας **22**

3.1 Δημιουργία του σώματος εκπαίδευσης και αξιολόγησης 22

3.2 Επιλογή ιδιοτήτων 26

3.2.1 Οι βασικές ιδιότητες 27

3.2.2 Υποψήφιες ιδιότητες που αντιστοιχούν σε n-γράμματα 29

3.2.3 Αξιολόγηση υποψηφίων ιδιοτήτων 31

3.3 Υλοποιήσεις SVM 32

3.3.1 WEKA (Waikato Environment for Knowledge Analysis) 32

3.3.2 SVM^{light} 33

3.3.3 libSVM 34

3.4 Η on-line μορφή του συστήματός μας 35

4 Πειραματικά αποτελέσματα **37**

4.1 Μεθοδολογία 37

4.2 Πειράματα 38

4.2.1 Πειράματα εντοπισμού βέλτιστου πλήθους ιδιοτήτων 38

4.2.2 Ενδιαφέροντα παραδείγματα απαντήσεων 43

4.2.3 Πειράματα με μεταβλητό μέγεθος ερωτήσεων εκπαίδευσης 44

5 Μελλοντικές επεκτάσεις και βελτιώσεις.....47
Αναφορές.....49
Παράρτημα.....50

Περίληψη

Οι μελλοντικές μηχανές αναζήτησης (search engines) δε θα αρκούνται στο να επιστρέφουν στο χρήστη συνδέσμους (links) προς τα έγγραφα που περιέχουν πληροφορίες σχετικές με την αναζήτησή του. Θα φιλτράρουν τα έγγραφα και θα επιστρέφουν μόνο εκείνα τα αποσπάσματά τους που χρειάζονται για την απάντηση της ερώτησης που τους τέθηκε. Η εργασία αυτή προσπαθεί να αναπτύξει μία τέτοια μηχανή αναζήτησης για αρχεία ελληνικών εφημερίδων που θα απαντά σε ερωτήματα ορισμού προσώπων της μορφής «Ποιος είναι ο Δερτούζος;».

Η συλλογή εγγράφων στην οποία ψάχνει το σύστημα της εργασίας είναι το αρχείο της εφημερίδας "Το Βήμα", όπως διατίθεται μέσω του ιστοτόπου της εφημερίδας. Το σύστημα χρησιμοποιεί την υπάρχουσα μηχανή αναζήτησης του ιστοτόπου και από τα κείμενα που αυτή επιστρέφει προσπαθεί να εξαγάγει ένα απόσπασμα που να περιέχει ένα σύντομο ορισμό του προσώπου του ερωτήματος. Για τον εντοπισμό του καλύτερου αποσπάσματος χρησιμοποιείται μια Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine) και βελτιωμένες μορφές τεχνικών που αναπτύχθηκαν σε προηγούμενες εργασίες για το χειρισμό ερωτήσεων ορισμού σε αγγλικές συλλογές εγγράφων. Το σύστημα της εργασίας είναι δυνατόν να τροποποιηθεί εύκολα, ώστε να χρησιμοποιηθεί με τον αντίστοιχο ιστότοπο οποιασδήποτε άλλης εφημερίδας.

Τα πειραματικά αποτελέσματα της εργασίας δείχνουν ότι το σύστημα καταφέρνει να απαντήσει ικανοποιητικά περίπου του 80% των ερωτήσεων.

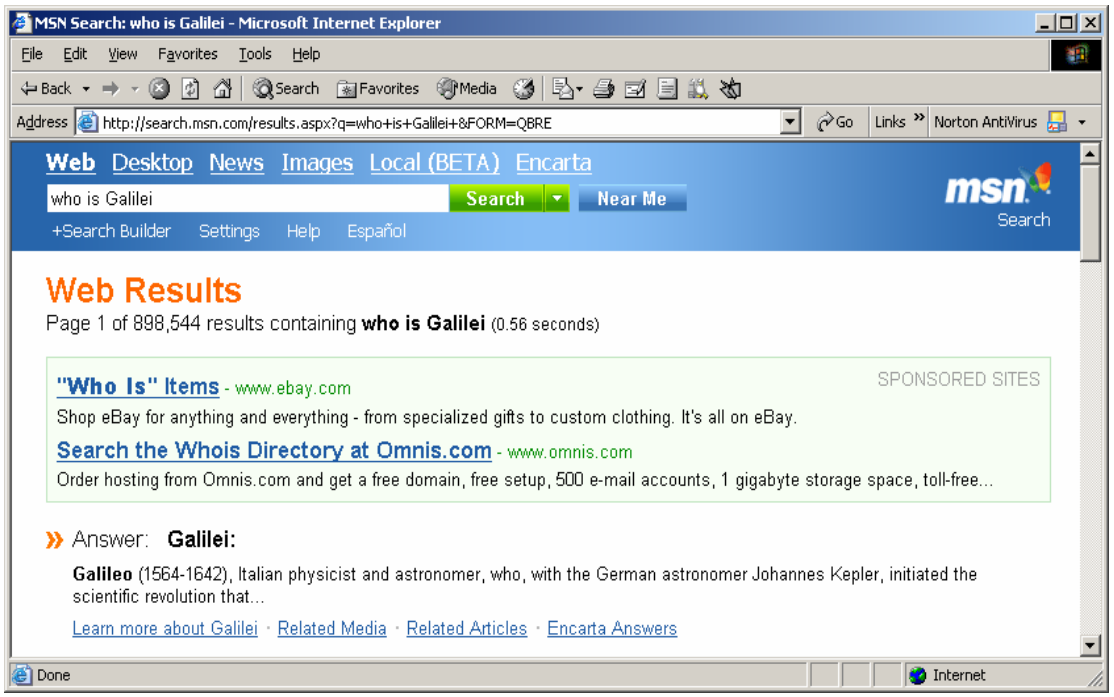
1 Εισαγωγή

'Η αρχή είναι το ήμισυ του παντός'

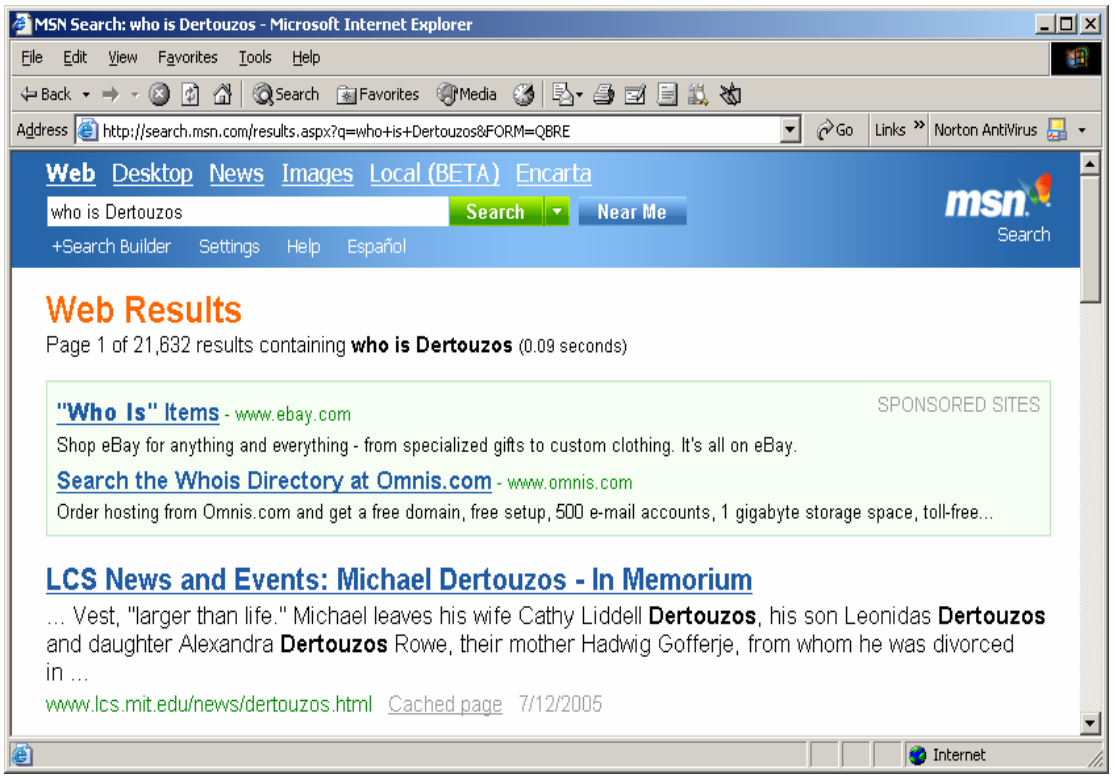
1.1 Αντικείμενο και στόχοι της εργασίας

Ολοένα και περισσότερες μηχανές αναζήτησης προσφέρουν στους χρήστες τους τη δυνατότητα ερωταποκρίσεων, δηλαδή δέχονται μια ερώτηση σε φυσική γλώσσα και επιστρέφουν μια συμβολοσειρά στην οποία περιέχεται η απάντηση (είτε απόσπασμα από κάποιο έγγραφο, είτε συνθέτουν με μηχανικό τρόπο την απάντηση). Ιδιαίτερα συχνές είναι οι ερωτήσεις ορισμού (π.χ. «Τι είναι η θαλασσαιμία;», «Ποιος ήταν ο Γαλιλαίος;») [1].

Στις περιπτώσεις ερωτήσεων ορισμού, η MSN search engine επιχειρεί να βρει τον επιθυμητό ορισμό στην εγκυκλοπαίδεια Encarta. Αν όμως δεν τον βρει εκεί, κάτι που είναι δυνατόν να συμβεί για παράδειγμα με ονόματα της επικαιρότητας που δεν βρίσκονται σε εγκυκλοπαίδειες, επιστρέφει απλά συνδέσμους προς ιστοσελίδες που περιέχουν τον όρο της ερώτησης και αποσπάσματά τους (βλ. παρακάτω παραδείγματα), χωρίς, όπως φαίνεται, να λαμβάνει υπόψη της ότι πρόκειται για ερώτηση ορισμού.



Η επιτυχημένη απάντηση που επιστρέφει η MSN search engine στην ερώτηση "Ποιος είναι ο Γαλιλαίος".



Η αποτυχημένη απάντηση που επιστρέφει η MSN search engine στην ερώτηση "Ποιος είναι ο Δερτούζος".

Η μηχανή αναζήτησης Google παρέχει μια αντίστοιχη δυνατότητα για ερωτήσεις ορισμού, που ενεργοποιείται όταν ο χρήστης εισάγει ερωτήματα της μορφής "define:

Dertouzos". Στην περίπτωση αυτή, η αναζήτηση φαίνεται πως γίνεται σε ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια που διατίθενται στον Ιστό. Δημιουργείται και πάλι, όμως, πρόβλημα, με ερωτήσεις που ζητούν να οριστούν όροι που δεν περιλαμβάνονται σε εγκυκλοπαίδειες και γλωσσάρια.

Στην εργασία αυτή επιχειρούμε να εντοπίσουμε ορισμούς σε κείμενα εφημερίδων, κάτι που είναι χρήσιμο σε περιπτώσεις που δε βρεθεί ο ζητούμενος ορισμός σε εγκυκλοπαίδειες και λεξικά. Δεν ασχοληθήκαμε με το σύνολο όλων των ερωτήσεων ορισμού, αλλά με ένα σημαντικό υποσύνολό τους, αυτό των ερωτήσεων που ζητούν να οριστούν πρόσωπα (π.χ. «Ποιος είναι ο Δερτούζος;»). Στηριχθήκαμε στα αποτελέσματα της πτυχιακής εργασίας της Σπ. Μηλιαράκη [2], που ανέπτυξε τεχνικές χειρισμού ερωτήσεων ορισμού για αγγλικές συλλογές κειμένων, και προσπαθήσαμε να τα προσαρμόσουμε για την ελληνική γλώσσα και κείμενα εφημερίδων που ανακτώνται από τον ιστότοπο μιας εφημερίδας. Πιο συγκεκριμένα, αναπτύξαμε ένα σύστημα που ψάχνει για ορισμούς προσώπων στο αρχείο της εφημερίδας "Το Βήμα", όπως διατίθεται μέσω του ιστοτόπου της εφημερίδας. Το σύστημα χρησιμοποιεί την υπάρχουσα μηχανή αναζήτησης του ιστοτόπου και από τα κείμενα που αυτή επιστρέφει προσπαθεί να εξαγάγει ένα απόσπασμα που να περιέχει ένα σύντομο ορισμό του προσώπου του ερωτήματος. Για τον εντοπισμό του καλύτερου αποσπάσματος χρησιμοποιείται μια Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine, SVM) που επιχειρεί να κατατάξει τα αποσπάσματα σε δύο κατηγορίες, εκείνα που πραγματικά αποτελούν απάντηση στο ερώτημα, δηλαδή περιέχουν αποδεκτό ορισμό του ζητούμενου όρου, και σε εκείνα που δεν αποτελούν αποδεκτή απάντηση. Για κάθε απόσπασμα, η απόκριση του SVM συνοδεύεται και από ένα ποσοστό βεβαιότητας, που δείχνει πόσο σίγουρο είναι το SVM ότι το απόσπασμα ανήκει στην κατηγορία στην οποία το κατέταξε. Το σύστημα επιστρέφει τελικά το απόσπασμα που κατετάγη στην κατηγορία των αποδεκτών απαντήσεων με τη μεγαλύτερη βεβαιότητα.

Τα αποτελέσματα της εργασίας είναι ιδιαίτερα ενθαρρυντικά, αφού δείχνουν ότι το σύστημα καταφέρνει να απαντήσει ικανοποιητικά περίπου του 80% των ερωτήσεων.

1.2 Διάρθρωση της εργασίας

Η υπόλοιπη εργασία είναι χωρισμένη σε 4 κεφάλαια:

- Το Κεφάλαιο 2 παρέχει το θεωρητικό υπόβαθρο στο οποίο στηρίζεται η εργασία και χωρίζεται σε 2 μέρη. Στο πρώτο μέρος δίνονται πληροφορίες σχετικά με το τι είναι ένα σύστημα ερωταποκρίσεων, ενώ στο δεύτερο παρουσιάζεται η μαθηματική θεμελίωση των SVMs.
- Στο Κεφάλαιο 3 περιγράφεται ο τρόπος με τον οποίο δημιουργήθηκαν τα δεδομένα εκπαίδευσης και αξιολόγησης της μεθόδου που υλοποιήσαμε, καθώς και ο τρόπος με τον οποίο επιλέγονται οι ιδιότητες (attributes) που χρησιμοποιεί το SVM. Επίσης παρουσιάζονται διάφορες βιβλιοθήκες κώδικα που χρησιμοποιήσαμε για τη δημιουργία του συστήματός μας.
- Στο Κεφάλαιο 4 αναλύονται διεξοδικά τα πειράματα της εργασίας και τα αποτελέσματά τους.
- Στο Κεφάλαιο 5 παρουσιάζονται πιθανές μελλοντικές επεκτάσεις και βελτιώσεις του συστήματος, σύμφωνα με τα συμπεράσματα που προέκυψαν από την εργασία.

1.3 Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον υπεύθυνο καθηγητή της εργασίας κ. Ίωνα Ανδρουτσόπουλο, για την πολύτιμη καθοδήγησή του. Όσες φορές η εργασία αντιμετώπιζε προβλήματα ήταν πάντα παρών και με την εμπειρία του βοήθησε στο να ξεπερασθούν τα όποια εμπόδια. Επίσης θα ήθελα να ευχαριστήσω τον κ. Θεόδωρο Καλαμπούκη για κάποιες ιδέες που μου έδωσε στα πλαίσια της υλοποίησης της μονάδας αποκοπής καταλήξεων (stemmer) καθώς και για τη γόνιμη ανταλλαγή ιδεών πάνω στα θέματα ανάκτησης πληροφοριών. Ακόμα θα ήθελα να ευχαριστήσω όλα τα παιδιά στα εκπαιδευτικά εργαστήρια του Ο.Π.Α. που με ανέχτηκαν (κυρίως την περίοδο των πειραμάτων) και ιδιαίτερα τους συμφοιτητές μου Μακίδη Μιχάλη και Μακούδη Γεωργία.

2 Θεωρητικό Υπόβαθρο

'The important thing is not to stop questioning.'
(Albert Einstein)

'I' m sorry, Dave. I' m afraid I can't do that…'
(HAL 9000, 2001: A Space Odyssey)

Στο κεφάλαιο αυτό θα προσπαθήσουμε να δώσουμε τις βάσεις που απαιτούνται για την καλύτερη κατανόηση της εργασίας. Αρχικά θα αναλυθούν τα συστήματα ερωταποκρίσεων (τι είδη ερωτήσεων υπάρχουν, τι είναι ένα σύστημα ερωταποκρίσεων) και έπειτα θα αναλυθεί ο τρόπος με τον οποίο λειτουργούν τα SVMs και με ποιον τρόπο χρησιμοποιούνται στην παρούσα εργασία.

2.1 Σύστημα ερωταποκρίσεων

2.1.1 Κατηγορίες ερωτήσεων

Οι πιο συνηθισμένες ερωτήσεις προς ένα σύστημα ερωταποκρίσεων μπορούν να χωριστούν σε 3 βασικές κατηγορίες με βάση την απάντηση που επιδέχονται:

- ερωτήσεις με καθορισμένη απάντηση (factual question), που περιλαμβάνουν με τη σειρά τους τις ακόλουθες πιο συνηθισμένες υποκατηγορίες:
    - τοποθεσίας: π.χ. «Πού γεννήθηκε ο Όμηρος;»
    - χρόνου: π.χ. «Πότε περπάτησε ο πρώτος άνθρωπος στο φεγγάρι;»
    - ονόματος προσώπου: π.χ. «Ποιος έφτιαξε τον πρώτο ιό για τους υπολογιστές;»
    - ποσότητας: π.χ. «Πόσο ζυγίζει ένας μέσος άνθρωπος στο φεγγάρι;»
    - ορισμού: π.χ. «Τι είναι υπολογιστής;»
- ερωτήσεις γνώμης (opinion question): π.χ. «Ποια θα έπρεπε να είναι η στάση των δισκογραφικών εταιριών απέναντι στα peer-to-peer προγράμματα;»
- ερωτήσεις περίληψης (summary question): π.χ. «Ποια είναι η πλοκή του βιβλίου "The Hitchhiker's guide to the Galaxy";»

Στην εργασία αυτή ασχοληθήκαμε μόνο με τις ερωτήσεις ορισμού και συγκεκριμένα με τις ερωτήσεις που ζητούν να οριστεί ένα πρόσωπο. Η επιθυμητή απάντηση είναι ένας σύντομος ορισμός, που συνήθως περιλαμβάνει την επαγγελματική ιδιότητα του προσώπου ή κάποια άλλη ιδιότητά του για την οποία είναι γνωστό. Η γενική μορφή αυτών των ερωτήσεων είναι: «Ποιος είναι *άρθρο όνομα_οντότητας*;». Για παράδειγμα «Ποιος είναι ο Bill Gates;».

2.1.2 Οι απαντήσεις σε ένα σύστημα ερωταποκρίσεων

Οι απαντήσεις των συστημάτων ερωταποκρίσεων δεν πρέπει να είναι ολόκληρα έγγραφα, αλλά περιορισμένες σε μήκος συμβολοσειρές, συνήθως 50 ή 250 χαρακτήρων. Επίσης, είναι αρκετά συνηθισμένο το σύστημα να επιτρέπεται να επιστρέψει μόνο μία ή το πολύ μέχρι 5 απαντήσεις (συμβολοσειρές) ανά ερώτηση· στη δεύτερη περίπτωση, η απόκριση του συστήματος θεωρείται σωστή αν τουλάχιστον μία από τις 5 απαντήσεις είναι αποδεκτή. Στην παρούσα εργασία οι απαντήσεις περιορίζονται στους 250 χαρακτήρες, ενώ για την αξιολόγηση των αποκρίσεων του συστήματος χρησιμοποιούνται δύο μέτρα, ένα το οποίο λαμβάνει υπόψη του τις 5 κορυφαίες απαντήσεις του συστήματος, δηλαδή τις 5 απαντήσεις που το σύστημα θεωρεί πιθανότερο να αποτελούν αποδεκτές απαντήσεις, και ένα δεύτερο το οποίο λαμβάνει υπόψη του μόνο την κορυφαία απάντηση του συστήματος.

Εν γένει οι απαντήσεις ενός συστήματος ερωταποκρίσεων μπορεί να είναι είτε αποσπάσματα από τα κείμενα από τα οποία εξήχθησαν, είτε να έχουν παραχθεί από το σύστημα με τη βοήθεια τεχνικών παραγωγής φυσικής γλώσσας, κάνοντας ακόμα και συρραφή κειμένων από διαφορετικά έγγραφα. Στην παρούσα εργασία οι απαντήσεις που επιστρέφονται είναι αποσπάσματα από τα κείμενα των εφημερίδων και το κάθε απόσπασμα προέρχεται από ένα μόνο κείμενο.

Παράδειγμα

Ερώτηση: Ποιος είναι ο Βαϊσμίλερ;

Λανθασμένη απάντηση: *ίωσης στην ουσία) των ιδιωτικών σχολείων. Η ζούγκλα του μαυροπίνακα χρειάζεται επειγόντως έναν «Ταρζάν» - τύπου Βαϊσμίλερ και όχι*

Μάκη - να βάλει τέλος στη ζωώδη κατάσταση που επικρατεί. Θα ήταν εκτός τόπου και χρόνου βέβαια όποιος θα έκανε έκκλησ

Ορθή απάντηση: *ς σήμερα τον θρυλικό «Λόρδο των πιθήκων» σε περισσότερες από 40 ταινίες, δεν είναι άλλος από τον ολυμπιονίκη Τζόνι Βαισμίλερ που εγκατέλειψε το 1932 τις πισίνες για να βουτήξει μαζί με την αγαπημένη του Τζέιν (Μορίν Ο' Σάλιβαν) στην άγρια ζούγκλα της*

Η δεύτερη απάντηση παραπάνω θεωρείται σωστή επειδή αναφέρει πως ο Βαϊσμίλερ ήταν ολυμπιονίκης, αν και η απάντηση θα μπορούσε να θεωρηθεί ημιτελής γιατί δεν κάνει απολύτως σαφές πως πρόκειται επίσης για τον ηθοποιό που υποδύθηκε τον Ταρζάν, μια δεύτερη ιδιότητα για την οποία είναι γνωστό το συγκεκριμένο πρόσωπο. Προφανώς η διάκριση μεταξύ ορθών και λανθασμένων απαντήσεων δεν είναι πάντα εύκολη και ενδέχεται να διαφέρει ανάλογα με τον κριτή των απαντήσεων.

2.1.3 Ο τρόπος που λειτουργεί ένα σύστημα ερωταποκρίσεων

Το παρακάτω διάγραμμα αναπαριστά την τυπική αρχιτεκτονική ενός συστήματος ερωταποκρίσεων, στην οποία και βασίσθηκε το σύστημα της εργασίας.

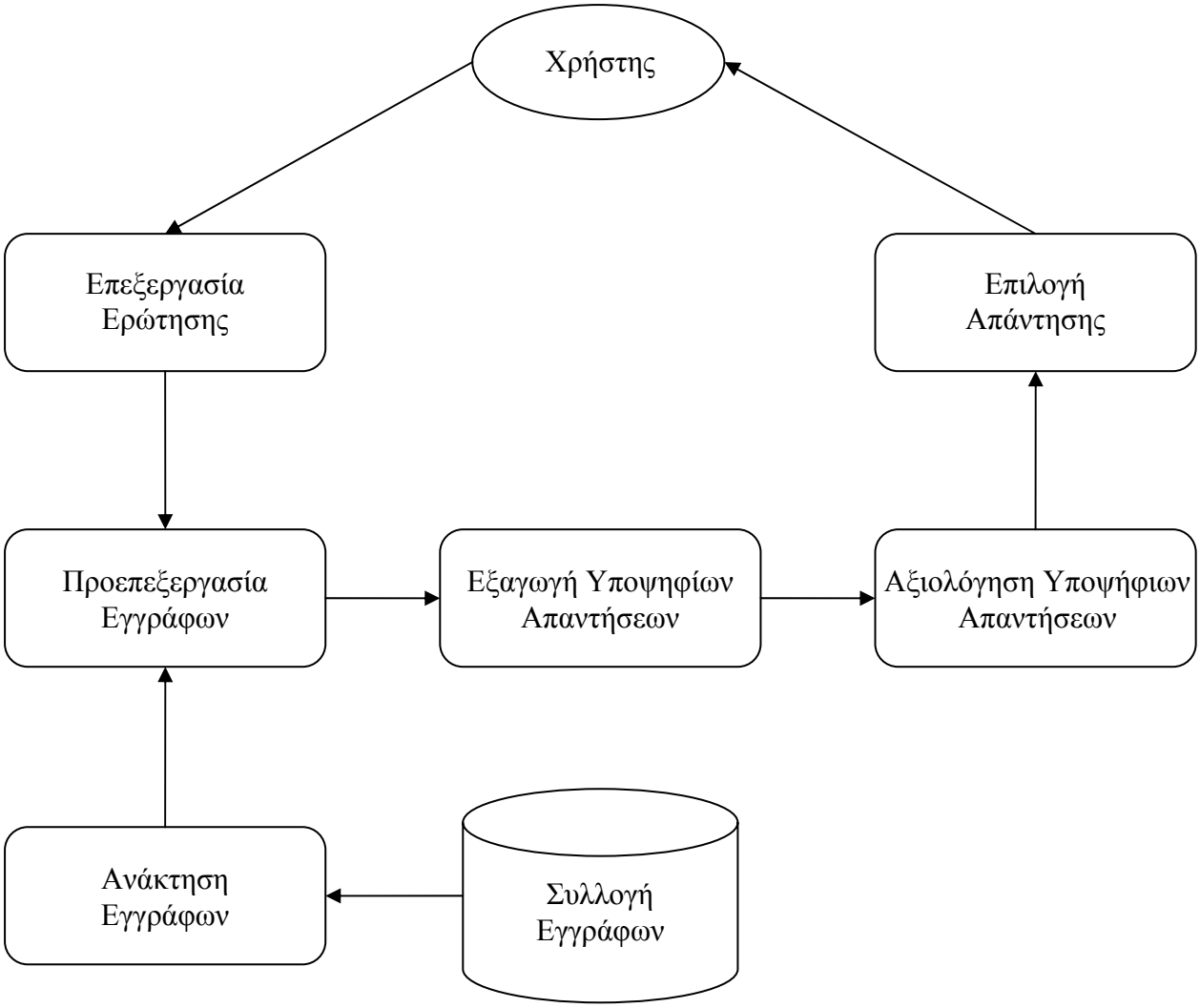
Αναλυτικά:

Ο χρήστης αρχικά δίνει μια ερώτηση στο σύστημα.

- **Επεξεργασία της Ερώτησης:** Το σύστημα προσπαθεί να εντοπίσει σε ποια από τις παραπάνω κατηγορίες ερωτήσεων ανήκει η ερώτηση που του τέθηκε, ώστε να ακολουθήσει τον αντίστοιχο αλγόριθμο (για παράδειγμα, συνήθως οι ερωτήσεις ονομάτων προσώπων ή τοποθεσιών απαιτούν την εύρεση αποσπασμάτων που να περιέχουν τουλάχιστον ένα όνομα προσώπου ή τοποθεσίας, αντίστοιχα, το καθένα, κάτι που δεν απαιτείται στις ερωτήσεις ορισμού). Επίσης στο στάδιο αυτό το σύστημα βρίσκει τους όρους (terms) του ερωτήματος (εξαιρώντας, για παράδειγμα συνηθισμένες λέξεις, γνωστές ως stop words) και πιθανόν άλλα στοιχεία που θα βοηθήσουν το σύστημα στην εύρεση της απάντησης (π.χ. μέρη του λόγου των λέξεων της ερώτησης, σε ορισμένα συστήματα συντακτικό δέντρο της ερώτησης κ.λ.π). Στην περίπτωση του συστήματος της εργασίας, όλες οι ερωτήσεις είναι ερωτήσεις

ορισμού προσώπων, οπότε δεν απαιτείται διαχωρισμός των ερωτήσεων σε κατηγορίες. Επίσης, ο χρήστης γράφει κατευθείαν το όνομα του προσώπου που θέλει να οριστεί (π.χ. γράφει «Βαϊσμίλερ» αντί «Ποιος ήταν ο Βαϊσμίλερ;»), οπότε δεν απαιτείται ούτε εξεύρεση των όρων της ερώτησης. Στη συνέχεια αναφερόμαστε στο όνομα που εισάγει ο χρήστης ως «**όνομα-στόχο**».

- **Ανάκτηση Εγγράφων:** Οι όροι της ερώτησης (στην περίπτωσή μας, ο όρος του οποίου ζητείται ο ορισμός) δίνονται ως λέξεις-κλειδιά σε μια κλασική μηχανή αναζήτησης, η οποία ανασύρει από τη συλλογή εγγράφων (π.χ. τα αρχεία μιας εφημερίδας ή ολόκληρο τον Παγκόσμιο Ιστό) έγγραφα που πιθανώς σχετίζονται με την ερώτηση του χρήστη και τα επιστρέφει με κάποια σειρά κατάταξης (ranked list). Επειδή η μηχανή αναζήτησης ενδέχεται να επιστρέψει μεγάλο πλήθος εγγράφων, το σύστημα επιλέγει μόνο τα Χ κορυφαία έγγραφα της σειράς κατάταξης. Στην παρούσα εργασία περιοριζόμαστε στα 12 κορυφαία έγγραφα που επιστρέφει η μηχανή αναζήτησης της εφημερίδας «Το Βήμα».
- **Προεπεξεργασία Εγγράφων:** Τα έγγραφα που προέκυψαν από τη συλλογή πιθανόν να χρειάζονται κάποια επεξεργασία, όπως αφαίρεση ετικετών (tags) ή μετατροπή τους σε κάποια άλλη επιθυμητή μορφή, ώστε να γίνει η μετέπειτα επεξεργασία τους από το σύστημα. Επίσης, ενδέχεται να απαιτείται ο εντοπισμός μέσα στα έγγραφα ονομάτων συγκεκριμένων κατηγοριών (π.χ. ονόματα τοποθεσιών στην περίπτωση ερωτήσεων τοποθεσιών), η συντακτική ανάλυση των εγγράφων (σε συστήματα που μετρούν πόσο μοιάζει το συντακτικό δέντρο της ερώτησης με το συντακτικό δέντρο κάθε υποψήφιας απάντησης) κ.λ.π.



Διαγραμματική αναπαράσταση της τυπικής αρχιτεκτονικής ενός συστήματος ερωταποκρίσεων

- **Εξαγωγή Υποψήφιων Απαντήσεων:** Από τα ανακτηθέντα έγγραφα επιλέγονται εκείνα τα μέρη τους που ενδέχεται να αποτελούν και απαντήσεις στην ερώτηση (π.χ. προτάσεις που περιέχουν ονόματα τοποθεσιών, στην περίπτωση ερωτήσεων τοποθεσιών). Στην παρούσα εργασία, επιλέγονται ως υποψήφιες απατήσεις όλα τα τμήματα κειμένων μήκους 250 χαρακτήρων που περιέχουν στο μέσο τους το όνομα-στόχο. Στο εξής αναφερόμαστε στα τμήματα αυτά ως «**παράθυρα**» του ονόματος-στόχου.
- **Αξιολόγηση Υποψήφιων Απαντήσεων:** Οι υποψήφιες απαντήσεις αξιολογούνται (στην περίπτωσή μας χρησιμοποιώντας το SVM) και παράγεται μια σειρά κατάταξης (ranked list) που δείχνει ποιες υποψήφιες

απαντήσεις το σύστημα θεωρεί καταλληλότερες, πιθανόν μαζί με το βαθμό βεβαιότητας του συστήματος για κάθε απάντηση.

- **Επιλογή Απάντησης:** Στο στάδιο αυτό το σύστημα επιλέγει την απάντηση ή τις απαντήσεις (αν π.χ. επιτρέπονται 5 απαντήσεις) που θεωρεί καταλληλότερες και τις επιστρέφει στον χρήστη. Αν χρησιμοποιείται και κάποια τεχνική παραγωγής φυσική γλώσσας, η απάντηση ενδέχεται να αποτελεί σύνθεση πολλών από τις κορυφαίες υποψήφιες απαντήσεις.

2.2 Μηχανική μάθηση

2.2.1 Κατηγοριοποίηση με επιβλεπόμενη μηχανική μάθηση

Η εργασία χρησιμοποιεί τεχνικές επιβλεπόμενης μηχανικής μάθησης (supervised learning) για να κατατάξει παράθυρα του ονόματος-στόχου (υποψήφιες απαντήσεις) σε δύο κατηγορίες: παράθυρα που αποτελούν ή όχι αποδεκτούς ορισμούς. Στην επιβλεπόμενη μηχανική μάθηση παρέχονται αρχικά στο σύστημα παραδείγματα εκπαίδευσης (στην περίπτωσή μας παράθυρα) για τα οποία είναι από πριν γνωστή η κατηγορία στην οποία ανήκουν. Το σύστημα επεξεργάζεται τα παραδείγματα εκπαίδευσης και παράγει ένα μοντέλο των κατηγοριών (π.χ. κανόνες κατάταξης, δέντρο απόφασης, στατιστικό μοντέλο), το οποίο χρησιμοποιείται στη συνέχεια για να καταταγούν στις κατηγορίες νέες περιπτώσεις (στην περίπτωσή μας, νέα παράθυρα) των οποίων δεν είναι γνωστές οι κατηγορίες.

2.2.2 Διανυσματική αναπαράσταση

Στους περισσότερους αλγορίθμους επιβλεπόμενης μάθησης, τόσο τα παραδείγματα εκπαίδευσης όσο και οι νέες περιπτώσεις που πρέπει να καταταγούν χρησιμοποιώντας το μοντέλο που προκύπτει από την εκπαίδευση, παριστάνονται ως διανύσματα. Κάθε διάνυσμα αποτελείται από n+1 τιμές, όπου n είναι το πλήθος των ιδιοτήτων που χρησιμοποιούνται για να περιγράψουν κάθε περίπτωση που πρέπει να καταταγεί. Η μία επιπλέον τιμή δείχνει σε ποια κατηγορία ανήκει το παράδειγμα (η τιμή αυτή είναι γνωστή στη διάρκεια της εκπαίδευσης και άγνωστη κατά την κατάταξη νέων περιπτώσεων μετά την εκπαίδευση).

Έστω, για παράδειγμα, ότι θέλουμε να αγοράσουμε ένα φορητό υπολογιστή και προσπαθούμε να εκπαιδεύσουμε ένα σύστημα να διακρίνει τους υπολογιστές που μας προτείνουν σε αυτούς που ικανοποιούν τις απαιτήσεις μας και σε εκείνους που δεν τις ικανοποιούν. Για απλούστευση, ας θεωρήσουμε ότι οι μόνες πληροφορίες (ιδιότητες) που έχουμε για κάθε υπολογιστή είναι οι εξής (στην πράξη θα υπήρχαν περισσότερες ιδιότητες):

- Το βάρος του υπολογιστή σε κιλά.
- Αν διαθέτει ο υπολογιστής chip γραφικών με δική του μνήμη ή όχι.
- Η χωρητικότητα της μνήμη RAM του υπολογιστή.

Έστω ότι μας προτείνουν αρχικά τα προϊόντα Α, Β και Γ με τα εξής χαρακτηριστικά, τα οποία κατατάσσουμε χειρωνακτικά στις δύο κατηγορίες (μας ικανοποιεί ή όχι), ώστε να χρησιμοποιηθούν ως παραδείγματα εκπαίδευσης (στην.

Υπολογιστής	Βάρος	Chip γραφικών	Μνήμη RAM	Ικανοποιεί;
Α	2.8	Ναι (1)	768	Ναι (1)
Β	3.8	Ναι (1)	2048	Ναι (1)
Γ	3	Όχι (0)	1024	Όχι (0)

Τα διανύσματα που περιγράφουν τα παραπάνω παραδείγματα είναι:

Υπολογιστής	Διάνυσμα
Α	{2.8, 1, 768, 1}
Β	{3.8, 1, 2048, 1}
Γ	{3, 0, 1024, 0}

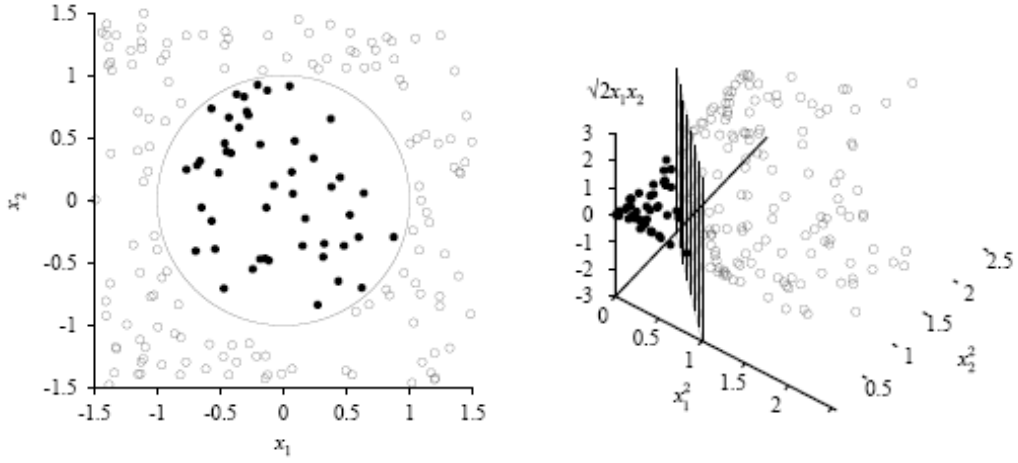
Από αυτά τα παραδείγματα το σύστημα θα μπορούσε να συμπεράνει ότι οι υπολογιστές που δε διαθέτουν chip γραφικών με δική του μνήμη θα πρέπει να απορρίπτονται, ενώ οι υπόλοιποι είναι αποδεκτοί. Δίνοντας περισσότερα παραδείγματα εκπαίδευσης και περισσότερες ιδιότητες, το σύστημα θα μπορούσε να κατασκευάσει ένα ακριβέστερο μοντέλο των προτιμήσεών μας.

2.2.3 Μηχανές Διανυσμάτων Υποστήριξης (SVM)[1]

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines, SVMs [17, 16, 33]) είναι μία σχετικά καινούρια μέθοδος επιβλεπόμενης μηχανικής μάθησης, η οποία μπορεί να εφαρμοστεί και σε προβλήματα κατηγοριοποίησης και η οποία έχει επιτύχει εξαιρετικά αποτελέσματα σε πολλές εφαρμογές. Στην απλούστερή τους μορφή, που χρησιμοποιούμε εδώ, οι ΜΔΥ μαθαίνουν να διαχωρίζουν περιπτώσεις δύο κατηγοριών. Ουσιαστικά προβάλλουν, με τη χρήση μίας συνάρτησης μετασχηματισμού, τα διανύσματα ιδιοτήτων σε ένα χώρο περισσοτέρων διαστάσεων και στη συνέχεια προσπαθούν να βρουν ένα γραμμικό διαχωριστή, δηλαδή ένα υπερεπίπεδο, που να διαχωρίζει τις δύο κατηγορίες με μέγιστο περιθώριο (margin) στο νέο διανυσματικό χώρο.

Η μετάβαση στο νέο χώρο περισσοτέρων διαστάσεων διευκολύνει την εύρεση γραμμικού διαχωριστή. Για παράδειγμα, στο παρακάτω σχήμα αριστερά φαίνεται μία περίπτωση, όπου δεν υπάρχει γραμμικός διαχωριστής (ευθεία) στο επίπεδο (εδώ υπάρχουν δύο μόνο ιδιότητες). Χρησιμοποιώντας όμως τη συνάρτηση μετασχηματισμού $\vec{F}(\vec{x}) = \left\langle x_1^2, x_2^2, \sqrt{2} \cdot x_1 \cdot x_2 \right\rangle$ και μεταβαίνοντας στις τρεις διαστάσεις παρατηρούμε ότι υπάρχει ένα επίπεδο που διαχωρίζει τα διανύσματα (σχήμα στα δεξιά). Στην περίπτωση περισσοτέρων ιδιοτήτων, ο διαχωριστής θα είναι ένα υπερεπίπεδο.

[1] Το κείμενο αυτής της ενότητας προέρχεται από την εργασία του Γιώργου Λουκαρέλλι [3] και περιλαμβάνεται σε αυτήν την εργασία έπειτα από συνεννόηση με τον επιβλέποντα καθηγητή του, κ. Ίωνα Ανδρουτσόπουλο.



Μετασχηματισμός από τις δύο διαστάσεις στις τρεις [1]

Γενικά, η εξίσωση του υπερεπιπέδου διαχωρισμού θα είναι της ακόλουθης μορφής, όπου F η συνάρτηση μετασχηματισμού:

$$\vec{w} \cdot \vec{F}(\vec{x}) + b = 0$$

Το υπερεπίπεδο διαχωρισμού τοποθετείται στο μέσον της απόστασης δύο παράλληλων υπερεπιπέδων, τα οποία διαχωρίζουν πλήρως τα παραδείγματα εκπαίδευσης και εφάπτονται με τουλάχιστον ένα παράδειγμα εκπαίδευσης, διαφορετικής κατηγορίας για το κάθε ένα από τα δύο υπερεπίπεδα. Τα $\vec{w}$ ($\vec{w} \in R^l$, όπου l ο αριθμός των ιδιοτήτων στο νέο χώρο) και b μπορούν να επιλεγούν (με scaling) ώστε τα δύο παράλληλα εφαπτόμενα υπερεπίπεδα να έχουν εξισώσεις:

$$\vec{w} \cdot \vec{F}(\vec{x}) + b = \pm 1$$

οπότε η απόσταση μεταξύ των δύο εφαπτόμενων υπερεπιπέδων είναι $2/\left\|\vec{w}\right\|$. Η απόσταση αυτή είναι το «περιθώριο» του υπερεπιπέδου διαχωρισμού, που τοποθετείται στο μέσον της απόστασης των δύο εφαπτόμενων υπερεπιπέδων. Όπως αναφέρθηκε ήδη, ο στόχος των ΜΔΥ είναι να βρουν το υπερεπίπεδο διαχωρισμού με το μέγιστο περιθώριο. Οπότε, προκύπτει τα παρακάτω πρόβλημα βελτιστοποίησης:

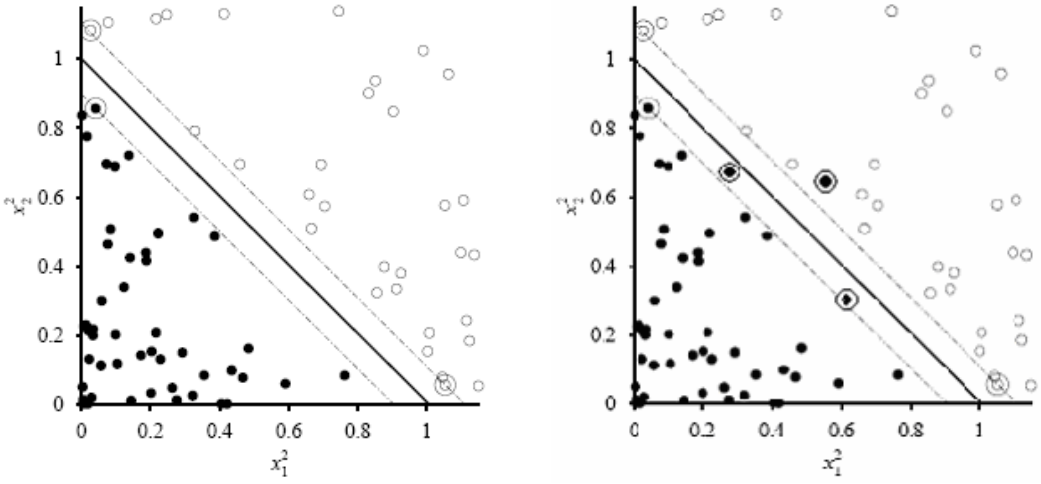
$$\min \left\|\vec{w}\right\|^2 / 2$$

[1] Τα σχήματα είναι από το βιβλίο των Stuart Russell και Peter Norvig *"Artificial Intelligence: A Modern Approach (Second Edition)"*, Prentice Hall, 2002.

$$(\vec{w} \cdot \vec{F}(\overrightarrow{x_j}) + b) \cdot y_j \geq 1$$

όπου x_j με $1 \leq j \leq n$ είναι το διάνυσμα του j -οστού παραδείγματος εκπαίδευσης και $y_j \in \{1,-1\}$ είναι η κατηγορία του j -οστού διανύσματος εκπαίδευσης.

Οι περιορισμοί του παραπάνω προβλήματος βελτιστοποίησης επιβάλλουν όλα τα διανύσματα εκπαίδευσης να βρίσκονται έξω ή το πολύ στα όρια του περιθωρίου και από τη σωστή πλευρά του υπερεπιπέδου, ανάλογα με την κατηγορία τους, όπως φαίνεται στο παρακάτω σχήμα αριστερά. Οι περιορισμοί αυτοί, όμως, είναι πολύ αυστηροί. Για παράδειγμα, ενδέχεται να μην είναι δυνατή η εύρεση γραμμικού διαχωριστή που να διαχωρίζει πλήρως τα παραδείγματα εκπαίδευσης, παρά τη μετάβαση στο νέο χώρο διαστάσεων. Ή ενδέχεται να προτιμούμε ένα υπερεπίπεδο διαχωρισμού που έχει μεγαλύτερο περιθώριο αλλά κατατάσσει λανθασμένα ή εντός του περιθωρίου κάποια παραδείγματα εκπαίδευσης (όπως στο παρακάτω σχήμα στα δεξιά) από κάποιο άλλο που ικανοποιεί όλους τους περιορισμούς αλλά έχει μικρότερο περιθώριο.



Υπερεπίπεδο με μέγιστο περιθώριο[1]

Για τους λόγους αυτούς, είναι δυνατόν οι περιορισμοί να χαλαρώσουν, με αποτέλεσμα να κατασκευαστεί ένα ανεκτικότερο πρόβλημα βελτιστοποίησης, το οποίο ορίζεται ως εξής:

[1] Τα σχήματα είναι από το βιβλίο των Stuart Russell και Peter Norvig *"Artificial Intelligence: A Modern Approach (Second Edition)"*, Prentice Hall, 2002.

$$\min \left\| \vec{w} \right\|^2 / 2 + C \cdot \sum_j \xi_j$$
$$(\vec{w} \cdot \vec{F}(\vec{x_j}) + b) \cdot y_j \geq 1 - \xi_j$$
$$\xi_j \geq 0$$

όπου το ξ_j είναι το σφάλμα για κάθε διάνυσμα εκπαίδευσης (το πόσο απέχουμε από το να ικανοποιείται ο αντίστοιχος περιορισμός) και C το κόστος (ανοχή) που δίνεται στο συνολικό σφάλμα. Στην περίπτωση αυτή, όπως φαίνεται στο παραπάνω σχήμα δεξιά, υπάρχει καλύτερη δυνατότητα γενίκευσης και ο διαχωριστής είναι πιο ανεκτικός σε λάθη επισημείωσης των δεδομένων εκπαίδευσης.

Τελικά, επιλύνοντας το παραπάνω πρόβλημα ελαχιστοποίησης, προκύπτει $\vec{w}$ της μορφής:

$$\vec{w} = \sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j})$$

Οπότε η εξίσωση του υπερεπιπέδου γίνεται:

$$\left(\sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j}) \right) \cdot \vec{F}(\vec{x}) + b = 0$$

ή

$$\left(\sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x}) \right) + b = 0$$

όπου οι τιμές των a_j είναι διάφορες του μηδενός μόνο για τα «διανύσματα υποστήριξης», δηλαδή τα διανύσματα εκπαίδευσης που βρίσκονται πάνω στα δύο εφαπτόμενα υπερεπίπεδα και (στην περίπτωση που ανεχόμαστε σφάλματα) τα διανύσματα που κατατάσσονται λανθασμένα ή εντός του περιθωρίου. Τα παραδείγματα που δεν είναι διανύσματα υποστήριξης ουσιαστικά αγνοούνται.
Η συνάρτηση μετασχηματισμού συμμετέχει μόνο σε εσωτερικά γινόμενα $\vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$. Ορίζουμε, λοιπόν, ως πυρήνα της ΜΔΥ τη συνάρτηση:

$$K(\vec{x_j}, \vec{x_i}) = \vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$$

Σύμφωνα με το θεώρημα του Mercer, κάθε συνάρτηση $K(\overrightarrow{x_i},\overrightarrow{x_j})$ για την οποία ο πίνακας $K_{ij} = K(\overrightarrow{x_i},\overrightarrow{x_j})$ είναι θετικά ορισμένος[1] υπολογίζει το εσωτερικό γινόμενο των $\overrightarrow{x_i},\overrightarrow{x_j}$ σε κάποιο νέο διανυσματικό χώρο, δηλαδή μπορεί να χρησιμοποιηθεί ως πυρήνας μιας ΜΔΥ. Το ενδιαφέρον είναι ότι σε πολλές περιπτώσεις είναι δυνατόν να υπολογιστούν οι τιμές του πυρήνα χωρίς να υπολογιστεί πρώτα η τιμή των $\vec{F}(\overrightarrow{x_j})$ και $\vec{F}(\overrightarrow{x_i})$, δηλαδή χωρίς να υπολογίσουμε τις (συνήθως πολύ περισσότερες) ιδιότητες των διανυσμάτων στο νέο χώρο, κάτι που επιτρέπει τη χρήση πυρήνων που υπολογίζουν εσωτερικά γινόμενα σε νέους χώρους πολύ μεγάλου αριθμού διαστάσεων. Για παράδειγμα, στην περίπτωση του αρχικού παραδείγματος μετασχηματισμού με $\vec{F}(\vec{x}) = \left\langle x_1^2, x_2^2, \sqrt{2}\cdot x_1 \cdot x_2 \right\rangle$ ο πυρήνας έχει τη μορφή $K(\overrightarrow{x_i},\overrightarrow{x_j}) = F(\overrightarrow{x_i})\cdot F(\overrightarrow{x_j}) = (\overrightarrow{x_i}\cdot\overrightarrow{x_j})^2$, δηλαδή οι τιμές του μπορούν να υπολογισθούν με βάση μόνο τις τιμές των ιδιοτήτων του αρχικού χώρου. Τελικά, αντικαθιστώντας το εσωτερικό γινόμενο $\vec{F}(\overrightarrow{x_j})\cdot\vec{F}(\overrightarrow{x_i})$ με τη συνάρτηση πυρήνα $K(\overrightarrow{x_j},\overrightarrow{x_i})$ η εξίσωση του υπερεπιπέδου γίνεται:

$$\left(\sum_j a_j \cdot y_j \cdot K(\overrightarrow{x_j},\vec{x})\right) + b = 0$$

Παραδείγματα πυρήνων που χρησιμοποιούνται είναι τα εξής:

- γραμμικός: $K(\overrightarrow{x_i},\overrightarrow{x_j}) = \overrightarrow{x_i}\cdot\overrightarrow{x_j}$
- πολυωνυμικός: $K(\overrightarrow{x_i},\overrightarrow{x_j}) = (\overrightarrow{x_i}\cdot\overrightarrow{x_j}+1)^d$
- ακτινωτής βάσης (radial base function – RBF):
  $K(\overrightarrow{x_i},\overrightarrow{x_j}) = \exp\left(-\gamma\cdot\left\|\overrightarrow{x_i}-\overrightarrow{x_j}\right\|^2\right), \gamma > 0$
- σιγμοειδής: $K(\overrightarrow{x_i},\overrightarrow{x_j}) = \tanh(\overrightarrow{x_i}\cdot\overrightarrow{x_j}+r)$

[1] Ο πίνακας $A \in R^{n\times n}$ είναι ένας θετικά ορισμένος πίνακας αν για όλα τα μη μηδενικά διανύσματα $x \in R^n$ ισχύει $x^T \cdot A \cdot x > 0$, όπου x^T είναι το ανάστροφο διάνυσμα.

όπου τα γ, r και d είναι παράμετροι κάθε πυρήνα. Ο γραμμικός πυρήνας (που δεν προκαλεί μετάβαση σε νέο διανυσματικό χώρο) είναι ειδική περίπτωση του πυρήνα ακτινωτής βάσης (Keerthi και Lin 2003). Επίσης, ο σιγμοειδής πυρήνας συμπεριφέρεται όπως ο πυρήνας ακτινωτής βάσης για συγκεκριμένες παραμέτρους (Lin και Lin 2003).

Τέλος, η απόφαση για την κατηγορία ενός καινούριου διανύσματος, δεδομένου ενός εκπαιδευμένου ταξινομητή, λαμβάνεται με βάση το πρόσημο της ακόλουθης παράστασης:

$$sign\left(\sum_j a_j \cdot y_j \cdot K(\vec{x}_j, \vec{x}) + b\right)$$

3 Το Σύστημα της Εργασίας

‘I haven't failed, I've found 10,000 ways that don't work’
(Thomas Edison)

Στο κεφάλαιο αυτό αναλύεται ο τρόπος με τον οποίο δημιουργήθηκαν τόσο τα δεδομένα εκπαίδευσης όσο και τα δεδομένα αξιολόγησης. Περιγράφεται επίσης λεπτομερώς η προσέγγισή μας στο πρόβλημα της επιλογής των ιδιοτήτων που χρησιμοποιούνται κατά την κατάταξη των υποψηφίων απαντήσεων και παρουσιάζονται τα εργαλεία που χρησιμοποιήσαμε για την υλοποίηση του συστήματος.

3.1 Δημιουργία του σώματος εκπαίδευσης και αξιολόγησης

Το σύστημά μας αναλαμβάνει να απαντήσει σε ερωτήσεις της μορφής «Ποιος/ποια/ποιο είναι <άρθρο> <Όνομα Προσώπου>» ψάχνοντας στο αρχείο μιας εφημερίδας. Η τρέχουσα υλοποίηση του συστήματος ψάχνει, όπως προαναφέρθηκε, στο αρχείο της εφημερίδας «Το Βήμα» [5], που περιλαμβάνει φύλλα πολλών ετών. Το σύστημα μπορεί, όμως, να τροποποιηθεί εύκολα, ώστε να ψάχνει στα αρχεία και άλλων εφημερίδων, αρκεί αυτά να είναι διαθέσιμα στον Παγκόσμιο Ιστό μέσω ιστοτόπων που παρέχουν δυνατότητες ανάκτησης άρθρων με τη χρήση λέξεων-κλειδιών.

Για τη δημιουργία των δεδομένων εκπαίδευσης και αξιολόγησης επιλέξαμε 200 ονόματα προσώπων για τα οποία θα μπορούσαμε να βρούμε πληροφορίες στα αρχεία μιας εφημερίδας. Συγκεκριμένα ανατρέξαμε σε πρόσφατα φύλλα της εφημερίδας «Το Βήμα» και καταγράψαμε με τυχαίο τρόπο 200 ονόματα που συναντήσαμε. Τα ονόματα αυτά ήταν:

Αβραμόπουλος	Γούλφοβιτς	Κουστουρίτσα	Νικολακοπούλου	Σβάιτσερ
Αγκάθα	Γραμματικάκης	Κουστώ	Νίξον	Σεφέρης
Αγραπίδης	Γρυλλάκης	Κραουνάκης	Νταλί	Σιράκ
Αλαβάνος	Δερτούζος	Κρίκ	Ντασσέν	Σλίμαν
Αλαφούζος	Δήμας	Κρόνενμπεργκ	Ντελ Πόντε	Σπανουδάκης
Αλιάγα	Διώτης	Κυριακού	Ντελακρουά	Σπίλμπεργκ
Αλμοδόβαρ	Εγγονόπουλος	Κωστόπουλος	Ντεμιρέλ	Σρέντερ
Αλόνσο	Ειρηναίος	Λάγιος	Ντενκτάς	Ταλαμπανί
Αμπάς	Ερντογάν	Λάτσης	Ντίβατς	Τεγόπουλος
Ανάν	Ζαγοράκης	Λένιν	Ντίσνεϋ	Τέρνερ
Αρβελέρ	Ζιντάν	Λεπέν	Ξενάκης	Τζιοβάνι
Αριστοτέλης	Ζολά	Λυκουρέζος	Οζάλ	Τζόινερ
Αρλέτα	Θάτσερ	Λυμιέρ	Ολιβιέ	Τζουγκάνοβιτς
Άρμστρονγκ	Θεωδοράκης	Μαντόνα	Ολμπράιτ	Τίτο
Ασλάνης	Ικτίνος	Μαραντόνα	Οπενχάιμερ	Τόλκιν
Βαβύλης	Καβάφης	Ματίς	Όργουελ	Τρίτσης
Βαϊσμίλερ	Καζάν	Μερκούρη	Ουμπέρτο	Τρούμαν
Βαντίμ	Καζαντζάκης	Μέρντοκ	Ουστίνοφ	Τσαϊκόφσκι
Βαρδινογιάννης	Καλατράβα	Μικρούτσικος	Παβαρότι	Τσαουσέσκου
Βενιέρης	Καμύ	Μίλλερ	Πάλμε	Τσάπλιν
Βέρν	Κανέλλη	Μιλόσεβιτς	Παναγούλης	Τσαρούχας
Βερσάτσε	Καπράλος	Μινέλι	Πάουελ	Τσολάκης
Βολταίρος	Κάρατζιτς	Μιτεράν	Παπαζάχος	Τσόμσκι
Βούγιας	Καρεμπέ	Μονρόε	Παπαθανασίου	Φαραντούρη
Βουλγαράκης	Καρέρας	Μουρ	Πατουλίδου	Φασουλής
Βούλγαρης	Κάρτερ	Μούσχουρη	Πελέ	Φειδίας
Βουρλούμης	Καρυωτάκης	Μπάγεβιτς	Πικάσο	Φιορίνα
Γαλιλαίος	Κατσιμίχας	Μπαλάφας	Πολύζος	Φίσερ
Γέλτσιν	κεντέρης	Μπαμπινιώτης	Πολυζωγόπουλος	Φούρας
Γεωργουσόπουλος	Κέπλερ	Μπαρόζο	Πουτσίνι	Φυντανίδης
Γιακουμάκης	Κιμούλης	Μπασινάς	Ραγκούζα	Φώσκολος
Γιακούμπ	Κληρίδης	Μπατίστ	Ράμσφελντ	Χάιντερ
Γκάι	Κοέλο	Μπέκαμ	Ραχμάνινοφ	Χαλεπάς
Γκάντι	Κομανέτσι	Μπέργκμαν	Ρεχάγκελ	Χάντιγκτον
Γκέιτς	Κοντομηνάς	Μπερνς	Ρουγκόβα	Χατζηδάκις
Γκέμπελς	Κόπερφιλντ	Μπόμπολας	Σαγκάν	Χατζηνικολάου
Γκρίνσπαν	Κούγιας	Μπράντο	Σαμαράκης	Χίτσκοκ
Γκρίφιθ	Κούκ	Μπρέζνιεφ	Σάμαρανκ	Χόκινς
Γουέλς	Κουμεντάκης	Μπρεχτ	Σαρόν	Χούβερ
Γούλφερσον	Κουροσάβα	Νεμπιόλο	Σαρτρ	Χούκλη

Το επόμενο βήμα ήταν να μαζέψουμε το σύνολο των άρθρων που αναφέρονται σε αυτά τα πρόσωπα, κάτι που επιτύχαμε χρησιμοποιώντας την υπάρχουσα μηχανή αναζήτησης του ιστοτόπου της εφημερίδας. Καθώς τα άρθρα που μας επέστρεφε η

μηχανή αναζήτησης ήταν πολλά, κρατήσαμε μόνο τα 12 πρώτα για κάθε όνομα. Έτσι δημιουργήσαμε ένα σώμα κειμένων (corpus) που αποτελείται από 2400 άρθρα. Τα άρθρα αυτά είναι σε μορφή HTML, όπως επιστρέφονται από τη μηχανή αναζήτησης της εφημερίδας.[1] Στα πειράματα της εργασίας, το σώμα κειμένων των 2400 άρθρων χρησιμοποιήθηκε τόσο για την εκπαίδευση όσο και για την αξιολόγηση του συστήματος, με την τεχνική της διασταυρωμένης επικύρωσης (cross-validation)· περισσότερες πληροφορίες δίνονται στο 4ο κεφάλαιο.

Τα 2400 άρθρα αποθηκεύτηκαν έτσι ώστε η εκπαίδευση και η αξιολόγηση του συστήματος να είναι δυνατό να γίνει off-line. (Ωστόσο, όπως θα περιγραφεί σε επόμενες ενότητες, μετά την εκπαίδευσή του το σύστημα είναι δυνατόν να χρησιμοποιηθεί και on-line. Στην περίπτωση αυτή, ο χρήστης εισάγει ένα όνομα προσώπου που επιθυμεί να οριστεί και το σύστημα ανακτά άρθρα εκείνη τη στιγμή από το αρχείο της εφημερίδας και εντοπίζει μέσα σε αυτά κατάλληλους ορισμούς, τους οποίους παρουσιάζει στο χρήστη.) Η αποθήκευση των άρθρων στο τοπικό σκληρό δίσκο έγινε με τη χρήση ενός προγράμματος σε Java, το οποίο αναλαμβάνει να στείλει ερωτήματα (στην περίπτωσή μας ονόματα προσώπων) στον εξυπηρετητή (server) της μηχανής αναζήτησης του ιστότοπου της εφημερίδας και να αποθηκεύσει στο σκληρό δίσκο τα άρθρα που επιστρέφει ο εξυπηρετητής. Το ίδιο πρόγραμμα χρησιμοποιείται και στην on-line μορφή του συστήματός μας, για να ανακτήσει από το αρχείο της εφημερίδας τα άρθρα που αναφέρονται στο όνομα-στόχο.[2]

Το επόμενο βήμα ήταν ο εντοπισμός των παραθύρων των ονομάτων-στόχων (ενότητα 2.1.3) μέσα στα 2400 άρθρα και η χειρωνακτική σημείωση των παραθύρων ως αποδεκτών απαντήσεων ή όχι. Για να σημειώσουμε τα παράθυρα χρησιμοποιήσαμε την ετικέτα (tag) **1**<όνομα-στόχος>**$$**, αν το παράθυρο αποτελούσε αποδεκτή απάντηση (αποδεκτό ορισμό του ονόματος-στόχου). Διαφορετικά, αν η απάντηση θεωρείτο μη αποδεκτή, η ετικέτα ήταν **0**<όνομα-στόχος>**$$**. Επιλέχθηκαν οι

[1] Λόγω προβλημάτων που συναντήσαμε κατά το χειρισμό των κειμένων των άρθρων σε Unicode, τα κείμενα των άρθρων μετατρέπονται και αποθηκεύονται σε Windows-1253 (η κωδικοποίηση που χρησιμοποιούν τα Windows για τις ιστοσελίδες που περιέχουν ελληνικά).

[2] Στη διάρκεια της εργασίας χρειάστηκε να γίνουν κάποιες μικροαλλαγές στον κώδικα του προγράμματος αυτού, λόγω αλλαγών στον ιστότοπο της εφημερίδας. Η ευκολία με την οποία έγιναν αυτές οι αλλαγές είναι ένα δείγμα του πόσο εύκολα μπορεί να μεταφερθεί ή το σύστημά μας και σε άλλες εφημερίδες.

συγκεκριμένες ετικέτες καθώς είναι απίθανο να εμφανιστούν σε ελληνικά άρθρα εφημερίδων.

Σε κάθε άρθρο του σώματος εκπαίδευσης και αξιολόγησης έχουν σημειωθεί μόνο τα πέντε πρώτα παράθυρα του ονόματος-στόχου, ενώ τα υπόλοιπα παράθυρα έχουν αγνοηθεί, δηλαδή δε θεωρούνται υποψήφιες απαντήσεις. Το ίδιο συμβαίνει και κατά την on-line χρήση του συστήματος: σε κάθε άρθρο που επιστρέφει η μηχανή αναζήτησης της εφημερίδας, υποψήφιες απαντήσεις θεωρούνται μόνο τα 5 πρώτα παράθυρα του ονόματος-στόχου. Αυτό γίνεται επειδή είναι απίθανο σε ένα άρθρο να οριστεί ένα όνομα αφού έχει ήδη εμφανιστεί 5 φορές. Αποκλείονται έτσι παράθυρα που είναι σίγουρο πως δεν αποτελούν ορισμούς. Με τον τρόπο αυτό μειώνεται και η διαφορά μεταξύ του πλήθους των παραθύρων που αποτελούν ορισμούς και του πλήθους των παραθύρων που δεν αποτελούν ορισμούς, η οποία, αν είναι πολύ μεγάλη, μπορεί να οδηγήσει τον αλγόριθμο μάθησης να μάθει να κατατάσσει όλα τα παράθυρα στην κατηγορία των μη ορισμών. Στην πράξη παρατηρήσαμε ότι τα περισσότερα άρθρα που επιστρέφει η μηχανή αναζήτησης της εφημερίδας αναφέρονται μόνο 2-3 φορές στο όνομα-στόχο, οπότε η επίδραση της απόφασής μας να αγνοούμε τα παράθυρα μετά το $5^{ο}$ είναι μικρή.

<u>Παραδείγματα σημειωμένων παραθύρων:</u>

Και στις δύο παρακάτω περιπτώσεις, το όνομα-στόχος ήταν «*Λυκουρέζος*»:

Θετικό παράδειγμα (ορισμός): *ο υπουργός Κοινωνικής Απασχόλησης κ. Π. Παναγιωτόπουλος, ο πρόεδρος της ΕΠΑΕ κ. Αλ. 1Λυκουρέζος$$. Για την Υπερνομαρχία Αθηνών-Πειραιώς τα πάντα είναι «ανοικτά» και η Ρηγίλλης θα αποφασίσει την ύστατη στιγμή, ενώ για τη Νομαρχία Αθηνών ακούγεται το όνομα*

Αρνητικό παράδειγμα (μη ορισμός): *Οι κκ. Ν. Κακλαμάνης (αριστερά) και Αλ. 0Λυκουρέζος$$ θεωρούνται ως ενδιαφερόμενοι για τον Δήμο Αθηναίων*

Η χειρωνακτική σημείωση είναι μια δύσκολη και επίπονη διαδικασία. Εκτός του ότι απαιτεί πάρα πολύ χρόνο, απαιτεί και ιδιαίτερη προσοχή, αφού τυχόν λανθασμένη σημείωση προσθέτει θόρυβο στα δεδομένα εκπαίδευσης του συστήματος. Εκτός αυτού, όπως προαναφέρθηκε, υπάρχουν περιπτώσεις όπου δεν είναι προφανές αν ένα

παράθυρο πρέπει να σημειωθεί ως αποδεκτός ορισμός ή όχι, κάτι που δυσκολεύει ακόμα περισσότερο τη διαδικασία της χειρωνακτικής σημείωσης.[1]

Καθώς τα ονόματα στην ελληνική γλώσσα κλίνονται, τα παράθυρα του ονόματος-στόχου δεν ήταν δυνατόν να εντοπισθούν μέσα στα άρθρα της εφημερίδας απευθείας με ταίριασμα συμβολοσειρών (string matching)· για παράδειγμα, οι λέξεις «Λυκουρέζος» και «Λυκουρέζου» θα θεωρούνταν διαφορετικά ονόματα. Έτσι χρειάστηκε να δημιουργήσουμε μια απλοϊκή αλλά ιδιαίτερα αποτελεσματική για ονόματα προσώπων μονάδα αποκοπής καταλήξεων (stemmer).[2] Η μονάδα αυτή αφαιρεί το τελευταίο γράμμα της λέξης αν είναι 'ν' ή 'ς'. Έπειτα αφαιρεί όλα τα φωνήεντα που βρίσκονται στο τέλος της λέξης μέχρι να ξανασυναντήσει σύμφωνο. Στο παράδειγμά μας, τα «Λυκουρέζος» και «Λυκουρέζου» θα γίνονταν και τα δύο «Λυκουρέζ». Έτσι, μετά την αποκοπή των καταλήξεων, μπορούμε να εντοπίσουμε σχεδόν όλα τα παράθυρα του ονόματος-στόχου με ταίριασμα συμβολοσειρών· προβλήματα δημιουργούνται μόνο σε σπάνιες περιπτώσεις, όπως οι συντομογραφίες ονομάτων.

3.2 Επιλογή ιδιοτήτων

Όπως προαναφέρθηκε, σε ένα σύστημα που εκτελεί κατηγοριοποίηση με επιβλεπόμενη μάθηση, οι προς κατάταξη περιπτώσεις (στην περίπτωσή μας οι υποψήφιες απαντήσεις) παριστάνονται ως διανύσματα ιδιοτήτων. Η επιλογή των ιδιοτήτων που θα περιλαμβάνονται στα διανύσματα είναι ένα από τα πιο σημαντικά και δύσκολα στάδια της ανάπτυξης του συστήματος, αφού πρέπει οι ιδιότητες να παρέχουν αρκετές πληροφορίες, ώστε να είναι δυνατός ο διαχωρισμός των

[1] Η εργασία [4] προτείνει μια μέθοδο για την αυτόματη σημείωση παραθύρων εκπαίδευσης ως ορισμών ή μη ορισμών, η οποία έχει χρησιμοποιηθεί σε συστήματα εντοπισμού αγγλικών ορισμών σε ιστοσελίδες. Η μέθοδος αυτή χρησιμοποιεί κατά την εκπαίδευση όρους-στόχους (στην περίπτωσή μας ονόματα-στόχους) για τους οποίους υπάρχουν ορισμοί ταυτόχρονα σε ηλεκτρονικές εγκυκλοπαίδειες και κείμενα της συλλογής (στην περίπτωσή μας άρθρα εφημερίδων). Στην περίπτωσή μας, όμως, τα ονόματα που ορίζονται στα άρθρα των εφημερίδων είναι κυρίως ονόματα της επικαιρότητας, για τα οποία δεν υπάρχουν ορισμοί σε εγκυκλοπαίδειες, οπότε είναι αμφίβολο αν η μέθοδος της εργασίας [4] θα μπορούσε να εφαρμοστεί επιτυχώς.

[2] Μια μονάδα αποκοπής καταλήξεων είναι ένα πρόγραμμα που χρησιμοποιώντας απλούς κανόνες ή ακόμα και λεξικά προσπαθεί να βρει τη ρίζα της λέξης που του δίνεται ως είσοδος. Για την ελληνική γλώσσα υπάρχει ο stemmer των Σ. Νικολαΐδη και Θ. Καλαμπούκη, του Οικονομικού Πανεπιστημίου Αθηνών (βλ. http://www.cs.aueb.gr/dpmlab/publicationsgr.htm).

περιπτώσεων σε κατηγορίες, αλλά όχι άσχετες πληροφορίες που εισάγουν μόνο θόρυβο και δυσχεραίνουν το διαχωρισμό.

3.2.1 Οι βασικές ιδιότητες

Το σύστημά μας χρησιμοποιεί σε όλες του τις μορφές 2 βασικές ιδιότητες, που εξηγούνται παρακάτω. Η πιο απλοϊκή μορφή του συστήματος, την οποία χρησιμοποιούμε ως μέθοδο αναφοράς (baseline), χρησιμοποιεί μόνο αυτές τις 2 ιδιότητες. Άλλες μορφές του συστήματος χρησιμοποιούν επιπρόσθετες ιδιότητες, που επιλέγονται με τρόπους που θα εξηγηθούν σε επόμενες ενότητες. Υπενθυμίζεται ότι όλες οι ιδιότητες παρέχουν πληροφορίες για τις υποψήφιες απαντήσεις (παράθυρα του όρου-στόχου) που το SVM επιχειρεί να διαχωρίσει σε αποδεκτούς και μη ορισμούς.

Οι δύο βασικές ιδιότητες, οι οποίες προέρχονται από την εργασία [2], είναι:

a) **Ο τακτικός αριθμός του παραθύρου στο άρθρο από το οποίο προέρχεται,** δηλαδή το αν το παράθυρο ήταν το 1ο, $2^{ο}$, $3^{ο}$ κ.λ.π. παράθυρο του ονόματος-στόχου στο άρθρο. Η ιδιότητα αυτή παρέχει χρήσιμες πληροφορίες, γιατί συνήθως οι πρώτες εμφανίσεις ενός όρου (στην περίπτωσή μας, ενός ονόματος προσώπου) σε ένα κείμενο είναι πιθανότερο να συνοδεύονται από ένα σύντομο ορισμό του όρου από ό,τι οι παρακάτω εμφανίσεις του όρου στο κείμενο. Ειδικά σε άρθρα εφημερίδων, παρατηρείται ότι στις πρώτες εμφανίσεις ενός ονόματος προσώπου σε ένα άρθρο (όχι πάντα στην $1^{η}$, που μπορεί π.χ. να αποτελεί μέρος τίτλου ή περίληψης) δίνεται συχνά και η ιδιότητα του προσώπου για την οποία είναι γνωστό (π.χ. το επάγγελμά του), που μπορεί να θεωρηθεί σύντομος ορισμός.
   Οι τιμές που μπορεί να πάρει αυτή η ιδιότητα είναι $n/5$, όπου n ο τακτικός αριθμός του παραθύρου, που παίρνει τιμές από 1 ως και 5 αφού εξετάζουμε μόνο τα 5 πρώτα παράθυρα. Διαιρούμε με το 5 ώστε η τιμή που προκύπτει να είναι πάντα ανάμεσα στο 0 και στο 1, σύμφωνα με τις συστάσεις των κατασκευαστών της υλοποίησης SVM (libSVM) που χρησιμοποιούμε (περισσότερες πληροφορίες για τη libSVM δίνονται στην ενότητα 3.3).

b) **Το πλήθος των πιο κοινών λέξεων που περιέχονται στο παράθυρο:** Σε κάθε παράθυρο W με όνομα-στόχο t, η τιμή αυτής της ιδιότητας υπολογίζεται ως εξής.

Συλλέγονται πρώτα όλα τα παράθυρα του *t* που βρίσκονται μέσα στα άρθρα που επέστρεψε η μηχανή αναζήτησης της εφημερίδας για το *t*. Στη συνέχεια καταγράφονται όλες οι λέξεις των παραθύρων αυτών, χωρίς να λαμβάνονται υπόψη οι πολύ συχνές λέξεις (stop words, τέτοιες λέξεις είναι τα άρθρα, οι σύνδεσμοι κ.λ.π. – βλ. Παράρτημα Α) και εφαρμόζεται πάνω τους η μονάδα αποκοπής καταλήξεων (ενότητα 3.1). Όμως αυτή τη φορά η μονάδα αποκοπής καταλήξεων αφαιρεί και τους τόνους.[1] Από τις λέξεις (ακριβέστερα, τις ρίζες λέξεων) που προκύπτουν, επιλέγονται οι 20 πιο συχνές, δηλαδή οι 20 λέξεις που εμφανίζονται συχνότερα στα παράθυρα του *t*. Στην εργασία [2], η τιμή της ιδιότητας στο παράθυρο *W* ήταν το ποσοστό των 20 πιο συχνών λέξεων που εμφανίζονται στο *W*. Η προσέγγιση αυτή έδινε την ίδια βαρύτητα και στις 20 συχνές λέξεις, ενώ κάποιες από αυτές ενδέχεται να είναι πολύ συχνότερες από κάποιες άλλες. Αντίθετα, στην παρούσα εργασία η τιμή της ιδιότητας υπολογίζεται ως εξής:

$$\frac{\sum_i f(w_i)}{\sum_j f(w_j)}$$

όπου το w_j είναι οι 20 συχνότερες λέξεις και w_i εκείνες από τις 20 συχνότερες λέξεις που εμφανίζονται στο παράθυρο *W*. Έτσι, οι συχνές λέξεις που εμφανίζονται στο *W* λαμβάνονται υπόψη με βαρύτητα ανάλογη της συχνότητάς τους στα παράθυρα του όρου-στόχου *t*.

Το σκεπτικό πίσω από τη χρήση της ιδιότητας αυτής είναι ότι παράθυρα ενός όρου-στόχου που περιέχουν πολλές λέξεις που συνεμφανίζονται συχνά με τον όρο-στόχο είναι πιθανότερο να αποτελούν ορισμούς του όρου-στόχου.

Στην εργασία [2] είχε χρησιμοποιηθεί ως ιδιότητα και η θέση του εγγράφου από το οποίο προέρχεται το παράθυρο στη σειρά κατάταξης (ranked list) της μηχανής αναζήτησης. Ωστόσο η σειρά κατάταξης της μηχανής αναζήτησης του Βήματος δεν φαίνεται να παρέχει ιδιαίτερα χρήσιμες πληροφορίες, οπότε δε θεωρήθηκε σκόπιμο να ληφθεί υπόψη.

[1] Στην περίπτωση του εντοπισμού των παραθύρων των ονομάτων-στόχων (ενότητα 3.1), δεν κρίναμε σκόπιμο να αφαιρέσουμε τους τόνους από τα ονόματα-στόχους, γιατί τότε θα έπρεπε να αφαιρεθούν οι τόνοι και από όλες τις λέξεις των άρθρων που επιστρέφει η μηχανή αναζήτησης, ώστε να λειτουργήσει το ταίριασμα συμβολοσειρών, πράγμα που θα επιβράδυνε το σύστημα. επίσης, διατηρώντας τους τόνους γίνεται συχνά καλύτερη διάκριση ανάμεσα σε ανδρικά και γυναικεία ονόματα. Για παράδειγμα, οι όροι-στόχοι «Αγγελοπούλου» (η κυρία Αγγελοπούλου) και «Αγγελόπουλος» γίνονται και οι δύο «Αγγελοπουλ» μετά την αποκοπή καταλήξεων αν αφαιρεθούν οι τόνοι.

3.2.2 Υποψήφιες ιδιότητες που αντιστοιχούν σε n-γράμματα

Εκτός από τις δύο βασικές ιδιότητες, πειραματιστήκαμε με επιπλέον δυαδικές (Boolean) ιδιότητες, κάθε μία από τις οποίες δείχνει αν υπάρχει (τιμή 1) ή όχι (τιμή 0) ένα συγκεκριμένο n-γραμμα λέξεων (με $n \in \{1,2,3\}$), δηλαδή μια συγκεκριμένη ακολουθία συνεχόμενων λέξεων (μήκους 1, 2 ή 3 λέξεων), μέσα στο παράθυρο αμέσως πριν ή μετά τον όρο-στόχο. Αρχικά υπάρχει μία υποψήφια ιδιότητα αυτού του είδους για κάθε ένα n-γραμμα που εμφανίζεται αμέσως πριν ή μετά τον όρο-στόχο στα παράθυρα εκπαίδευσης

Για παράδειγμα από το παράθυρο εκπαίδευσης[1]: *αποπειράται το βιβλίο του Αρη Τσαντηρόπουλου Η βεντέτα στη σύγχρονη ορεινή Κρήτη (εκδόσεις Πλέθρον). * Η ποίηση του Κ.Π. 1Καβάφη$$ και η ρωσική ποίηση των αρχών του 20ού αιώνα προσφέρουν μια ακόμη μαρτυρία για τη συγγένεια των αυτόνομων και παράλληλων*

προκύπτουν 6 υποψήφιες ιδιότητες, οι οποίες δείχνουν αν εμφανίζεται αμέσως πριν ή μετά τον όρο-στόχο το καθένα από τα ακόλουθα n-γράμματα. (Στο συγκεκριμένο παράθυρο, και οι 6 υποψήφιες ιδιότητες θα είχαν τιμή 1 αλλά σε άλλα παράθυρα οι τιμές τους δεν θα ήταν πάντα 1.)

1. **Μία λέξη πριν:** Κ.Π
2. **Δύο λέξεις πριν:** του Κ.Π.
3. **Τρεις λέξεις πριν**: ποίηση του Κ.Π.
4. **Μία λέξη μετά:** και
5. **Δύο λέξεις μετά:** και η
6. **Τρεις λέξεις μετά:** και η ρωσική

Οι υποψήφιες ιδιότητες που προκύπτουν με τον παραπάνω τρόπο περιέχουν συχνά κύρια ονόματα. Έτσι, για παράδειγμα, ενδέχεται να έχουμε ξεχωριστές υποψήφιες ιδιότητες για τα 3-γράμματα «πρόεδρος της Γαλλίας <όνομα-στόχος>», «πρόεδρος της Γερμανίας <όνομα-στόχος>», «πρόεδρος της ΗΡ <όνομα-στόχος>», ενώ όλα αυτά τα 3-γράμματα αποτελούν υποπεριπτώσεις ενός γενικότερου 3-γράμματος «πρόεδρος της <όνομα> <όνομα-στόχος>». Επίσης, οι τρεις ξεχωριστές υποψήφιες

[1] Από τα κείμενα έχουν πια αφαιρεθεί οι ετικέτες της HTML.

ιδιότητες θα είχαν όλες τιμή 0 αν συναντούσαμε σε ένα νέο παράθυρο το 3-γραμμα «πρόεδρος της Liverpool <όνομα-στόχος>», δηλαδή δεν θα μας παρείχαν καμία ένδειξη για την ομοιότητα που παρουσιάζει το νέο 3-γραμμα με τα προηγούμενα. Είναι προτιμότερο, λοιπόν, τα κύρια ονόματα που εμφανίζονται στα n-γράμματα των υποψηφίων ιδιοτήτων να αντικαθίστανται από γενικότερες ετικέτες (π.χ. <όνομα οργανισμού>, <όνομα προσώπου> κ.λ.π.). Αυτό μπορεί να επιτευχθεί με τη χρήση ενός συστήματος αναγνώρισης κυρίων ονομάτων (named entity recognizer). Επειδή, όμως, δεν διαθέταμε τέτοιο σύστημα για τα ελληνικά, χρησιμοποιήσαμε έναν πιο απλοϊκό αλγόριθμο αναγνώρισης ονομάτων, που περιγράφεται στην επόμενη παράγραφο, επιχειρώντας να αντικαταστήσουμε με ετικέτες κύρια ονόματα που βρίσκονται κοντά στο όνομα-στόχο. Για την ακρίβεια, επικεντρώσαμε την προσπάθειά μας στο να αντικαταστήσουμε με ετικέτες κύρια ονόματα που βρίσκονται στα n-γράμματα που εμφανίζονται αμέσως πριν το όνομα-στόχο, επειδή τέτοια κύρια ονόματα εμφανίζονται συχνά και επηρεάζουν σημαντικά τη γενικότητα των υποψήφιων ιδιοτήτων.

Αν η λέξη που βρίσκεται αμέσως πριν από το όνομα-στόχο ξεκινά με κεφαλαίο γράμμα και έχει μήκος μεγαλύτερο του 1 χαρακτήρα, τότε η λέξη θεωρείται κύριο όνομα και αντικαθίσταται από την ετικέτα [y] (π.χ. το «Γαλλίας <όνομα-στόχος» γίνεται «[y] <όνομα-στόχος>»). Υπάρχει γενικά ο κίνδυνος η λέξη που βρίσκεται αμέσως πριν από το όνομα-στόχο να ξεκινά με κεφαλαίο όχι επειδή είναι κύριο όνομα αλλά επειδή είναι η πρώτη λέξη μιας περιόδου (π.χ. «. Ο <όνομα-στόχος>», που δεν θέλουμε να γίνει «. [y] <όνομα-στόχος>»). Σε αυτήν την περίπτωση, στην ελληνική γλώσσα η λέξη αμέσως πριν το όνομα-στόχο, δηλαδή η πρώτη λέξη της περιόδου, είναι σχεδόν πάντα ένα από τα άρθρα «Ο» ή «Η», αφού το όνομα-στόχος είναι πάντα όνομα προσώπου. Οπότε απαιτώντας το μήκος της λέξης αμέσως πριν το όνομα-στόχο να είναι μεγαλύτερο της μονάδος, αποφεύγουμε τον κίνδυνο να αντικαταστήσουμε με [y] το άρθρο του ονόματος-στόχου.

Επίσης, στην περίπτωση που η πρώτη λέξη πριν το όνομα-στόχο έχει γίνει [y] και η δεύτερη λέξη πριν το όνομα-στόχο ξεκινά και αυτή με κεφαλαίο γράμμα και πριν από αυτή δεν προηγείται τελεία, τότε γίνεται [y] και η δεύτερη λέξη πριν το όνομα-στόχο . Μάλιστα συνενώνονται σε ένα [y], αφού αποτελούν κοινό ονοματικό σύνολο, και η αρχικά τρίτη λέξη πριν το όνομα-στόχο γίνεται δεύτερη. Για παράδειγμα «ο πρόεδρος

της Hewlett Packard <όνομα-στόχος>» γίνεται «ο πρόεδρος της [y] <όνομα-στόχος>».

Επιπλέον, αν η λέξη που προηγείται του ονόματος-στόχου τελειώνει σε τελεία και ξεκινά με κεφαλαίο, τότε η λέξη αυτή παραλείπεται, αφού θεωρείται ότι αποτελεί μέρος του ονόματος-στόχου. Για παράδειγμα, το «η ποίηση του Κ.Π. <όνομα-στόχος>» στο παράθυρο που δόθηκε στην αρχή αυτής της ενότητας γίνεται «η ποίηση του <όνομα-στόχος>». Μετά από την εφαρμογή αυτού του κανόνα, όσα σημεία στίξης παραμένουν θεωρούνται αυτόνομες λέξεις. Για παράδειγμα, στο «ο μαθητής του Αριστοτέλη, Πλάτωνας ...», όπου το όνομα-στόχος είναι «Αριστοτέλης», το 1-γραμμα που τελειώνει αμέσως μετά το όνομα-στόχο είναι «,» και το 2-γραμμα «, Πλάτωνας».

Από όλες τις υποψήφιες ιδιότητες που αντιστοιχούν σε n-γράμματα, επιλέγονται τελικά οι καλύτερες, με τον τρόπο που περιγράφεται στην επόμενη ενότητα.

3.2.3 Αξιολόγηση υποψηφίων ιδιοτήτων

Για την αξιολόγηση των υποψήφιων ιδιοτήτων μπορούν να χρησιμοποιηθούν πολλά κριτήρια, όπως TF-IDF, πληροφοριακό κέρδος (information gain), αμοιβαία πληροφορία (mutual information), κριτήριο X^2, κ.λ.π. [6] Ακολουθώντας την εργασία [2], αξιολογούμε τις υποψήφιες ιδιότητες με το μέτρο της ακρίβειας (precision).[1] Στην περίπτωσή μας, η ακρίβεια ορίζεται ως το πηλίκο TP/(TP+FP), όπου TP (true positives) το πλήθος των παραθύρων εκπαίδευσης που περιέχουν το n-γραμμα της ιδιότητας και είναι αποδεκτές απαντήσεις (ορισμοί) και FP (false positives) το πλήθος των παραθύρων εκπαίδευσης που περιέχουν το n-γραμμα της ιδιότητας χωρίς να είναι αποδεκτές απαντήσεις (μη ορισμοί). Μέσω της ακρίβειας, λοιπόν, έχουμε μια εκτίμηση του πόσο πιθανό είναι ένα παράθυρο που περιέχει το n-γραμμα της ιδιότητας να αποτελεί και αποδεκτή απάντηση. Αφού υπολογισθεί η ακρίβεια για όλες τις υποψήφιες ιδιότητες, επιλέγονται οι καλύτερες m, όπου η «βέλτιστη» τιμή του m θα προκύψει από τα πειραματικά μας αποτελέσματα.

[1] Η εργασία [2] κατέληξε πειραματικά στο συμπέρασμα ότι η αξιολόγηση των ιδιοτήτων των παραθύρων με το μέτρο της ακρίβειας είναι προτιμότερη από τη χρήση του πληροφοριακού κέρδους.

Ένα πρόβλημα που εμφανίζεται κατά την αξιολόγηση με το μέτρο της ακρίβειας είναι ότι, για παράδειγμα, αν το n-γραμμα μίας υποψήφιας ιδιότητας έχει εμφανιστεί μία μόλις φορά στο σύνολο των παραθύρων εκπαίδευσης και έτυχε αυτό το παράθυρο να είναι και αποδεκτή απάντηση, η ακρίβεια της υποψήφιας ιδιότητας θα είναι 1. Αν το n-γραμμα μιας άλλης υποψήφιας ιδιότητας έχει εμφανιστεί σε 10 παράθυρα εκπαίδευσης, από τα οποία μόνο τα 8 είναι αποδεκτοί ορισμοί, τότε η ακρίβεια αυτής της υποψήφιας ιδιότητας θα είναι 0,8, οπότε θα προτιμηθεί η πρώτη ιδιότητα. Η εκτίμηση, όμως, για την ακρίβεια της πρώτης ιδιότητας δεν είναι αξιόπιστη, γιατί βασίζεται σε πολύ μικρό δείγμα εμφανίσεων του n-γράμματός της και εκτός αυτού οι ιδιότητες που αντιστοιχούν σε πολύ σπάνια n-γράμματα συνεισφέρουν ελάχιστα στο διαχωρισμό των παραθύρων που αποτελούν ορισμούς από εκείνα που δεν είναι ορισμοί, αφού έχουν τιμή 0 σε σχεδόν όλα τα παράθυρα. Για αυτό το λόγο, ως υποψήφιες ιδιότητες θεωρούνται μόνο εκείνες που αντιστοιχούν σε n-γράμματα που εμφανίζονται σε τουλάχιστον k παράθυρα εκπαίδευσης, όπου το k είναι ένα κατώφλι. Στα πειράματα της εργασίας, το k ήταν 3. Σε περίπτωση που θέλουμε να κρατήσουμε περισσότερες ιδιότητες από όσες περνούν το κατώφλι, αφού επιλέξουμε όλες τις ιδιότητες που περνούν το κατώφλι, μειώνουμε το κατώφλι κατά μία μονάδα και εφαρμόζουμε το μέτρο της ακρίβειας στις νέες υποψήφιες ιδιότητες που προκύπτουν.

3.3 Υλοποιήσεις SVM

Στο διαδίκτυο υπάρχουν πολλές υλοποιήσεις για τα SVM. Οι τρεις που αξιολογήσαμε ήταν:

- WEKA [7]
- SVM^{light} [8]
- libSVM [9]

Απ' αυτές τελικά επιλέξαμε τη libSVM.

3.3.1 WEKA (Waikato Environment for Knowledge Analysis)

Το WEKA είναι ένα περιβάλλον ανάπτυξης εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης, το οποίο αναπτύχθηκε στο πανεπιστήμιο του Waikato στη Νέα Ζηλανδία. Είναι γραμμένο σε Java, ώστε να μπορεί να χρησιμοποιηθεί με όσο το

δυνατόν περισσότερα λειτουργικά συστήματα, και διατίθεται ελεύθερα (συμπεριλαμβανομένου του πηγαίου κώδικα). Παρέχει ένα ευρύ σύνολο από υλοποιήσεις αλγορίθμων μηχανικής μάθησης (τόσο για κατηγοριοποίηση όσο και για συσταδοποίηση, clustering) καθώς και μηχανισμούς για προ-επεξεργασία δεδομένων και μετα-επεξεργασία αποτελεσμάτων. Ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει τις υλοποιήσεις των αλγορίθμων είτε από τη γραμμή εντολών είτε από το γραφικό περιβάλλον το οποίο προσφέρει το WEKA, ενώ ο προγραμματιστής μπορεί να καλέσει τις υλοποιήσεις των αλγορίθμων από τα δικά του προγράμματα. Έτσι το WEKA μπορεί να λειτουργήσει σαν μια βιβλιοθήκη υλοποιήσεων αλγορίθμων μηχανικής μάθησης, που μπορεί να χρησιμοποιηθεί για την δημιουργία νέων προγραμμάτων. Επίσης, καθώς παρέχει μια πλήρη βιβλιοθήκη με κώδικα για αξιολόγηση αποτελεσμάτων (π.χ. για διασταυρωμένη επικύρωση – βλ. επόμενες ενότητες), μπορούν πολύ εύκολα να συγκριθούν νέες μέθοδοι με ήδη υπάρχουσες.

Το WEKA προσφέρει μια υλοποίηση SVM (τάξη SMO) αλλά η υλοποίηση αυτή αποδείχτηκε πολύ αργή (ίσως λόγω της υλοποίησης σε Java). Επίσης, επιτρέπει τη χρήση μόνο πολυωνυμικού πυρήνα και RBF και δεν παρέχει αυτοματοποιημένο τρόπο επιλογής των παραμέτρων του SVM (π.χ. παράμετρος d του πολυωνυμικού πυρήνα, παράμετρος C του προβλήματος βελτιστοποίησης – βλ. ενότητα 2.2.4).

3.3.2 SVM^{light}

Το SVM^{light} είναι μια υλοποίηση των SVM σε C που παρέχεται ως βιβλιοθήκη κώδικα. Δίνεται ελεύθερα σε πηγαία και εκτελέσιμη μορφή για ερευνητικούς και εκπαιδευτικούς σκοπούς. Ωστόσο για την εμπορική της χρήση θα πρέπει πρώτα να υπάρχει συγκατάθεση του δημιουργού. Η υλοποίηση είναι ιδιαίτερα γρήγορη και παρέχει στο χρήστη πάρα πολλές επιλογές. Ο χρήστης μπορεί να φτιάξει ακόμα και το δικό του πυρήνα και να τον χρησιμοποιήσει χωρίς να αντιμετωπίσει προβλήματα. Επίσης, υπολογίζονται αυτόματα πολλά στατιστικά μεγέθη για την καλύτερη αξιολόγηση των αποτελεσμάτων. Όμως το πλήθος των παραμέτρων που προσφέρονται και το ότι δεν παρέχεται αυτοματοποιημένος τρόπος επιλογής τιμών των παραμέτρων ούτε σχετική γραφική διεπαφή καθιστούν δύσκολη τη ρύθμιση (tuning) του SVM. Ακόμα, καθώς είναι γραμμένο σε C, είναι δυσκολότερο να χρησιμοποιηθεί με διαφορετικά λειτουργικά συστήματα από ό,τι το Weka, αν και

προσφέρονται εκδόσεις του (σε μορφή πηγαίου και εκτελέσιμου κώδικα) για Solaris, Windows, Linux και Cygwin. Ωστόσο αποτελεί τη πληρέστερη υλοποίηση SVM από τις τρεις που δοκιμάσαμε.

3.3.3 libSVM

Η βιβλιοθήκη libSVM δίνεται ελεύθερα για οποιαδήποτε χρήση. Παρέχονται υλοποιήσεις της σε πολλές γλώσσες προγραμματισμού (C++, Java, C# .NET) και συνεχώς εξελίσσεται (κατά τη διάρκεια της εργασίας βγήκαν δύο νέες εκδόσεις, ενώ δημιουργήθηκαν και νέα εργαλεία για την καλύτερη χρήση της βιβλιοθήκης)

Το libSVM προσφέρει ένα πλήθος παραμέτρων (πιο περιορισμένο από εκείνο του SVM^{light} αλλά καλύπτει τις βασικές ανάγκες). Ο χρήστης μπορεί να επιλέξει ανάμεσα σε 4 βασικούς πυρήνες (γραμμικός, πολυωνυμικός, RBF, σιγμοειδής – βλ. ενότητα 2.2.3) και να καθορίσει τις παραμέτρους για κάθε έναν από αυτούς. Επίσης δίνεται η δυνατότητα να επιστρέφεται όχι μόνο η κατηγορία στην οποία κατατάσσει το SVM κάθε περίπτωση αλλά και ο βαθμός βεβαιότητας της απόφασης του SVM

Δεν υπάρχει γραφική διεπαφή για τη βιβλιοθήκη, ωστόσο οι δημιουργοί του libSVM παρέχουν κώδικα οπτικοποίησης των διανυσμάτων εκπαίδευσης και του διαχωριστή που μαθαίνει το SVM για διάφορες τιμές των παραμέτρων του.

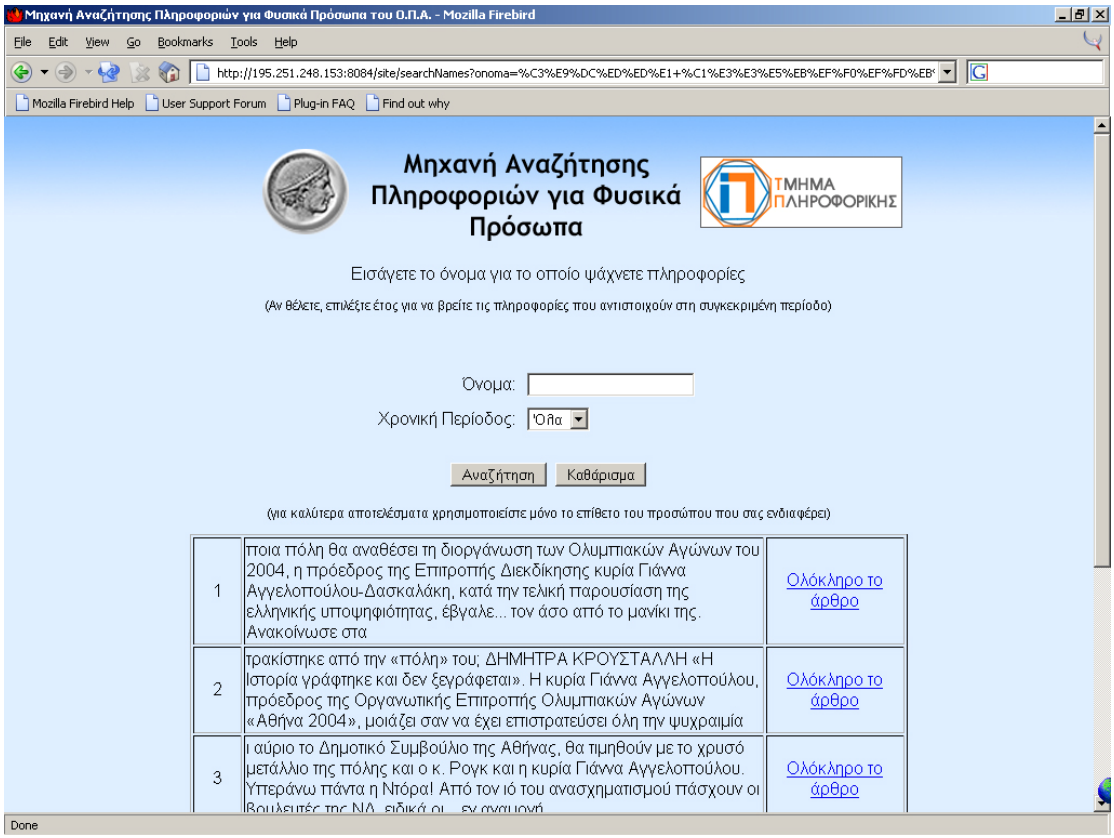
Στο σύστημά μας χρησιμοποιήσαμε τον πυρήνα RBF, που είναι αυτός που συνιστούν οι κατασκευαστές του LibSVM. Ο συγκεκριμένος πυρήνας απαιτεί να επιλεγούν οι τιμές 2 παραμέτρων (C και γ – βλ. ενότητα 2.2.3). Για την ευκολότερη επιλογή των τιμών αυτών των παραμέτρων, οι δημιουργοί του libSVM παρέχουν ένα script σε Python το οποίο χρησιμοποιεί έναν αλγόριθμο αναζήτησης grid-search για να βρει το «βέλτιστο» συνδυασμό τιμών των δύο παραμέτρων. Κάθε συνδυασμός τιμών αξιολογείται με διασταυρωμένη επικύρωση (cross-validation, βλ. ενότητα 4.1 παρακάτω) στα δεδομένα εκπαίδευσης. Αν και η διαδικασία αυτή είναι χρονοβόρα (ειδικά για μεγάλα σύνολα δεδομένων εκπαίδευσης), καθώς είναι αυτοματοποιημένη διευκολύνει πάρα πολύ την επιλογή τιμών των παραμέτρων.

3.4 Η on-line μορφή του συστήματός μας

Η ενότητα αυτή περιγράφει την on-line μορφή του συστήματος της εργασίας, δηλαδή τη μορφή που μπορεί να χρησιμοποιηθεί, μετά την εκπαίδευση του συστήματος, από χρήστες του Παγκόσμιου Ιστού.

Καθώς το σύστημά μας είναι γραμμένο σε Java, ένας εξυπηρετητής με TomCat αναλαμβάνει την παροχή της υπηρεσίας. Ο χρήστης αρχικά γράφει σε ένα textbox, όπως και σε κάθε άλλη μηχανή αναζήτησης, το όνομα για το οποίο ψάχνει πληροφορίες. Επειδή συχνά τα πρόσωπα έχουν πάνω από μία ιδιότητες, ανάλογα με τη χρονική στιγμή που εξετάζουμε, δίνεται στον επισκέπτη η δυνατότητα να επιλέξει από ένα listbox την χρονική περίοδο στην οποία θέλει να περιορίσει την αναζήτηση. Με αυτόν τον τρόπο το σύστημά μας θα ανατρέξει μόνο σε εκείνα τα άρθρα που γράφτηκαν τη συγκεκριμένη χρονική περίοδο.

Ο εξυπηρετητής, με τη μέθοδο get, παίρνει το ερώτημα του επισκέπτη και το στέλνει (με post) στον εξυπηρετητή της ιστοσελίδας του Βήματος. Αφού κατεβάσει τα 12 πρώτα κείμενα που επέστρεψε η μηχανή αναζήτησης της εφημερίδας, εκτελεί τον αλγόριθμό μας, επιλέγει τις 5 εκείνες συμβολοσειρές που θεωρεί ως καλύτερες απαντήσεις για το ερώτημα που του τέθηκε και τις εμφανίζει.



Στιγμιότυπο του on-line συστήματός μας. Ο χρήστης είχε εισαγάγει το όνομα ‘Γιάννα Αγγελοπούλου’

Η εκπαίδευση της on-line μορφής του συστήματός μας έχει γίνει στο σύνολο των εγγράφων που είχαμε στη διάθεσή μας (2400) και έχουν επιλεγεί οι 600 ιδιότητες με την υψηλότερη ακρίβεια.

4 Πειραματικά αποτελέσματα

'All that we are is the result of what we have thought'
(Buddha)

Στο κεφάλαιο αυτό θα περιγράψουμε τα πειράματα που πραγματοποιήθηκαν και τα αποτελέσματά τους.

4.1 Μεθοδολογία

Το σύστημα χρειάζεται μια συλλογή δεδομένων εκπαίδευσης και μια συλλογή δεδομένων αξιολόγησης. Και οι δύο συλλογές πρέπει να περιέχουν ονόματα-στόχους και τα αντίστοιχα παράθυρά τους από τα άρθρα που επέστρεψε η μηχανή αναζήτησης της εφημερίδας για κάθε όνομα-στόχο. Επειδή όμως το πλήθος των ονομάτων-στόχων που είχαμε ήταν σχετικά μικρό (200 ονόματα-στόχοι, βλ. ενότητα 3.1), δεν μπορούσαμε να χωρίσουμε απλά τα δεδομένα μας σε 2 μέρη, αφού έτσι και οι δύο συλλογές δεδομένων θα ήταν ακόμα μικρότερες. Γι' αυτό καταφύγαμε στην τεχνική της διασταυρωμένης επικύρωσης (cross-validation). Χωρίσαμε το σύνολο των δεδομένων μας (ονόματα-στόχοι και τα αντίστοιχα παράθυρά τους) σε 10 ίσα μέρη (10-fold cross validation) και επαναλάβαμε τα πειράματά μας 10 φορές. Σε κάθε επανάληψη τα διανύσματα που παριστάνουν τα παράθυρα των 9 μερών αποτελούν τα δεδομένα εκπαίδευσης του SVM και τα διανύσματα που παριστάνουν τα παράθυρα του 10ου μέρους αποτελούν τα δεδομένα αξιολόγησης. (Σε κάθε επανάληψη χρησιμοποιείται διαφορετικό μέρος για τα δεδομένα αξιολόγησης.) Έτσι όλα τα μέρη χρησιμοποιούνται τελικά τόσο για εκπαίδευση όσο και για αξιολόγηση, χωρίς όμως ποτέ το σύστημα να αξιολογείται σε δεδομένα που έχουν χρησιμοποιηθεί για την εκπαίδευσή του. Τα τελικά αποτελέσματα είναι ο μέσος όρος των αποτελεσμάτων των επαναλήψεων.

Η μέτρηση της επίδοσης του συστήματος έγινε με δύο τρόπους (ενότητα 2.1.2):

Ποσοστό ορθών πρώτων απαντήσεων: Στον πρώτο τρόπο λαμβάνουμε υπόψη μας σε κάθε ερώτηση (όνομα-στόχο) αξιολόγησης μόνο την κορυφαία απάντηση που μας

επιστρέφει το σύστημα, δηλαδή το παράθυρο για το οποίο το SVM είναι περισσότερο βέβαιο ότι αποτελεί ορισμό. Αν το παράθυρο αποτελεί αποδεκτό ορισμό, η απόκριση του συστήματος θεωρείται σωστή, διαφορετικά θεωρείται λανθασμένη. Σε κάθε επανάληψη της διασταυρωμένης επικύρωσης μετράμε το ποσοστό ορθών αποκρίσεων του συστήματος. Το τελικό αποτέλεσμα είναι ο μέσος όρος του ποσοστού ορθών αποκρίσεων στις 10 επαναλήψεις.

MRR: Στο δεύτερο τρόπο λαμβάνουμε σε κάθε ερώτηση αξιολόγησης υπόψη μας τις κορυφαίες 5 απαντήσεις του συστήματος (τα 5 παράθυρα για τα οποία το SVM είναι περισσότερο βέβαιο ότι αποτελούν ορισμούς) και αξιολογούμε την επίδοση του συστήματος με το μέτρο Mean Reciprocal Rank (MRR): Αν υπάρχει κάποια σωστή απάντηση (αποδεκτός ορισμός) στις 5 που επέστρεψε το σύστημα, τότε το σύστημα βαθμολογείται με 1/r, όπου r η θέση (ranking) της πρώτης (πιο πάνω) σωστής απάντησης. Αν δεν υπάρχει καμία σωστή απάντηση στις 5 που επέστρεψε το σύστημα, τότε το σύστημα βαθμολογείται με 0. Σε κάθε μία από τις 10 επαναλήψεις τις διασταυρωμένης επικύρωσης, το MRR του συστήματος είναι το άθροισμα των βαθμών που έλαβε στις ερωτήσεις αξιολόγησης δια τον αριθμό των ερωτήσεων αξιολόγησης. Το τελικό αποτέλεσμα είναι ο μέσος όρος των MRR των επαναλήψεων.

Σημειώνουμε ότι σε κάθε επανάληψη της διασταυρωμένης επικύρωσης υπολογίζονται εκ νέου οι «βέλτιστες» τιμές των παραμέτρων C και γ του SVM από τα δεδομένα εκπαίδευσης εκείνης της επανάληψης, χρησιμοποιώντας το σχετικό script των δημιουργών της βιβλιοθήκης libSVM (βλ. ενότητα 3.3.3).

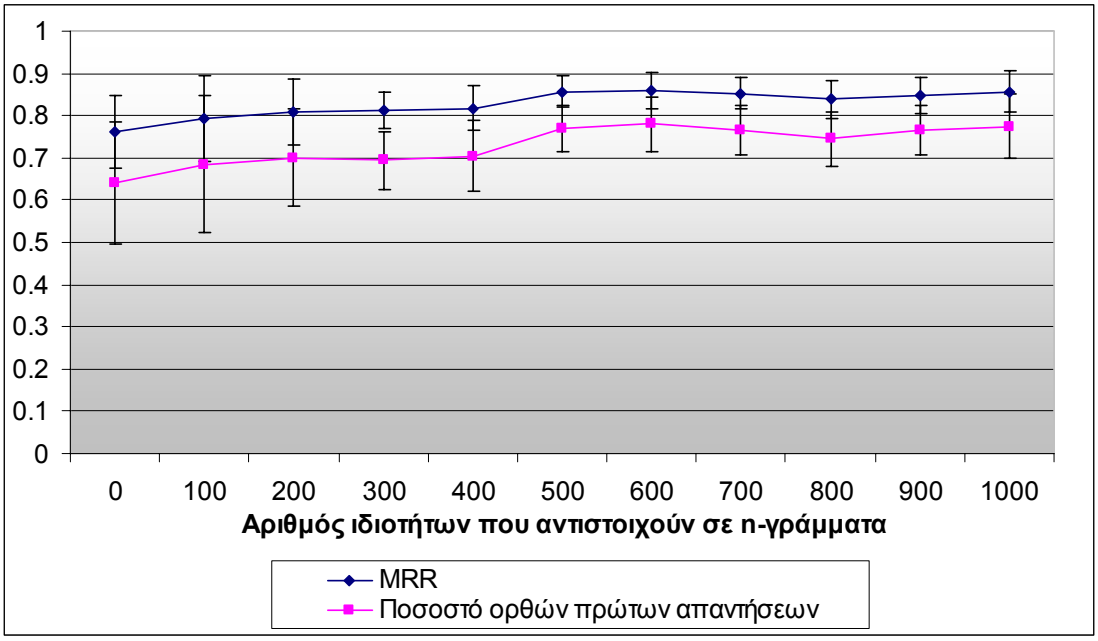
4.2 Πειράματα

4.2.1 Πειράματα εντοπισμού βέλτιστου πλήθους ιδιοτήτων

Αρχικά έπρεπε να βρούμε ποιος είναι ο αριθμός ιδιοτήτων που αντιστοιχούν σε n-γράμματα (βλ. ενότητες 3.2.2 και 3.2.3) ο οποίος οδηγεί στα καλύτερα αποτελέσματα. Πειραματισθήκαμε με 100, 200, 300, 400, 500, 600, 700, 800, 900 και 1000 ιδιότητες αυτού του είδους. Για κάθε αριθμό ιδιοτήτων επαναλάβαμε τη διασταυρωμένη επικύρωση και υπολογίσαμε τις μέσες τιμές των δύο μέτρων αξιολόγησης (ποσοστό ορθών πρώτων αποκρίσεων και MRR) καθώς και τη διακύμανσή τους. Ως μέθοδο

αναφοράς (baseline) χρησιμοποιήσαμε τη μορφή του συστήματος που δε χρησιμοποιεί καθόλου ιδιότητες που αντιστοιχούν σε n-γράμματα αλλά μόνο τις δύο βασικές ιδιότητες (ενότητα 3.2.1).

Τα αποτελέσματα φαίνονται στο παρακάτω διάγραμμα, όπου τα error bars έχουν μήκος ίσο με την τυπική απόκλιση των αποτελεσμάτων στις 10 επαναλήψεις της διασταυρωμένης επικύρωσης.



Αποτελέσματα με μεταβλητό αριθμό ιδιοτήτων που αντιστοιχούν σε n-γράμματα

Συγκεκριμένα τα αποτελέσματα που προέκυψαν είναι:

	0 ιδιότητες		100 ιδιότητες		200 ιδιότητες	
επανάληψη 1	0.55 (11)	0.72	0.45 (9)	0.66	0.5 (10)	0.7
επανάληψη 2	0.45 (9)	0.65	0.4 (8)	0.61	0.6 (12)	0.76
επανάληψη 3	0.6 (12)	0.75	0.75 (15)	0.84	0.65 (13)	0.76
επανάληψη 4	0.6 (12)	0.74	0.75 (15)	0.85	0.8 (16)	0.89
επανάληψη 5	0.6 (12)	0.76	0.65 (13)	0.82	0.75 (15)	0.85
επανάληψη 6	0.75 (15)	0.82	0.85 (17)	0.86	0.55 (11)	0.68
επανάληψη 7	0.55 (11)	0.72	0.7 (14)	0.8	0.8 (16)	0.9
επανάληψη 8	0.9 (18)	0.94	0.9 (18)	0.95	0.75 (15)	0.85
επανάληψη 9	0.55 (11)	0.67	0.6 (12)	0.72	0.8 (16)	0.84
επανάληψη 10	0.85 (17)	0.85	0.8 (16)	0.83	0.8 (16)	0.85

Μέση τιμή	0.64	0.762	0.685	0.794	0.7	0.808
Διακύμανση	0.021	0.007551	0.026694	0.01036	0.013333	0.006018

	300 ιδιότητες		400 ιδιότητες		500 ιδιότητες	
επανάληψη 1	0.6 (12)	0.75	0.55 (11)	0.75	0.7 (14)	0.833
επανάληψη 2	0.75 (15)	0.85	0.8 (16)	0.88	0.75 (15)	0.86
επανάληψη 3	0.7 (14)	0.8	0.65 (13)	0.76	0.75 (15)	0.83
επανάληψη 4	0.7 (14)	0.82	0.7 (14)	0.82	0.8 (16)	0.89
επανάληψη 5	0.75 (15)	0.85	0.7 (14)	0.79	0.7 (14)	0.8
επανάληψη 6	0.6 (12)	0.75	0.65 (13)	0.78	0.75 (15)	0.84
επανάληψη 7	0.65 (13)	0.8	0.75 (15)	0.86	0.85 (17)	0.88
επανάληψη 8	0.8 (16)	0.88	0.85 (17)	0.91	0.75 (15)	0.85
επανάληψη 9	0.75 (15)	0.83	0.7 (14)	0.8	0.8 (16)	0.85
επανάληψη 10	0.65 (13)	0.78	0.7 (14)	0.83	0.85 (17)	0.93
Μέση τιμή	0.695	0.811	0.705	0.818	0.77	0.8563
Διακύμανση	0.004694	0.001877	0.006917	0.002751	0.002889	0.001321

	600 ιδιότητες		700 ιδιότητες		800 ιδιότητες	
επανάληψη 1	0.8 (16)	0.88	0.8 (16)	0.88	0.75 (15)	0.84
επανάληψη 2	0.8 (16)	0.88	0.75 (15)	0.86	0.75 (15)	0.85
επανάληψη 3	0.7 (14)	0.8	0.75 (15)	0.82	0.7 (14)	0.8
επανάληψη 4	0.8 (16)	0.89	0.7 (14)	0.84	0.75 (15)	0.86
επανάληψη 5	0.65 (13)	0.78	0.8 (16)	0.85	0.65 (13)	0.78
επανάληψη 6	0.75 (15)	0.83	0.65 (13)	0.78	0.65 (13)	0.77
επανάληψη 7	0.85 (17)	0.88	0.85 (17)	0.9	0.85 (17)	0.9
επανάληψη 8	0.8 (16)	0.88	0.8 (16)	0.89	0.8 (16)	0.88
επανάληψη 9	0.8 (16)	0.85	0.75 (15)	0.83	0.75 (15)	0.83
επανάληψη 10	0.85 (17)	0.92	0.8 (16)	0.88	0.8 (16)	0.87
Μέση τιμή	0.78	0.859	0.765	0.853	0.745	0.838
Διακύμανση	0.004	0.001899	0.003361	0.001357	0.004139	0.001862

	900 ιδιότητες		1000 ιδιότητες	
επανάληψη 1	0.75 (15)	0.84	0.75 (15)	0.84
επανάληψη 2	0.85 (17)	0.9	0.85 (17)	0.92
επανάληψη 3	0.75 (15)	0.82	0.75 (15)	0.83
επανάληψη 4	0.8 (16)	0.88	0.8 (16)	0.88
επανάληψη 5	0.75 (15)	0.84	0.75 (15)	0.84
επανάληψη 6	0.65 (13)	0.76	0.6 (12)	0.75
επανάληψη 7	0.8 (16)	0.87	0.85 (17)	0.89
επανάληψη 8	0.8 (16)	0.88	0.8 (16)	0.88
επανάληψη 9	0.7 (14)	0.81	0.75 (15)	0.83
επανάληψη 10	0.8 (16)	0.88	0.85 (17)	0.91
Μέση τιμή	0.765	0.848	0.775	0.857
Διακύμανση	0.003361	0.001818	0.005694	0.00249

Αναλυτικά αποτελέσματα με μεταβλητό αριθμό ιδιοτήτων που αντιστοιχούν σε n-γράμματα

Παρατηρούμε, λοιπόν, ότι, ενώ η μέθοδος αναφοράς έχει πιθανότητα μόλις 64% να είναι σωστή η πρώτη απάντηση που επιστρέφει, χρησιμοποιώντας 100 ιδιότητες που αντιστοιχούν σε n-grams το ποσοστό ανεβαίνει στο 68,5%. Στη συνέχεια, παρ' ότι

υπάρχει μια διακύμανση των αποτελεσμάτων, φαίνεται πως γενικά υπάρχει μια ανοδική τάση καθώς αυξάνονται οι ιδιότητες και το ποσοστό ορθών πρώτων απαντήσεων για 600 ιδιότητες γίνεται 78%. Από το σημείο αυτό και πέρα βλέπουμε ότι τα αποτελέσματα της μεθόδου σταθεροποιούνται σε αυτό το ποσόστο. Ακόμα παρατηρούμε ότι η διακύμανση έχει επίσης μια πτωτική τάση: ενώ στη μέθοδο αναφοράς είναι 2,1, για 600 χαρακτηριστικά είναι 0,4.

Στην περίπτωση της δεύτερης μεθόδου αξιολόγησης (MRR), και πάλι τα αποτελέσματα έχουν την ίδια μορφή. Για 0 ιδιότητες n-γραμμάτων το ποσοστό επιτυχίας είναι 7,62 (με διακύμανση 0,75), ενώ για 100 και 600 είναι αντίστοιχα 7,94 (διακύμανση 1) και 8,59 (διακύμανση 0,19).

Παράδειγμα επιλεγμένων ιδιοτήτων για 100 ιδιότητες στην $10^{η}$ επανάληψη

πρόεδρο [y]	προέδρου [y]	μιλάει στο	ποίηση του	στρατηγός
Χρηματιστηρίου κ .	πρόεδρος	, στο	καθηγητής [y]	, μπορεί
του	υπουργός [y]	στη θέση του	πρωθυπουργός [y]	ΣΥΝ κ .
πρόεδρος του [y]	Ανάπτυξης κ .	, Πικαμπιά ,	» κ .	Πικαμπιά ,
« Η γέννηση	(ο	, όπου ο	Τάξης κ .	κάλεσε
σκηνοθετεί ο [y]	στρατηγού	μιλάει	, μια	, Μασόν
, υπεύθυνος του	πρόεδρο της [y]	πρωθυπουργό [y]	□ΜΖΉ	Πειραιά κ .
μιλάει στο«	σε μετάφραση του	« Madama Butterfly	είναι ο	βιβλίο του [y]
, πρόεδρος της	σκηνοθέτησε και πρωταγωνιστεί	ΟΤΕ κ .	και«	πρόεδρος της
Ελ Γκρέκο στον	. Κατά	, στον	σκηνοθέτησε	έγραψε ο
, ενώ ο	, ομότιμος	, ενός	όπερα του	σκηνοθέτης [y]
είναι καθηγητής Γλωσσολογίας	γεννήθηκε	, ως	ταινίες του [y]	, Μασόν,
επί	, ένας	στη θέση	Ί . □ΜΖΉ	σκηνοθεσία [y]
ΓΣΕΕ κ .	όπερας του	, ομότιμος καθηγητής	ήταν εκείνος που	σκηνοθέτησε και
πρόεδρος της [y]	» (	της όπερας του	ο υφυπουργός [y]	Αθηνών κ .
« Ο γύρος	ταινία του [y]	Ζαν - [y]	ΠαΣοΚ κ .	παρουσιάζει
και τους	Τύραννος » [y]	τραγούδια του [y]	Γκρέκο στον	Τουρισμού κ .
, αλλά και	είναι πρόεδρος	υπογράφει ο [y]	. Ο ίδιος	Κ
πρόεδρος [y]	. Παίζουν	προέδρου της [y]	« Madama	πρόεδρος της Alpha
πρόεδρός της [y]	ήταν εκείνος	μια	. Παίζουν:	. □ΜΖΉ

Παρατηρούμε ότι στο σύνολο των 100 ιδιοτήτων, κάποιες ιδιότητες φαίνονται άσχετες, για παράδειγμα «Madama Butterfly». Αυτό συμβαίνει επειδή το μέτρο αξιολόγησης ιδιοτήτων που χρησιμοποιήσαμε επιβραβεύει εκείνες τις ιδιότητες των οποίων τα n-γράμματα εμφανίζονται συχνά σε παράθυρα ορισμού και δεν εμφανίζονται συχνά σε παράθυρα μη-ορισμού. Στο παράδειγμα του «Madama Butterfly», στις ερωτήσεις εκπαίδευσης περιέχεται ο Πουτσίνι και το «Madama Butterfly» εμφανιζόταν σε αρκετά παράθυρα ορισμού του Πουτσίνι. Αντίθετα δεν εμφανίστηκε ποτέ σε κάποιο παράθυρο μη ορισμού, με αποτέλεσμα η ακρίβειά της ιδιότητας που αντιστοιχεί στο «Madama Butterfly» να είναι ιδιαίτερα υψηλή. Το πρόβλημα είναι πως επιλέγονται έτσι και ιδιότητες που αντιστοιχούν σε πολύ σπάνια n-γράμματα. Για να αντιμετωπιστεί αυτό το πρόβλημα χρησιμοποιήσαμε το κατώφλι εμφανίσεων k (ενότητα 3.2.3). Επειδή, όμως, το σύνολο των παραθύρων εκπαίδευσης (σε κάθε επανάληψη της διασταυρωμένης επικύρωσης) ήταν σχετικά περιορισμένο, η τιμή του k ήταν αναγκαστικά μικρή και η συγκεκριμένη ιδιότητα τελικά κατάφερε να το ξεπεράσει.

Ακόμα βλέπουμε ότι έχουν επιλεγεί ιδιότητες όπως «πρόεδρος του [y]», «πρόεδρος της [y]» κ.α., που αντιστοιχούν σε n-γράμματα των οποίων η εμφάνιση σε ένα παράθυρο αποτελεί σημαντική ένδειξη πως πρόκειται για παράθυρο ορισμού. Ίσως θα μπορούσαμε να ομαδοποιήσουμε ιδιότητες αυτού του είδους, χρησιμοποιώντας ένα σύστημα εντοπισμού μερών του λόγου (part-of-speech tagger), έτσι ώστε να αντιστοιχούν όλες μαζί σε ένα πρότυπο της μορφής «[άρθρο] [y]».

4.2.2 Ενδιαφέροντα παραδείγματα απαντήσεων

Παρακάτω φαίνονται οι απαντήσεις που επέστρεψε το σύστημα στην τέταρτη επανάληψη της διασταυρωμένης επικύρωσης, για την ερώτηση «Κουκ», με 100 και 600 ιδιότητες:

100 ιδιότητες	600 ιδιότητες
ν κεντρική αντίφαση της ελληνικής κληρονομιάς. ΤΣΑΡΛΣ ΦΡΙΜΑΝ - Το ελληνικό επίτευγμα Μετάφραση Μαίρη Περαντάκου-**Κουκ**, Εκδόσεις Κέδρος, σελ. 784, τιμή 28 ευρώ φάκελος Να καταδείξουν πώς λειτούργησαν και πού μας πήγαν οι Αγώνες της Α	μόνες τους. Πολύ ουσιαστικότερο είναι το μήνυμα που στέλνεται στο ΝΑΤΟ, παρατηρούν ο βρετανός πρώην υπουργός Ρόμπερτ **Κουκ** και ο αμερικανός στρατηγός Γουέσλι Κλαρκ. Ο πρώτος θεωρεί ως «υπονόμευση» της Συμμαχίας τη μεταφορά του ναυτικού αρχηγείου του
μόνες τους. Πολύ ουσιαστικότερο είναι το μήνυμα που στέλνεται στο ΝΑΤΟ, παρατηρούν ο βρετανός πρώην υπουργός Ρόμπερτ **Κουκ** και ο αμερικανός στρατηγός Γουέσλι Κλαρκ. Ο πρώτος θεωρεί ως «υπονόμευση» της Συμμαχίας τη μεταφορά του ναυτικού αρχηγείου του	

Το σύστημα με τις 100 ιδιότητες χρειάστηκε 2 απαντήσεις για να απαντήσει ορθά «Ποιος είναι ο Κουκ;». Αντίθετα το σύστημα με τις 600 ιδιότητες είχε επιτυχία με την πρώτη απάντηση. Χαρακτηριστικό είναι ότι και τα δύο συστήματα επιστρέφουν την ίδια σωστή απάντηση. Ωστόσο το δεύτερο σύστημα εξετάζοντας μεγαλύτερο εύρος ιδιοτήτων δίνει μεγαλύτερη βαρύτητα στις ιδιότητα ‘και ο’ και ‘υπουργός [y]’ από την ύπαρξη του ‘-’ πριν από το όνομα.

100 ιδιότητες	600 ιδιότητες
ντευξη με τον Κλάιβ Μπάρκερ, τον μοναδικό συγγραφέα που απείλησε κάποτε το βασίλειο του Στίβεν Κινγκ. *** Ο Ιρβιν **Γουέλς** θα βρίσκεται στην Αθήνα στις 18 και 19 Ιανουαρίου - σας το λέω από τώρα για να	βουλευτής κ. Φώτης Κουβέλης. * Ο συγγραφέας Ερβιν **Γουέλς** (Trainspotting) θα παρουσιάσει το έργο του και θα διαλέξει μουσική στο Bios, Πειραιώς 84, στις 9 μ.μ. Η είσοδος είναι ελεύθερη

μην προγραμματίσετε κάτι άλλο... ΤΟ ΒΗΜΑ	ς ημέρες μεταξύ των ιδρυτικών κρατών-μελών τους.
ι του «Ταξιτζή», ο οποίος έχει επηρεάσει τα μέγιστα γενιές ολόκληρες νεότερών του κινηματογραφιστών. Οπως ο Ορσον **Γουέλς** και ο Αλφρεντ Χίτσκοκ, ο Σκορσέζε μέχρι στιγμής ανήκει στους «αδικημένους» και, αν δεν βραβευθεί εφέτος, ίσως να μη βραβευθεί	

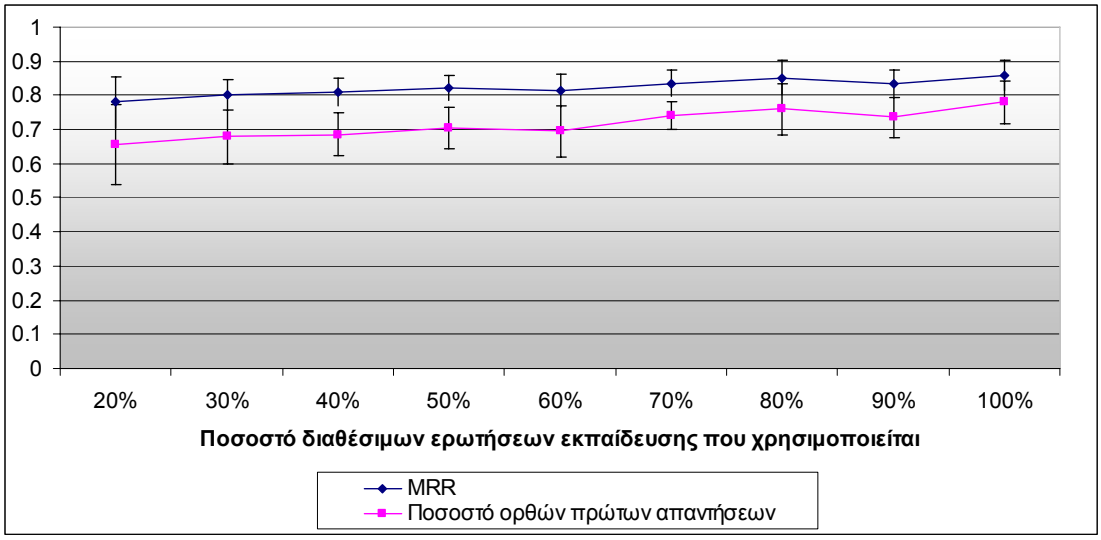
Στο παραπάνω παράδειγμα, όπου η ερώτηση ήταν «Γουέλς», βλέπουμε ότι το σύστημα με τις 100 ιδιότητες χρειάστηκε και πάλι δυο προσπάθειες, ενώ αυτό με τις 600 χρειάστηκε μόλις μία. Στο δεύτερο παράθυρο του πρώτου συστήματος δίνεται μεγαλύτερη βαρύτητα στην εμφάνιση των n-γράμματος ‘Όπως ο’ και ‘και ο’. Αντίθετα το δεύτερο σύστημα θεωρεί ότι οι παρενθέσεις που ακολουθούν το όνομα είναι καλύτερη ένδειξη ότι πρόκειται για ορισμό, αφού μέσα σε παρενθέσεις μετά από ένα όνομα δίνεται συχνά η ιδιότητα του αντίστοιχου προσώπου. Και οι δύο απαντήσεις είναι σωστές, αν και στην περίπτωση του δεύτερου συστήματος αυτό φαίνεται να οφείλεται στην τύχη. Μία άλλη παρατήρηση είναι ότι αναζητώντας πληροφορίες για το όνομα «Γουέλς», το σύστημα μας επέστρεψε ορισμούς για διάφορα άτομα με το ίδιο επίθετο. Η χρησιμοποίηση και του μικρού ονόματος, εκτός από το επώνυμο, ώστε να τα διακρίνουμε, θα δημιουργούσε αρκετές δυσκολίες στο ταίριασμα συμβολοσειρών, που χρησιμοποιείται για τον εντοπισμό του παραθύρου του ονόματος-στόχου: σε άλλα κείμενα θα εμφανίζεται πρώτα το μικρό όνομα και έπειτα το επίθετο, σε άλλα η σειρά θα είναι αντίστροφη και σε άλλα κείμενα δε θα εμφανίζεται καθόλου το μικρό όνομα.

4.2.3 Πειράματα με μεταβλητό μέγεθος ερωτήσεων εκπαίδευσης

Στο δεύτερο στάδιο των πειραμάτων μας θέλαμε να δούμε πώς μεταβάλλονται οι επιδόσεις του συστήματος όταν μεταβάλλεται ο αριθμός των ερωτήσεων εκπαίδευσης. Γι’ αυτό το λόγο ξανακάναμε τα πειράματά μας για 600 ιδιότητες n-γραμμάτων, αλλά χρησιμοποιώντας μόνο το 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% και 100% των διαθεσίμων ερωτήσεων εκπαίδευσης σε κάθε επανάληψη της διασταυρωμένης επικύρωσης, δηλαδή 40, 60, 80, 100, 120, 140, 160, 180 και 200 ερωτήσεις αντίστοιχα. Ο αριθμός των ερωτήσεων αξιολόγησης παρέμεινε

αμετάβλητος, χρησιμοποιούσαμε, δηλαδή, σε κάθε επανάληψη και τις 20 ερωτήσεις αξιολόγησης.

Τα αποτελέσματα για 600 ιδιότητες είναι:



Αποτελέσματα με μεταβλητό αριθμό ερωτήσεων εκπαίδευσης για 600 ιδιότητες (το 100% αντιστοιχεί σε 200 ερώτησεις εκπαίδευσης)

Τα αναλυτικά αποτελέσματα για 600 ιδιότητες είναι:

	20% (40)		30% (60)		40% (80)	
επανάληψη 1	0.6 (12)	0.78	0.6 (12)	0.77	0.65 (13)	0.8
επανάληψη 2	0.5 (10)	0.7	0.6 (12)	0.75	0.7 (14)	0.82
επανάληψη 3	0.5 (10)	0.68	0.75 (15)	0.82	0.65 (13)	0.77
επανάληψη 4	0.55 (11)	0.71	0.6 (12)	0.77	0.7 (14)	0.79
επανάληψη 5	0.7 (14)	0.78	0.75 (15)	0.82	0.75 (15)	0.84
επανάληψη 6	0.65 (13)	0.79	0.7 (14)	0.83	0.7 (14)	0.85
επανάληψη 7	0.75 (15)	0.85	0.6 (12)	0.75	0.6 (12)	0.77
επανάληψη 8	0.85 (17)	0.92	0.8 (16)	0.89	0.7 (14)	0.84
επανάληψη 9	0.7 (14)	0.76	0.65 (13)	0.8	0.8 (16)	0.87
επανάληψη 10	0.75 (15)	0.83	0.75 (15)	0.82	0.6 (12)	0.74

Μέση τιμή	0.655	0.78	0.68	0.802	0.685	0.809
Διακύμανση	0.013583	0.005422	0.006222	0.001884	0.003917	0.001743

	50% (100)		60% (120)		70% (140)	
επανάληψη 1	0.7 (14)	0.83	0.7 (14)	0.84	0.75 (15)	0.86
επανάληψη 2	0.65 (13)	0.8	0.7 (14)	0.83	0.75 (15)	0.87
επανάληψη 3	0.75 (15)	0.82	0.75 (15)	0.81	0.75 (15)	0.82
επανάληψη 4	0.65 (13)	0.79	0.6 (12)	0.77	0.65 (13)	0.78
επανάληψη 5	0.7 (14)	0.82	0.75 (15)	0.86	0.75 (15)	0.85
επανάληψη 6	0.6 (12)	0.75	0.55 (11)	0.72	0.7 (14)	0.79
επανάληψη 7	0.7 (14)	0.84	0.75 (15)	0.88	0.8 (16)	0.89
επανάληψη 8	0.8 (16)	0.88	0.65 (13)	0.8	0.75 (15)	0.85
επανάληψη 9	0.75 (15)	0.85	0.7 (14)	0.78	0.75 (15)	0.8

επανάληψη 10	0.75 (15)	0.83	0.8 (16)	0.86	0.75 (15)	0.85

Μέση τιμή	0.705	0.821	0.695	0.815	0.74	0.836
Διακύμανση	0.003583	0.001254	0.005806	0.002406	0.001556	0.001338

	80% (160)		90% (180)		100% (200)	
επανάληψη 1	0.8 (16)	0.89	0.75 (15)	0.85	0.8 (16)	0.88
επανάληψη 2	0.85 (17)	0.91	0.75 (15)	0.86	0.8 (16)	0.88
επανάληψη 3	0.75 (15)	0.82	0.75 (15)	0.82	0.7 (14)	0.8
επανάληψη 4	0.75 (15)	0.85	0.6 (12)	0.78	0.8 (16)	0.89
επανάληψη 5	0.65 (13)	0.78	0.7 (14)	0.81	0.65 (13)	0.78
επανάληψη 6	0.7 (14)	0.81	0.75 (15)	0.82	0.75 (15)	0.83
επανάληψη 7	0.9 (18)	0.94	0.75 (15)	0.84	0.85 (17)	0.88
επανάληψη 8	0.75 (15)	0.87	0.8 (16)	0.89	0.8 (16)	0.88
επανάληψη 9	0.75 (15)	0.82	0.7 (14)	0.78	0.8 (16)	0.85
επανάληψη 10	0.7 (14)	0.83	0.8 (16)	0.89	0.85 (17)	0.92

Μέση τιμή	0.76	0.852	0.735	0.834	0.78	0.859
Διακύμανση	0.005444	0.002484	0.003361	0.00156	0.004	0.001899

Αναλυτικά αποτελέσματα με μεταβλητό αριθμό ερωτήσεων εκπαίδευσης για 600 ιδιότητες

Από τα αποτελέσματα παρατηρούμε ότι αρχικά όσο αυξάνεται το μέγεθος της συλλογής εκπαίδευσης (αριθμός ερωτήσεων εκπαίδευσης) βελτιώνονται και οι επιδόσεις του συστήματός μας, Ωστόσο από ένα σημείο και μετά (στο 80% των ερωτήσεων –160 ερωτήσεις) έχουν αρχίσει να εμφανίζονται σημάδια κορεσμού (ειδικά στο δείκτη MRR η μεταβολή από το 80% στο 100% είναι μόλις 0,7%).

5 Μελλοντικές επεκτάσεις και βελτιώσεις

'Let's boldly go where no man has gone before'
(Star Trek)

Η μέθοδός μας φαίνεται να λειτουργεί ικανοποιητικά για τις ερωτήσεις ορισμού φυσικών προσώπων στην ελληνική γλώσσα και κείμενα εφημερίδων. Όταν επιτρέπεται μόνο μία απάντηση ανά ερώτηση, τα ποσοστά επιτυχίας φτάνουν το 78%, ενώ όταν επιτρέπονται 5 απαντήσεις το μέτρο MRR αγγίζει το 86%. Μια μελλοντική εργασία θα μπορούσε να διερευνήσει τις επιδόσεις του συστήματος για διαφορετικές τιμές για το κατώφλι (ενότητα 3.2.3). Θα ήταν ενδιαφέρον, επίσης, η αναζήτηση των ορισμών να γίνεται σε αρχεία περισσότερων εφημερίδων, κάτι που όπως προαναφέρθηκε είναι εύκολο να γίνει για εφημερίδες που οι ιστότοποί τους παρέχουν δυνατότητες αναζήτησης παλαιών άρθρων με λέξεις-κλειδιά.

Μια άλλη πιθανή βελτίωση θα ήταν να εφαρμόζεται συσταδοποίηση (clustering) στις κορυφαίες απαντήσεις (π.χ. τις πρώτες 10) που παράγει το σύστημα σε κάθε ερώτηση, ώστε απαντήσεις που μοιάζουν πολύ (π.χ. έχουν πολλές κοινές λέξεις) να θεωρούνται ίδιες και οι βαθμοί βεβαιότητάς τους να αθροίζονται. Μπορεί, για παράδειγμα, η 3η και $5^{η}$ απάντηση του συστήματος σε κάποια ερώτηση να είναι πολύ παρόμοιες, επειδή λένε το ίδιο πράγμα και αν αθροιστούν οι βαθμοί βεβαιότητας που έχει για αυτές το SVM ο συνολικός βαθμός βεβαιότητας των δύο απαντήσεων να ξεπερνά αυτόν της $1^{ης}$ απάντησης, οπότε να προτιμηθούν η $3^{η}$ και $5^{η}$ απάντηση (ή ένας αντιπρόσωπος αυτών), στην περίπτωση όπου επιτρέπεται μόνο μία απάντηση.

Ακόμα, τόσο ο αλγόριθμος αποκοπής καταλήξεων όσο και αυτός της αναγνώρισης ονομάτων οντοτήτων θα μπορούσαν να βελτιωθούν. Ειδικότερα, για την αναγνώριση ονομάτων προσώπων, θα μπορούσε να χρησιμοποιηθεί το σύστημα της εργασίας [3], το οποίο έχει εκπαιδευθεί σε άρθρα της ίδιας εφημερίδας.

Μία άλλη κατεύθυνση που θα ήταν ενδιαφέρον να εξεταστεί είναι να αυξηθεί το μέγιστο μήκος των n-γραμμάτων που αντιστοιχούνται σε ιδιότητες. Αυτό θα αύξανε κατά πολύ τον αριθμό των υποψηφίων ιδιοτήτων αλλά θα μπορούσε να οδηγήσει

στον εντοπισμό περισσότερων χρήσιμων προτύπων (patterns) που είναι ενδεικτικά παραθύρων ορισμών ή μη ορισμών. Τέλος, οι τιμές των δυαδικών ιδιοτήτων θα μπορούσαν να αντικατασταθούν από τις αντίστοιχες τιμές TF-IDF, ενώ η επιλογή των ιδιοτήτων θα μπορούσε να βασίζεται σε ένα συνδυασμό περισσοτέρων μέτρων αξιολόγησης ιδιοτήτων (ενότητα 3.2.3)· για παράδειγμα να επιλέγονται αρχικά οι πρώτες 500 ιδιότητες βάσει της ακρίβειας και έπειτα να χρησιμοποιούνται από αυτές οι 100 με το μεγαλύτερο πληροφοριακό κέρδος.

Αναφορές

[1] Voorhees Ellen M. «Overview of the TREC 2001 Question Answering Track», National Institute of Standards and Technology. 2001

[2] Σπ. Μηλιαράκη. «Χειρισμός Ερωτήσεων Ορισμού σε Συστήματα Ερωταποκρίσεων». Αθήνα 2003

[3] Γ. Λουκαρέλλι. «Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα». Αθήνα 2005

[4] Δ. Γαλάνης. «Αυτόματη κατασκευή παραδειγμάτων εκπαίδευσης για το χειρισμό ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων που χρησιμοποιούν μηχανική μάθηση». Αθήνα 2004

[5] http://tovima.dolnet.gr

[6] http://www.db-net.aueb.gr/index.php/corporate/courses/data_web_mining

[7] http://www.cs.waikato.ac.nz/~ml/

[8] http://svmlight.joachims.org/

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Παράρτημα

Οι συνηθισμένες λέξεις που αφαιρούνται (stop-words) είναι οι ακόλουθες. Έχουν ήδη αποκοπεί οι καταλήξεις (stemming).

0	ειν	{	εν	&	ουτ	²
1	τοτ	}		^	μειτ	%
2	αυτ	:	«	,	οπ	$
3	nbsp	|	·	.	ν	#
4	μερικ	\	κ	?	ταδ	@
5	Στ	/	η	;	δ	!
6	Μ	(	ή	~	$$	<
7	Απ	)	α	=	1	>
8	Πρ	+	ο	"	0	εχ
9	θ	-	τ	'	γ	ουδ
[	μον	_	αλλ	`	τωρ	μηδ
]	εν	*	ειτ	²°	ειχ	μαλλ
ω	π	οπ				