

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
εισ. 81304
Αρ.
ταξ.



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ**

Διπλωματική Εργασία

Μεταπτυχιακού Διπλώματος Ειδίκευσης

**«Ανάπτυξη συστήματος ερωταποκρίσεων για αρχεία
ελληνικών εφημερίδων»**

Μαρία-Ελένη Κολλιάρου

Επιβλέπων: Ιων Ανδρουτσόπουλος

Αθήνα, Ιούνιος 2007

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ**



ΠΕΡΙΕΧΟΜΕΝΑ

Περιεχόμενα.....	2
Περίληψη.....	4
1 Εισαγωγή.....	5
1.1 Αντικείμενο και στόχοι της εργασίας.....	5
1.2 Ευχαριστίες	6
2 Θεωρητικό υπόβαθρο.....	7
2.1 Αρχιτεκτονική συστημάτων ερωταποκρίσεων.....	7
2.1.1 Ανάλυση της ερώτησης.....	7
2.1.2 Ανάκτηση σχετικών εγγράφων.....	8
2.1.3 Επεξεργασία ανακτηθέντων εγγράφων.....	8
2.1.4 Εξαγωγή υποψηφίων απαντήσεων.....	8
2.1.5 Αξιολόγηση των υποψηφίων απαντήσεων.....	8
2.1.6 Επιλογή υποψήφιας απάντησης ή απαντήσεων.....	9
2.2 Υπάρχοντα συστήματα ερωταποκρίσεων.....	9
2.3 Μηχανές διανυσμάτων υποστήριξης.....	11
3 Το σύστημα της εργασίας	12
3.1 Ανάλυση ερωτήσεων.....	12
3.2 Ανάκτηση σχετικών εγγράφων.....	12
3.3 Επεξεργασία ανακτηθέντων εγγράφων.....	12
3.4 Εξαγωγή υποψηφίων απαντήσεων.....	13
3.5 Αξιολόγηση υποψηφίων απαντήσεων	13
3.6 Επιλογή υποψήφιας απάντησης ή απαντήσεων.....	19
4 Πειραματικά αποτελέσματα.....	21
4.1 Συλλογή δεδομένων	21
4.2 Διασταυρωμένη επικύρωση	22
4.3 Πειράματα.....	22

5	Τελική μορφή του συστήματος.....	30
6	Συμπεράσματα, Μελλοντικές επεκτάσεις.....	33
6.1	<i>Σύνοψη.....</i>	33
6.2	<i>Μελλοντικές επεκτάσεις</i>	33
7	Αναφορές	35



Περίληψη

Στη διάρκεια αυτής της εργασίας αναπτύχθηκε ένα σύστημα ερωταποκρίσεων φυσικής γλώσσας για αρχεία ελληνικών εφημερίδων. Το σύστημα υποστηρίζει ελληνικές ερωτήσεις προσώπων, οργανισμών και χρονικές ερωτήσεις. Επί του παρόντος, αναζητεί τις απαντήσεις των ερωτήσεων στα αρχεία των εφημερίδων «Τα Νέα» και «Το Βήμα», όπως διατίθενται μέσω των ιστότοπων των εφημερίδων, αλλά μπορεί να επεκταθεί, ώστε να χρησιμοποιεί και τα αρχεία άλλων εφημερίδων.

Το σύστημα χρησιμοποιεί τις υπάρχουσες μηχανές αναζήτησης των ιστότοπων των εφημερίδων. Σε κάθε ερώτηση, υποψήφιες απαντήσεις είναι τα ονόματα προσώπων, οργανισμών ή οι χρονικές εκφράσεις, ανάλογα με την κατηγορία της ερώτησης, που εντοπίζονται στα κείμενα που επέστρεψαν οι μηχανές αναζήτησης. Η κατηγορία της ερώτησης είναι δυνατόν να προσδιορίζεται είτε χειρωνακτικά από τον ίδιο το χρήστη, είτε αυτόματα χρησιμοποιώντας λογισμικό προηγούμενης εργασίας, το οποίο κατατάσσει ελληνικές ερωτήσεις σε κατηγορίες. Οι υποψήφιες απαντήσεις εντοπίζονται επιστρατεύοντας λογισμικό προηγούμενων εργασιών, το οποίο εντοπίζει ονόματα προσώπων, οργανισμών και χρονικές εκφράσεις σε ελληνικά κείμενα. Μεταξύ των υποψηφίων απαντήσεων, το σύστημα επιλέγει αυτές που θεωρεί καλύτερες, χρησιμοποιώντας Μηχανές Διανυσμάτων Υποστήριξης που εκπαιδεύονται να διαχωρίζουν τις ορθές υποψήφιες απαντήσεις από τις λανθασμένες. Τα πειραματικά αποτελέσματα της εργασίας δείχνουν ότι το σύστημα καταφέρνει να απαντήσει σωστά στο 65% των ερωτήσεων προσώπων, το 58% των ερωτήσεων οργανισμών και το 49% των χρονικών ερωτήσεων, όταν επιτρέπονται πέντε απαντήσεις ανά ερώτηση.

1 ΕΙΣΑΓΩΓΗ

1.1 ΑΝΤΙΚΕΙΜΕΝΟ ΚΑΙ ΣΤΟΧΟΙ ΤΗΣ ΕΡΓΑΣΙΑΣ

Καθώς ο όγκος των δεδομένων που υπάρχουν στο διαδίκτυο γίνεται όλο και μεγαλύτερος, η εύρεση πληροφοριών σε αυτό με χρήση των υπαρχουσών μηχανών αναζήτησης γίνεται όλο και δυσκολότερη. Για την εύρεση της επιθυμητής πληροφορίας, ο χρήστης του διαδικτύου πρέπει να υποβάλει λέξεις-κλειδιά σε κάποια μηχανή αναζήτησης και στη συνέχεια να εντοπίσει την πληροφορία ο ίδιος στα έγγραφα που επιστρέφει η μηχανή. Τα συστήματα ερωταποκρίσεων (Question Answering Systems) [1, 2, 3] επιχειρούν να διευκολύνουν το χρήστη, επιτρέποντάς του να εισάγει ερωτήσεις φυσικής γλώσσας, αντί για λέξεις-κλειδιά, και επιστρέφοντας ακριβείς απαντήσεις και συνδέσμους προς τα έγγραφα που τις περιέχουν, αντί για ταξινομημένες λίστες εγγράφων. Οι απαντήσεις που επιστρέφονται μπορεί να είναι μεμονωμένες λέξεις, φράσεις ή σύντομα αποσπάσματα κειμένου.

Αντικείμενο αυτής της εργασίας ήταν η κατασκευή ενός συστήματος ερωταποκρίσεων, το οποίο να εντοπίζει σε ηλεκτρονικά αρχεία ελληνικών εφημερίδων που διατίθενται στον Παγκόσμιο Ιστό απαντήσεις σε τρεις κατηγορίες ερωτήσεων: α) ερωτήσεις των οποίων η απάντηση είναι όνομα προσώπου, β) ερωτήσεις των οποίων η απάντηση είναι όνομα οργανισμού και γ) ερωτήσεις των οποίων η απάντηση είναι χρονική έκφραση. Το σύστημα αυτής της εργασίας χρησιμοποιεί το σύστημα αναγνώρισης ονομάτων οντοτήτων (named entity recognizer) των εργασιών [4], [5], [6] και [7], καθώς και το σύστημα κατάταξης ερωτήσεων σε κατηγορίες της εργασίας [8].

Πιο συγκεκριμένα, το σύστημα που αναπτύχθηκε στη διάρκεια αυτής της εργασίας δέχεται από το χρήστη μια ερώτηση στα ελληνικά, η οποία πρέπει να ανήκει σε κάποια από τις τρεις προαναφερθείσες κατηγορίες, και στη συνέχεια, χρησιμοποιεί τις μηχανές αναζήτησης των εφημερίδων «Τα Νέα»¹ και «Το Βήμα»², προκειμένου να βρει άρθρα τα οποία να σχετίζονται με την ερώτηση. Υποψήφιες απαντήσεις είναι όλες οι εκφράσεις του τύπου που ζητά η ερώτηση (π.χ. ονόματα προσώπων, αν η ερώτηση ζητά όνομα προσώπου) οι οποίες περιέχονται στα άρθρα που επέστρεψαν οι μηχανές αναζήτησης. Για τον εντοπισμό της καλύτερης από τις υποψήφιες απαντήσεις, το σύστημα χρησιμοποιεί μια Μηχανή Διανυσμάτων Υποστήριξης (MΔΥ, Support Vector Machine - SVM), η οποία κατατάσσει τις υποψήφιες απαντήσεις σε σωστές και λανθασμένες, την κάθε μία με κάποιο βαθμό βεβαιότητας. Τελικά, το σύστημα επιστρέφει στο χρήστη την υποψήφια απάντηση για την οποία η MΔΥ ήταν πιο βέβαιη ότι ανήκει στην κατηγορία των σωστών απαντήσεων.

Τα πειραματικά αποτελέσματα του συστήματος διαφέρουν ανάλογα με την κατηγορία της ερώτησης. Το μεγαλύτερο ποσοστό επιτυχίας επιτεύχθηκε στην κατηγορία των ερωτήσεων προσώπου (65%), ακολουθούμενο από το ποσοστό επιτυχίας των ερωτήσεων οργανισμών (58%), ενώ το μικρότερο ποσοστό επιτυχίας μετρήθηκε στην κατηγορία των χρονικών ερωτήσεων (49%).

¹ Βλ. <http://www.tanea.gr/Find.aspx?d=20070606>

² Βλ. <http://tovima.dolnet.gr/search.php>

Το επόμενο κεφάλαιο παραθέτει περισσότερες πληροφορίες για τα συστήματα ερωταποκρίσεων και τις μεθόδους μηχανικής μάθησης που χρησιμοποιήθηκαν στην εργασία. Κατόπιν, το τρίτο κεφάλαιο περιγράφει το σύστημα της εργασίας, ενώ το τέταρτο παρουσιάζει τα πειράματα της εργασίας και τα αποτελέσματά τους. Τέλος, το πέμπτο κεφάλαιο συνοψίζει τα αποτελέσματα της εργασίας και προτείνει μελλοντικές επεκτάσεις της.

1.2 ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της εργασίας κ. Ιωνα Ανδρουτσόπουλο για την καθοδήγησή του καθ' όλη τη διάρκεια της εργασίας και την πολύτιμη βοήθειά του σε όλα τα προβλήματα που αντιμετώπισα. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Θεόδωρο Καλαμπούκη που δέχτηκε το ρόλο του δεύτερου αξιολογητή της εργασίας.

Ακόμα θα ήθελα να ευχαριστήσω το Γιώργο Λουκαρέλλι για τη βοήθειά του σε θέματα της libSVM και σε βελτιώσεις που χρειάστηκε να γίνουν στο σύστημα αναγνώρισης ονομάτων (ΣΑΟΟ), προκειμένου να χρησιμοποιηθεί στην εργασία. Για τη βοήθειά του στη βελτίωση του ΣΑΟΟ ευχαριστώ επίσης τον Ξενοφώντα Βασιλάκο. Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα τον Παντελή Φραγκούδη για τη βοήθειά του στη δημιουργία του κώδικα ανάκτησης των κειμένων από τα αρχεία των εφημερίδων και την επίβλεψη των πειραμάτων μου στο εργαστήριο.

2 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΩΝ ΕΡΩΤΑΠΟΚΡΙΣΕΩΝ

Ένα σύστημα ερωταποκρίσεων μπορεί είτε να επιστρέφει αυτούσια κομμάτια από κείμενα της συλλογής στην οποία αναζητεί την απάντηση (π.χ. ονόματα, ημερομηνίες ή ολόκληρες προτάσεις), είτε να παράγει το ίδιο την απάντηση χρησιμοποιώντας τεχνικές παραγωγής φυσικής γλώσσας. Στην εργασία αυτή περιοριζόμαστε στην πρώτη (και πιο διαδεδομένη) περίπτωση. Επίσης, συχνά επιτρέπεται στο σύστημα ερωταποκρίσεων να επιστρέψει περισσότερες από μία απαντήσεις ανά ερώτηση. Για παράδειγμα, στους διαγωνισμούς του Question-Answering Track [9] του διεθνούς συνεδρίου TREC (Text Retrieval and Evaluation Conference), επιτρέπονταν κατά καιρούς μέχρι πέντε απαντήσεις ανά ερώτηση.

Ένα τυπικό σύστημα ερωταποκρίσεων αποτελείται από μονάδες (modules), οι οποίες υλοποιούν τις εξής λειτουργίες:

2.1.1 Ανάλυση της ερώτησης

Η ερώτηση που δίνει ο χρήστης αναλύεται και προκύπτουν γι' αυτή πληροφορίες, όπως οι όροι της (οι κυριότερες λέξεις της, εξαιρώντας άρθρα, συνδέσμους και άλλες πολύ συχνές λέξεις), καθώς και η κατηγορία στην οποία ανήκει η ερώτηση και, επομένως, ο τύπος της απάντησης που αναμένεται. Οι ερωτήσεις μπορούν γενικά να διακριθούν σε τρεις κατηγορίες, χρησιμοποιώντας ως κριτήριο τον τύπο της απάντησης:

- ❖ *Eρωτήσεις των οποίων η απάντηση είναι αυστηρά καθορισμένη (factual questions)*
Οι ερωτήσεις αυτές διαιρούνται περαιτέρω σε υποκατηγορίες, ανάλογα με το είδος της απάντησης που απαιτούν, για παράδειγμα:
 - όνομα προσώπου, π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδας;»,
 - όνομα οργανισμού, π.χ. «Σε ποιο κόμμα είναι αρχηγός ο Κ. Καραμανλής?»,
 - χρονική έκφραση, π.χ. «Πότε έγιναν οι Ολυμπιακοί Αγώνες στην Ελλάδα?»,
 - τοποθεσία, π.χ. «Πού βρίσκεται ο πύργος του Αϊφελ?»,
 - ποσότητα, π.χ. «Πόσα χρόνια ήταν η Ελλάδα υπό γερμανική κατοχή?»,
 - ορισμός, π.χ. «Τι είναι η τεχνητή νοημοσύνη?».
- ❖ *Eρωτήσεις γνώμης (opinion questions),* π.χ. «Τι θα μπορούσε να αποτελέσει αιτία ενός Τρίτου Παγκοσμίου Πολέμου?».
- ❖ *Eρωτήσεις περιληψης (summary questions),* π.χ. «Ποια είναι η υπόθεση της τριλογίας “Ο άρχοντας των δαχτυλιδιών”?».

Η παρούσα εργασία ασχολείται μόνο με ερωτήσεις των οποίων η απάντηση είναι σαφώς καθορισμένη. Μεταξύ αυτών, όπως προαναφέρθηκε, η εργασία επικεντρώνεται στις ερωτήσεις προσώπων, οργανισμών και χρονικών

εκφράσεων, που αντιστοιχούν στα είδη των εκφράσεων που υποστηρίζει το σύστημα αναγνώρισης ονομάτων οντοτήτων που χρησιμοποιήθηκε.

2.1.2 Ανάκτηση σχετικών εγγράφων

Οι όροι της ερώτησης υποβάλλονται στη μηχανή αναζήτησης την οποία χρησιμοποιεί το σύστημα, στην περίπτωσή μας τη μηχανή αναζήτησης μιας εφημερίδας. Η μηχανή αναζήτησης επιστρέφει συνήθως μια ταξινομημένη λίστα από έγγραφα που θεωρεί σχετικά με τους όρους. Από αυτά, επιλέγονται τα X πρώτα, δηλαδή τα X πιο σχετικά με την ερώτηση, όπου το X είναι ένας προκαθορισμένος αριθμός.

2.1.3 Επεξεργασία ανακτηθέντων εγγράφων

Τα έγγραφα που ανακτήθηκαν ενδέχεται να χρειάζονται κάποιου είδους επεξεργασία, προκειμένου να χρησιμοποιηθούν στη συνέχεια από το σύστημα. Τέτοιους είδους επεξεργασία μπορεί να είναι, για παράδειγμα, η αφαίρεση ετικετών HTML, ο διαχωρισμός περιόδων (sentence splitting), η συντακτική ανάλυσή τους, η αναγνώριση ονομάτων οντοτήτων κ.ά. Στο σύστημα της εργασίας, η επεξεργασία περιλαμβάνει την αφαίρεση ετικετών HTML, το διαχωρισμό περιόδων και την αναγνώριση ονομάτων προσώπων, ονομάτων οργανισμών και χρονικών εκφράσεων.

2.1.4 Εξαγωγή υποψηφίων απαντήσεων

Από τα επεξεργασμένα έγγραφα επιλέγονται οι συμβολοσειρές οι οποίες είναι πιθανό να αποτελούν ή να περιέχουν τη ζητούμενη απάντηση. Στο σύστημα της εργασίας, αν η ερώτηση είναι, για παράδειγμα, ερώτηση προσώπου, υποψήφιες απαντήσεις είναι εν γένει τα ονόματα προσώπων των επεξεργασμένων εγγράφων, με κάποιους επιπλέον περιορισμούς που θα παρουσιαστούν σε επόμενα κεφάλαια.

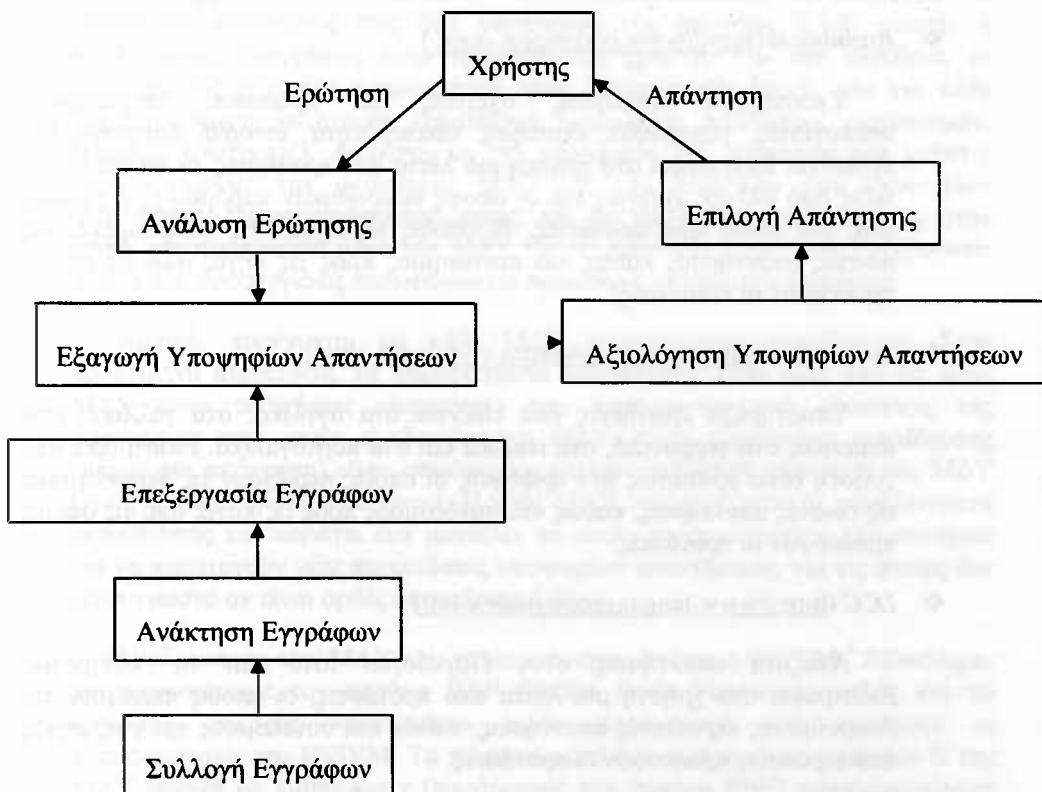
2.1.5 Αξιολόγηση των υποψηφίων απαντήσεων

Οι υποψήφιες απαντήσεις αξιολογούνται από το σύστημα και ταξινομούνται κατά φθίνουσα σειρά καταλληλότητας. Στο σύστημα της εργασίας, όπως προαναφέρθηκε, το στάδιο αυτό χρησιμοποιεί μια Μηχανή Διανυσμάτων Υποστήριξης (ΜΔΥ), η οποία κατατάσσει τις υποψήφιες απαντήσεις σε ορθές και λανθασμένες, επιστρέφοντας και ένα βαθμό βεβαιότητας για την κατάταξη κάθε υποψήφιας απάντησης. (Ο βαθμός βεβαιότητας είναι ουσιαστικά μια κανονικοποιημένη μορφή της απόστασης της διανυσματικής αναπαράστασης της υποψήφιας απάντησης από το υπερεπίπεδο διαχωρισμού της ΜΔΥ.) Καταλληλότερες θεωρούνται οι υποψήφιες απαντήσεις για τις οποίες η ΜΔΥ ήταν βεβαιότερη ότι αποτελούν ορθές απαντήσεις. Για την ακρίβεια, χρησιμοποιούνται τρεις διαφορετικές ΜΔΥ, μία για κάθε υποστηριζόμενη κατηγορία ερωτήσεων, όπως θα εξηγηθεί περαιτέρω στα επόμενα κεφάλαια.

2.1.6 Επιλογή υποψήφιας απάντησης ή απαντήσεων

Ανάλογα με το πόσες απαντήσεις επιτρέπεται να επιστρέψει το σύστημα, επιλέγεται ο αντίστοιχος αριθμός απαντήσεων από την κορυφή της ταξινομημένης λίστας των υποψηφίων απαντήσεων του προηγούμενου σταδίου. Οι επιλεγόμενες απαντήσεις επιστρέφονται στο χρήστη, συνοδευόμενες πιθανώς από τους αντίστοιχους βαθμούς βεβαιότητας του συστήματος και συνδέσμους προς τα έγγραφα από τα οποία προέρχονται.

Η Εικόνα 1 συνοψίζει την αρχιτεκτονική ενός τυπικού συστήματος ερωταποκρίσεων.



Εικόνα 1: Τυπική αρχιτεκτονική συστημάτων ερωταποκρίσεων

2.2 ΥΠΑΡΧΟΝΤΑ ΣΥΣΤΗΜΑΤΑ ΕΡΩΤΑΠΟΚΡΙΣΕΩΝ

Τα πρώτα συστήματα ερωταποκρίσεων αναπτύχθηκαν τη δεκαετία του '60. Μεταξύ αυτών ήταν το BASEBALL και το LUNAR [10]. Το πρώτο απαντούσε ερωτήσεις σχετικές με το πρωτάθλημα baseball στις Η.Π.Α., ενώ το δεύτερο απαντούσε ερωτήσεις σχετικές με τη γεωλογική σύνθεση ορυκτών από τις αποστολές των διαστημοπλοίων της σειράς Απόλλων στη σελήνη. Τα συστήματα αυτά, όπως και πολλά νεότερα [11], αναζητούσαν απαντήσεις σε βάσεις δεδομένων. Αντίθετα, τα τελευταία χρόνια το ενδιαφέρον εστιάζεται σε συστήματα ερωταποκρίσεων για συλλογές εγγράφων ή τον Παγκόσμιο Ιστό.

Στη συνέχεια, αναφέρονται κάποια από τα υπάρχοντα συστήματα ερωταποκρίσεων φυσικής γλώσσας για τον Παγκόσμιο Ιστό και παρουσιάζονται ορισμένα χαρακτηριστικά τους. Πολλά από τα συστήματα αυτά είναι εμπορικά προϊόντα, οπότε δεν είναι εύκολο να μάθει κανείς πώς ακριβώς λειτουργούν.

❖ *START* (<http://start.csail.mit.edu/>)

Υποστηρίζει ερωτήσεις σχετικές με γεωγραφία, επιστήμη, τέχνη, επικαιρότητα, ιστορία και πολιτισμό. Επιστρέφει στο χρήστη μια πρόταση, η οποία περιέχει τη θεωρούμενη ως σωστή απάντηση, καθώς και ένα σύνδεσμο προς την πηγή, από την οποία προέκυψε η πρόταση.

❖ *Brainboost* (<http://www.brainboost.com/>)

Υποστηρίζει ερωτήσεις σχετικές με πρόσωπα, επιχειρήσεις, υπολογιστές, γεωγραφία, επιστήμη, επικαιρότητα, ιστορία, διατροφή και προϊόντα. Επιστρέφει στο χρήστη μια λίστα από ερωτήσεις, οι οποίες έχουν τεθεί από άλλους χρήστες και οι οποίες θεωρήθηκαν παρόμοιες με τη δική του, μια λίστα από προτάσεις, οι οποίες περιέχουν τις θεωρούμενες ως σωστές απαντήσεις, καθώς και συνδέσμους προς τις πηγές από τις οποίες προέκυψαν οι προτάσεις.

❖ *AnswerBus* (<http://www.answerbus.com/index.shtml>)

Υποστηρίζει ερωτήσεις που τίθενται στα αγγλικά, στα γαλλικά, στα ισπανικά, στα γερμανικά, στα ιταλικά και στα πορτογαλικά. Επιστρέφει στο χρήστη δέκα προτάσεις ανά ερώτηση, οι οποίες περιέχουν τις θεωρούμενες ως σωστές απαντήσεις, καθώς και συνδέσμους προς τις πηγές από τις οποίες προέκυψαν οι προτάσεις.

❖ *LCC* (<http://www.languagecomputer.com/>)

Αναζητά απαντήσεις στον Παγκόσμιο Ιστό και τη Wikipedia. Επιστρέφει στο χρήστη μια λίστα από προτάσεις, οι οποίες περιέχουν τις θεωρούμενες ως σωστές απαντήσεις, καθώς και συνδέσμους προς τις πηγές από τις οποίες προέκυψαν οι προτάσεις.

❖ *ASU* (<http://qa.wpcarey.asu.edu/>)

Επιστρέφει στο χρήστη μια λίστα από τμήματα κειμένου, τα οποία περιέχουν τις θεωρούμενες ως σωστές απαντήσεις, καθώς και συνδέσμους προς τις πηγές από τις οποίες προέκυψαν τα τμήματα.

❖ *asked* (<http://asked.jp/>)

Υποστηρίζει ερωτήσεις που τίθενται στα αγγλικά, στα ιαπωνικά, στα κινέζικα, στα ρώσικα και στα σουηδικά. Επιστρέφει στο χρήστη πέντε απαντήσεις ανά ερώτηση, τις οποίες θεωρεί σωστές, τις προτάσεις που τις περιέχουν, καθώς και συνδέσμους προς τις πηγές από τις οποίες προέκυψαν.

❖ *NSIR* (<http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi>)

Ο χρήστης μπορεί να επιλέξει αν το σύστημα θα αναζητήσει την απάντηση στην ερώτησή του στον Παγκόσμιο Ιστό ή σε αρχεία, τα οποία υπάρχουν ήδη αποθηκευμένα στο σύστημα και αφορούν ερωτήσεις των διαγωνισμών του TREC. Επιπλέον, μπορεί να επιλέξει τον αριθμό των θεωρούμενων ως σωστών απαντήσεων που θα του επιστρέψει το σύστημα.

2.3 ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ

Η εργασία χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση (supervised machine learning) και συγκεκριμένα Μηχανές Διανυσμάτων Υποστήριξης (MΔΥ, Support Vector Machines - SVMs), προκειμένου να κατατάξει τις υποψήφιες απαντήσεις στις δύο κατηγορίες της ενότητας 2.1.5: σωστές ή λανθασμένες απαντήσεις στην ερώτηση του χρήστη. Για την ακρίβεια, το σύστημα της εργασίας χρησιμοποιεί τρεις διαφορετικές MΔΥ, μία για κάθε υποστηριζόμενη κατηγορία ερωτήσεων (ερωτήσεις προσώπων, οργανισμών, χρονικές ερωτήσεις). Ανάλογα με την κατηγορία της ερώτησης του χρήστη, χρησιμοποιούμε την αντίστοιχη MΔΥ για την κατάταξη των υποψηφίων απαντήσεων Πειραματιστήκαμε, όμως, και με μια δεύτερη προσέγγιση, στην οποία χρησιμοποιείται μόνο μία MΔΥ για τις ερωτήσεις όλων των κατηγοριών. Και οι δύο προσέγγισεις περιγράφονται περαιτέρω σε επόμενα κεφάλαια.

Αρχικά, παρέχονται σε κάθε MΔΥ παραδείγματα εκπαίδευσης. Στην προκειμένη περίπτωση, τα παραδείγματα εκπαίδευσης κάθε μίας από τις τρεις MΔΥ είναι υποψήφιες απαντήσεις που προέρχονται από ερωτήσεις της κατηγορίας στην οποία εξειδικεύεται η MΔΥ. Κάθε παράδειγμα εκπαίδευσης (υποψήφια απάντηση) είναι σημειωμένο με την επιθυμητή απόκριση της MΔΥ (ορθή ή λανθασμένη απάντηση). Η κάθε MΔΥ επεξεργάζεται τα παραδείγματα εκπαίδευσης και παράγει ένα μοντέλο, το οποίο χρησιμοποιείται στη συνέχεια για να καταταγούν νέες περιπτώσεις υποψηφίων απαντήσεων, για τις οποίες δεν είναι γνωστό αν είναι ορθές απαντήσεις ή όχι.

Η υλοποίηση των MΔΥ που χρησιμοποιήσαμε είναι η libSVM.³ Ειδικότερα, χρησιμοποιήσαμε τον πυρήνα RBF (Radial Basis Function), καθώς και το εργαλείο επιλογής «βέλτιστων» τιμών παραμέτρων που παρέχουν οι κατασκευαστές της libSVM. Το τελευταίο επιλέγει τιμές των παραμέτρων C της MΔΥ (ανοχή σε λάθη) και γ (παράμετρος του πυρήνα RBF) χρησιμοποιώντας αναζήτηση πλέγματος (grid search) και διασταυρωμένη επικύρωση (cross-validation) στα δεδομένα εκπαίδευσης.

Για περισσότερες πληροφορίες σχετικά με τη μηχανική μάθηση και τις Μηχανές Διανυσμάτων Υποστήριξης, ο αναγνώστης παραπέμπεται στα [12], και [13]. Συνοπτική εισαγωγή στις Μηχανές Διανυσμάτων Υποστήριξης περιέχεται στην εργασία [4].

³ Βλ. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

3 ΤΟ ΣΥΣΤΗΜΑ ΤΗΣ ΕΡΓΑΣΙΑΣ

Το σύστημα της εργασίας ακολουθεί την αρχιτεκτονική του προηγούμενου κεφαλαίου. Οι επόμενες ενότητες περιγράφουν αναλυτικότερα τα στάδια της λειτουργίας του συστήματος.

3.1 ΑΝΑΛΥΣΗ ΕΡΩΤΗΣΕΩΝ

Για κάθε ερώτηση, στη διάρκεια αυτού του σταδίου εκτελούνται τα παρακάτω βήματα:

- ❖ Προσδιορίζεται η κατηγορία της ερώτησης. Η κατηγορία είναι δυνατόν να προσδιοριστεί χρησιμοποιώντας το σύστημα κατάταξης ερωτήσεων της εργασίας [8]. Εναλλακτικά, είναι δυνατόν να προσδιοριστεί από τον ίδιο το χρήστη, μέσω της γραφικής διεπαφής του συστήματος (βλ. ενότητα 5).
- ❖ Αφαιρούνται από την ερώτηση οι πολύ συχνές λέξεις (stop-words), όπως άρθρα, σύνδεσμοι κλπ. Οι λέξεις που παραμένουν θεωρούνται όροι (terms) της ερώτησης. Για παράδειγμα, στην ερώτηση «Ποιος είναι ο πρωθυπουργός της Ελλάδας», οι όροι είναι «πρωθυπουργός» και «Ελλάδας».

3.2 ΑΝΑΚΤΗΣΗ ΣΧΕΤΙΚΩΝ ΕΓΓΡΑΦΩΝ

Στη διάρκεια αυτού του σταδίου, πραγματοποιούνται τα εξής:

- ❖ Το σύστημα στέλνει στις μηχανές αναζήτησης των εφημερίδων τους όρους της ερώτησης. Όπως προαναφέρθηκε, η τρέχουσα μορφή του συστήματος χρησιμοποιεί τις μηχανές αναζήτησης των εφημερίδων «Τα Νέα» και «Το Βήμα», αλλά πολύ εύκολα μπορεί να προσαρμοσθεί, ώστε να χρησιμοποιεί τις μηχανές αναζήτησης και άλλων εφημερίδων.
- ❖ Από τα κείμενα που επιστρέφει η κάθε μηχανή, επιλέγονται τα 10 πρώτα, θεωρώντας ότι οι μηχανές αναζήτησης επιστρέφουν πρώτα τα κείμενα που θεωρούν πιο σχετικά, κάτι που φαίνεται να ισχύει. Αν μια μηχανή επιστρέψει λιγότερα από 10 κείμενα, τότε απλά επιλέγονται όλα όσα επέστρεψε.

3.3 ΕΠΕΞΕΡΓΑΣΙΑ ΑΝΑΚΤΗΘΕΝΤΩΝ ΕΓΓΡΑΦΩΝ

Στη διάρκεια αυτού του σταδίου:

- ❖ Αφαιρούνται από τα κείμενα που επιλέχθηκαν στο προηγούμενο στάδιο οι ετικέτες HTML ή/και αντικαθίστανται από ετικέτες XML που απαιτεί το σύστημα αναγνώρισης ονομάτων οντοτήτων. Πιο συγκεκριμένα, οι ετικέτες <html> και </html> αντικαθίστανται από τις ετικέτες <ARTICLE> και </ARTICLE>, όλες οι ετικέτες κεφαλίδων αντικαθίστανται από ετικέτες <SUBTITLE> και </SUBTITLE>, οι ετικέτες <body> και </body> αντικαθίστανται από ετικέτες <BODY> και </BODY> κι όλες οι ετικέτες <p> αντικαθίστανται από ετικέτες <PARAGRAPH>. Οι ετικέτες </p> και όλες οι άλλες ετικέτες HTML αφαιρούνται.

- ❖ Τα κείμενα δίνονται κατόπιν στο σύστημα αναγνώρισης ονομάτων οντοτήτων, το οποίο σημειώνει με ετικέτες XML τα ονόματα προσώπων και οργανισμών, καθώς και τις χρονικές εκφράσεις. Το ίδιο σύστημα περιλαμβάνει έναν διαχωριστή περιόδων, ο οποίος προσθέτει ετικέτες XML που δείχνουν τα όρια των περιόδων. Για παράδειγμα, το κείμενο:

Η πρώτη επίσκεψη ως πρωθυπουργός του Κώστα Καραμανλή στην Κύπρο την ερχόμενη Παρασκευή δεν πρόκειται να αλλοιώσει τη συνειδητή επιλογή της Λευκωσίας να συντηρεί ένα ομιχλώδες τοπίο γύρω από τη στάση που θα τηρήσει στις 17 Δεκεμβρίου για τον καθορισμό ημερομηνίας ενταξιακών συνομιλιών της ΕΕ με την Τουρκία.

γίνεται:

<SENTENCE> Η πρώτη επίσκεψη ως πρωθυπουργός του <ENAMEX TYPE = "PERSON" CONF0 = "0.986072850630536" CONF1 = "0.9960208904332629"> Κώστα Καραμανλή </ENAMEX> στην Κύπρο την <TIMEX TYPE = "DATE"> ερχόμενη Παρασκευή </TIMEX> δεν πρόκειται να αλλοιώσει τη συνειδητή επιλογή της Λευκωσίας να συντηρεί ένα ομιχλώδες τοπίο γύρω από τη στάση που θα τηρήσει στις <TIMEX TYPE = "DATE"> 17 Δεκεμβρίου </TIMEX> για τον καθορισμό ημερομηνίας ενταξιακών συνομιλιών της <ENAMEX TYPE="ORGANIZATION" CONF0 = "0.945272923185093"> ΕΕ </ENAMEX> με την <ENAMEX TYPE = "ORGANIZATION" CONF0 = "0.013706128505628246"> Τουρκία </ENAMEX>. </SENTENCE>.

Οι ετικέτες των ονομάτων περιέχουν επιπλέον ιδιότητες (CONF0, CONF1, κλπ.) που δείχνουν πόσο βέβαιο είναι το σύστημα αναγνώρισης ονομάτων οντοτήτων ότι η κάθε λεκτική μονάδα του ονόματος όντως αποτελεί μέρος ονόματος της συγκεκριμένης κατηγορίας.

3.4 ΕΞΑΓΩΓΗ ΥΠΟΨΗΦΙΩΝ ΑΠΑΝΤΗΣΕΩΝ

Στις ερωτήσεις που ζητούν ονόματα προσώπων, υποψήφιες απαντήσεις είναι τα ονόματα προσώπων που εντόπισε το σύστημα αναγνώρισης ονομάτων οντοτήτων μέσα στα επιλεγέντα κείμενα της ενότητα 3.2. Αντιστοίχως ορίζονται οι υποψήφιες απαντήσεις στις περιπτώσεις των ερωτήσεων που ζητούν ονόματα οργανισμών ή χρονικές εκφράσεις. Στη διάρκεια των πειραμάτων δοκιμάσαμε και επιπλέον περιορισμούς κατά την επιλογή των υποψηφίων απαντήσεων, οι οποίοι περιγράφονται στο επόμενο κεφάλαιο.

3.5 ΑΞΙΟΛΟΓΗΣΗ ΥΠΟΨΗΦΙΩΝ ΑΠΑΝΤΗΣΕΩΝ

Στο στάδιο αυτό χρησιμοποιούνται οι τρεις ΜΔΥ που προαναφέρθηκαν, μία για κάθε κατηγορία ερωτήσεων, οι οποίες κατατάσσουν κάθε υποψήφια απάντηση ως ορθή ή λανθασμένη, επιστρέφοντας και ένα βαθμό βεβαιότητας για κάθε κατάταξη. Οι υποψήφιες απαντήσεις ταξινομούνται ως προς το πόσο βέβαιη ήταν η αντίστοιχη ΜΔΥ ότι αποτελούν ορθές απαντήσεις και επιλέγονται οι N απαντήσεις με τον υψηλότερο βαθμό βεβαιότητας, όπου N ο επιτρεπόμενος αριθμός απαντήσεων ανά ερώτηση. Σε κάποια πειράματα δοκιμάσαμε επίσης να χρησιμοποιήσουμε μόνο μία ΜΔΥ για όλες τις κατηγορίες ερωτήσεων, κάτι που επιτρέπει στη ΜΔΥ να εκπαιδευθεί σε περισσότερα δεδομένα εκπαίδευσης (τις υποψήφιες απαντήσεις όλων των ερωτήσεων εκπαίδευσης, αντί για τις υποψήφιες απαντήσεις μόνο μίας κατηγορίας ερωτήσεων).

Κάθε υποψήφια απάντηση παριστάνεται ως ένα διάνυσμα ιδιοτήτων (attributes) και οι ΜΔΥ μαθαίνουν να κατατάσσουν στην πραγματικότητα διανύσματα ιδιοτήτων, αντί για τις ίδιες τις υποψήφιες απαντήσεις. Μεγάλο μέρος της εργασίας αφιερώθηκε στην πειραματική εξεύρεση των καλύτερων ιδιοτήτων. Οι ιδιότητες πρέπει να παρέχουν επαρκείς πληροφορίες, ώστε να είναι δυνατή η ορθή κατάταξη των υποψηφίων απαντήσεων, αλλά δεν πρέπει να εισάγουν θόρυβο ή να οδηγούν σε υπερ-εφαρμογή (over-fitting) στα δεδομένα εκπαίδευσης. Οι ιδιότητες που χρησιμοποιήθηκαν, συνολικά 19, περιγράφονται αναλυτικά στις επόμενες ενότητες. Είναι οι ίδιες και για τις τρεις ΜΔΥ, αλλά κάθε ΜΔΥ μπορεί κατά την εκπαίδευσή της να μάθει να τους δίνει διαφορετικά βάρη. Στην περίπτωση που χρησιμοποιείται μόνο μία ΜΔΥ, υπάρχουν τρεις επιπλέον Boolean ιδιότητες, μία για κάθε κατηγορία ερώτησης. Ανάλογα με την κατηγορία της ερώτησης, η αντίστοιχη ιδιότητα παίρνει την τιμή true και οι υπόλοιπες δύο την τιμή false.

Σημειωτέον ότι κατά την εκπαίδευση της κάθε ΜΔΥ, τα ανακτηθέντα κείμενα κάθε ερώτησης επισημειώνονται χειρωνακτικά με ετικέτες της μορφής <ANSWER> απάντηση </ANSWER>, που δείχνουν ποιες υποψήφιες απαντήσεις είναι στην πραγματικότητα οι ορθές και επομένως ποιες είναι οι επιθυμητές αποφάσεις των ΜΔΥ για τα αντίστοιχα διανύσματα. Το απάντηση μπορεί να είναι μια λέξη ή μια ακολουθία λέξεων.

Στη συνέχεια, παρουσιάζονται οι 19 ιδιότητες, χωρισμένες σε 3 ομάδες.

3.5.1 Ιδιότητες που αφορούν την ίδια την υποψήφια απάντηση

➤ Οι πρώτες τέσσερις ιδιότητες αυτής της ομάδας σχετίζονται με τη βεβαιότητα του συστήματος αναγνώρισης ονομάτων οντοτήτων (ΣΑΟΟ). Υπενθυμίζεται ότι κάθε υποψήφια απάντηση είναι μία ή περισσότερες συνεχόμενες λέξεις που το ΣΑΟΟ θεώρησε όνομα προσώπου (αν η ερώτηση ζητά όνομα προσώπου), όνομα οργανισμού (αν η ερώτηση ζητά όνομα οργανισμού) ή χρονική έκφραση (αν η ερώτηση είναι χρονική). Κάθε λέξη της υποψήφιας απάντησης θα έχει επίσης σημειωθεί (μέσω των CONF0, CONF1, κλπ. της ενότητας 3.3) με έναν αριθμό που δείχνει το πόσο βέβαιο ήταν το ΣΑΟΟ για την απόφασή του, εκτός αν πρόκειται για χρονική έκφραση, οπότε το ΣΑΟΟ δεν επιστρέφει βαθμό βεβαιότητας και θεωρούμε ότι η βεβαιότητά του είναι 100% για όλες τις λέξεις της χρονικής έκφρασης. Οι ιδιότητες αυτές είναι:

- *Η μέση βεβαιότητα του ΣΑΟΟ ότι η υποψήφια απάντηση ανήκει στη σωστή κατηγορία.*

Αν η υποψήφια απάντηση αποτελείται από μια λέξη, τότε η μέση βεβαιότητα ταυτίζεται με τη βεβαιότητα του ΣΑΟΟ για τη λέξη αυτή. Αν η υποψήφια απάντηση αποτελείται από περισσότερες λέξεις, τότε αθροίζονται οι βεβαιότητες του ΣΑΟΟ για όλες τις λέξεις της υποψήφιας απάντησης και το άθροισμα διαιρείται με το πλήθος των λέξεων της υποψήφιας απάντησης.

- *Η τυπική απόκλιση της βεβαιότητας του ΣΑΟΟ ότι η υποψήφια απάντηση ανήκει στη ζητούμενη κατηγορία.*



Αν η υποψήφια απάντηση αποτελείται από μια λέξη, τότε η τυπική απόκλιση είναι 0. Αν η υποψήφια απάντηση αποτελείται από περισσότερες λέξεις, τότε ο υπολογισμός της τυπικής απόκλισης γίνεται από τον τύπο $\sqrt{\frac{\sum_{i=1}^n (p_i - \bar{p})^2}{n}}$, όπου n είναι το πλήθος των λέξεων της υποψήφιας απάντησης, p_i είναι η βεβαιότητα του ΣΑΟΟ για την i -οστή λέξη της υποψήφιας απάντησης και \bar{p} είναι η μέση τιμή της βεβαιότητας, δηλαδή η τιμή της προηγούμενης ιδιότητας.

- *Η μέγιστη βεβαιότητα του ΣΑΟΟ ότι η υποψήφια απάντηση ανήκει στη ζητούμενη κατηγορία.*

Αν η υποψήφια απάντηση αποτελείται από μια λέξη, τότε η μέγιστη βεβαιότητα ταυτίζεται με τη βεβαιότητά του για τη λέξη αυτή. Αν η υποψήφια απάντηση αποτελείται από περισσότερες λέξεις, τότε η μέγιστη βεβαιότητα είναι η μεγαλύτερη από τις βεβαιότητες που επέστρεψε το ΣΑΟΟ για τις λέξεις της υποψήφιας απάντησης.

- *Η ελάχιστη βεβαιότητα του ΣΑΟΟ ότι η υποψήφια απάντηση ανήκει στη ζητούμενη κατηγορία.*

Ορίζεται αντίστοιχα με την προηγούμενη περίπτωση.

Οι τέσσερις παραπάνω ιδιότητες έχουν αρχικά τιμές στο διάστημα [0, 1], αλλά κανονικοποιούνται, ώστε να πάρουν τιμές στο διάστημα [-1, 1].

- Οι υπόλοιπες δύο ιδιότητες της ομάδας σχετίζονται με το πόσες φορές εμφανίζεται η υποψήφια απάντηση μεταξύ των υποψηφίων απαντήσεων. Οι ιδιότητες αυτές είναι:
 - *To ποσοστό εμφανίσεων της υποψήφιας απάντησης μεταξύ των υποψηφίων απαντήσεων, όταν δεν εφαρμόζεται αποκοπή καταλήξεων (stemming) στις λέξεις των υποψηφίων απαντήσεων.*

Μεταξύ των υποψηφίων απαντήσεων μπορεί να περιλαμβάνονται πολλές εμφανίσεις της ίδιας υποψήφιας απάντησης, αν η υποψήφια απάντηση εμφανίζεται πολλές φορές στα ανακτηθέντα κείμενα της ενότητας 3.2. Η τιμή αυτής της ιδιότητας προκύπτει ως ο λόγος των εμφανίσεων της υποψήφιας απάντησης προς το συνολικό αριθμό εμφανίσεων των υποψηφίων απαντήσεων.

- *To ποσοστό εμφανίσεων της υποψήφιας απάντησης μεταξύ των υποψηφίων απαντήσεων, όταν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις των υποψηφίων απαντήσεων.*

Όπως η προηγούμενη ιδιότητα, αλλά αποκόπτονται οι καταλήξεις των λέξεων των υποψηφίων απαντήσεων.

Οι δύο παραπάνω ιδιότητες παίρνουν αρχικά τιμές στο διάστημα [0, 1], αλλά κανονικοποιούνται, ώστε να πάρουν τιμές στο διάστημα [-1, 1].

3.5.2 Ιδιότητες που αφορούν το κείμενο από το οποίο προέκυψε η υποψήφια απάντηση

- *H εφημερίδα από την οποία ανακτήθηκε το κείμενο.*

Η ιδιότητα αυτή παίρνει την τιμή -1, αν η εφημερίδα από την οποία ανακτήθηκε το άρθρο είναι «Τα Νέα», και την τιμή +1, αν η εφημερίδα, από την οποία ανακτήθηκε το άρθρο είναι «Το Βήμα».

- *H κατάταξη (ranking) των κειμένου από τη μηχανή αναζήτησης της εφημερίδας, μεταξύ των κειμένων από τα οποία εξάγονται οι υποψήφιες απαντήσεις.*

Υπενθυμίζεται ότι κρατάμε τα δέκα κορυφαία κείμενα ανά ερώτηση, μεταξύ εκείνων που επιστρέφει η κάθε μηχανή αναζήτησης. Υπάρχουν, όμως, ερωτήσεις για τις οποίες μια μηχανή αναζήτησης επιστρέφει λιγότερα από 10 άρθρα. Έστω n το πλήθος των άρθρων που επέστρεψε η μηχανή αναζήτησης της εφημερίδας από τα κείμενα της οποίας προέρχεται η υποψήφια απάντηση. Η τιμή αυτής της ιδιότητας δείχνει από ποιο από τα n κείμενα προέρχεται η υποψήφια απάντηση, αντιστοιχίζοντας το πρώτο κείμενο στην τιμή 1. Επομένως, η ιδιότητα παίρνει τιμές στο διάστημα $[1, n]$ και κανονικοποιείται, έτσι ώστε να πάρει τιμές στο διάστημα $[-1, 1]$.

- *H κατάταξη των κειμένου με βάση την παλαιότητά του.*

Για όλα τα κείμενα (το πολύ δέκα) τα οποία ανακτήθηκαν και επιλέχθηκαν από τη συγκεκριμένη εφημερίδα για τη συγκεκριμένη ερώτηση, βρέθηκε η ημερομηνία κατά την οποία γράφηκαν και υπολογίστηκε η χρονική της απόσταση από την τρέχουσα ημερομηνία. Με βάση τους υπολογισμούς αυτούς, τα κείμενα κατατάχτηκαν κατά αύξουσα σειρά παλαιότητας, δηλαδή το πρώτο κείμενο στην κατάταξη ήταν το πιο πρόσφατο και το τελευταίο κείμενο στην κατάταξη ήταν το πιο παλιό. Έστω n το πλήθος των καταταγμένων κειμένων. Όπως στην προηγούμενη ιδιότητα, η ελάχιστη τιμή του n είναι 1. Επομένως, η ιδιότητα παίρνει τιμές στο διάστημα $[1, n]$ και κανονικοποιείται, έτσι ώστε να πάρει τιμές στο διάστημα $[-1, 1]$.

3.5.3 Ιδιότητες που αφορούν τις υπόλοιπες λέξεις του κειμένου από το οποίο προέκυψε η υποψήφια απάντηση

- Οι πρώτες δύο ιδιότητες αυτής της ομάδας σχετίζονται με το πόσες λέξεις της ερώτησης δεν εμφανίζονται στο κείμενο από το οποίο προέρχεται η υποψήφια απάντηση. Οι ιδιότητες αυτές είναι:
 - *To ποσοστό των μη εμφανιζόμενων στο κείμενο λέξεων της ερώτησης, όταν δεν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.*

- *To ποσοστό των μη εμφανιζόμενων στο κείμενο λέξεων της ερώτησης, όταν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.*

Και οι δύο παραπάνω ιδιότητες, εφόσον είναι ποσοστά, παίρνουν τιμές στο διάστημα [0, 1] και κανονικοποιούνται, έτσι ώστε να πάρουν τιμές στο διάστημα [-1, 1].

- Οι επόμενες δύο ιδιότητες αυτής της ομάδας σχετίζονται με το πόσες λέξεις (ακριβέστερα, εμφανίσεις λέξεων) του κειμένου αποτελούν λέξεις της ερώτησης. Οι ιδιότητες αυτές είναι:
 - *To ποσοστό των λέξεων (ακριβέστερα, εμφανίσεων λέξεων) του κειμένου που αποτελούν λέξεις της ερώτησης, όταν δεν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.*
 - *To ποσοστό των λέξεων του κειμένου, οι οποίες αποτελούν λέξεις της ερώτησης, όταν εφαρμόζεται stemming στις λέξεις του κειμένου και στις λέξεις της ερώτησης.*

Και οι δύο παραπάνω ιδιότητες, εφόσον είναι ποσοστά, παίρνουν τιμές στο διάστημα [0, 1] και κανονικοποιούνται, έτσι ώστε να πάρουν τιμές στο διάστημα [-1, 1].

- Ορίζουμε ως λέξεις-στόχους του κειμένου από το οποίο προέρχεται η υποψήφια απάντηση τις λέξεις (ακριβέστερα, τις εμφανίσεις λέξεων) του κειμένου που περιέχονται και στην ερώτηση. Οι επόμενες δύο ιδιότητες αυτής της ομάδας σχετίζονται με τη μέση απόσταση από την υποψήφια απάντηση των λέξεων-στόχων του κειμένου που δεν περιλαμβάνονται στην υποψήφια απάντηση. Ο υπολογισμός της μέσης απόστασης γίνεται από τον τύπο $\frac{\sum_{i=1}^n \frac{1}{d_i}}{n}$, όπου n είναι το πλήθος των λέξεων-στόχων του κειμένου (ακριβέστερα, των εμφανίσεών τους) και d_i είναι η απόσταση (μετρούμενη σε λέξεις) της i -οστής λέξης-στόχου από την υποψήφια απάντηση (από το κοντινότερο άκρο της).
- *H μέση απόσταση από την υποψήφια απάντηση των λέξεων-στόχων του κειμένου, όταν δεν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και στις λέξεις της ερώτησης.*

Στην περίπτωση αυτή, μια λέξη του κειμένου θεωρείται λέξη-στόχος, αν ταυτίζεται με κάποια λέξη της ερώτησης.

- *H μέση απόσταση από την υποψήφια απάντηση των λέξεων-στόχων του κειμένου, όταν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και στις λέξεις της ερώτησης.*

Στην περίπτωση αυτή, μια λέξη του κειμένου θεωρείται λέξη-στόχος, αν έχει την ίδια ρίζα με κάποια λέξη της ερώτησης.

Αφού η ελάχιστη και η μέγιστη τιμή τις οποίες μπορεί να πάρει η ποσότητα $\sum_{i=1}^n \frac{1}{d_i}$ είναι, αντίστοιχα, 0 και n , τότε η ελάχιστη και η

μέγιστη τιμή τις οποίες μπορεί να πάρει η ποσότητα $\frac{\sum_{i=1}^n \frac{1}{d_i}}{n}$ θα είναι, αντίστοιχα, 0 και 1. Επομένως, και οι δύο παραπάνω ιδιότητες παίρνουν τιμές στο διάστημα $[0, 1]$ και κανονικοποιούνται, έτσι ώστε να πάρουν τιμές στο διάστημα $[-1, 1]$

- Οι επόμενες δύο ιδιότητες σχετίζονται με το μέσο πλήθος λέξεων που υπάρχουν μεταξύ δύο λέξεων-στόχων του κειμένου. Μεταξύ των 2 λέξεων-στόχων, οι οποίες εξετάζονται κάθε φορά, είναι πιθανό να μεσολαβούν και άλλες λέξεις-στόχοι. Για κάθε λέξη-στόχο του κειμένου υπολογίζουμε την απόστασή της από όλες τις επόμενές της λέξεις-στόχους. Ο υπολογισμός του μέσου πλήθους λέξεων μεταξύ δύο λέξεων-στόχων γίνεται από τον τύπο $\frac{\sum_{i=1}^n \sum_{j=i+1}^n d_{ij}}{N}$, όπου n είναι το πλήθος των λέξεων-στόχων του κειμένου, d_{ij} είναι η απόσταση (μετρούμενη σε λέξεις) μεταξύ της i-οστής και της j-οστής λέξης-στόχου και N είναι το πλήθος των διαφορετικών ζευγών λέξεων-στόχων. Οι ιδιότητες αυτές είναι:
 - Το μέσο πλήθος λέξεων μεταξύ δύο λέξεων-στόχων του κειμένου, όταν δεν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.
 - Το μέσο πλήθος λέξεων μεταξύ δύο λέξεων-στόχων του κειμένου, όταν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.

Κατά τον υπολογισμό των αποστάσεων μεταξύ ζευγών λέξεων-στόχων, βρίσκουμε την ελάχιστη και τη μέγιστη απόσταση μεταξύ δύο λέξεων-στόχων σε αυτό το κείμενο, έστω minDistance και maxDistance αντίστοιχα. Οι δύο παραπάνω ιδιότητες παίρνουν τιμές στο διάστημα $[minDistance, maxDistance]$ και κανονικοποιούνται, έτσι ώστε το προηγούμενο διάστημα να μετασχηματιστεί στο διάστημα $[-1, 1]$.

- Οι τελευταίες δύο ιδιότητες αυτής της ομάδας σχετίζονται με το μέσο μήκος ακολουθιών λέξεων-στόχων, οι οποίες εμφανίζονται συνεχόμενες τόσο στο κείμενο όσο και στην ερώτηση. Οι ιδιότητες αυτές είναι:
 - Το μέσο μήκος των ακολουθιών λέξεων-στόχων οι οποίες εμφανίζονται συνεχόμενες τόσο στο κείμενο όσο και στην ερώτηση, όταν δεν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.

Βρίσκουμε όλες τις (μακρύτερες δυνατές) ακολουθίες συνεχόμενων λέξεων-στόχων του κειμένου, οι οποίες αποτελούν και ακολουθίες συνεχόμενων λέξεων της ερώτησης. Στη συνέχεια, αθροίζουμε τα μήκη των ακολουθιών αυτών (μετρούμενων σε λέξεις) και διαιρούμε με το πλήθος των ακολουθιών.

- Το μέσο μήκος των ακολουθιών λέξεων-στόχων οι οποίες εμφανίζονται συνεχόμενες τόσο στο κείμενο όσο και στην ερώτηση, όταν εφαρμόζεται αποκοπή καταλήξεων στις λέξεις του κειμένου και της ερώτησης.

Βρίσκουμε όλες τις (μακρύτερες δυνατές) ακολουθίες συνεχόμενων λέξεων-στόχων του κειμένου, οι λέξεις των οποίων έχουν τις ίδιες ρίζες με τις λέξεις κάποιας ακολουθίας λέξεων της ερώτησης. Στη συνέχεια, αθροίζουμε τα μήκη των ακολουθιών αυτών και διαιρούμε με το πλήθος των ακολουθιών.

Κατά τον υπολογισμό των τιμών των παραπάνω δύο ιδιοτήτων, βρίσκουμε το ελάχιστο και το μέγιστο μήκος των ακολουθιών, έστω \min_{Num} και \max_{Num} αντίστοιχα. Οι δύο ιδιότητες παίρνουν τιμές στο διάστημα $[\min_{\text{Num}}, \max_{\text{Num}}]$ και κανονικοποιούνται, έτσι ώστε το προηγούμενο διάστημα να μετασχηματιστεί στο $[-1, 1]$.

3.5.4 Ιδιότητες που αφορούν την κατηγορία της ερώτησης

Στην περίπτωση που χρησιμοποιείται μόνο μία ΜΔΥ, υπάρχουν τρεις επιπλέον Boolean ιδιότητες, μία για κάθε κατηγορία ερώτησης. Κάθε ιδιότητα παίρνει την τιμή 1 (παριστάνει το *true*) αν η ερώτηση ανήκει στην αντίστοιχη κατηγορία και την τιμή -1 (*false*) διαφορετικά

3.6 ΕΠΙΛΟΓΗ ΥΠΟΨΗΦΙΑΣ ΑΠΑΝΤΗΣΗΣ Η ΑΠΑΝΤΗΣΕΩΝ

Στο τέλος του προηγούμενου σταδίου, σε κάθε υποψήφια απάντηση έχει δοθεί ένας βαθμός βεβαιότητας, που δείχνει το κατά πόσον το σύστημα «πιστεύει» ότι η υποψήφια απάντηση είναι ορθή. Μία υποψήφια απάντηση, όμως, μπορεί να έχει παραχθεί πολλές φορές, αν εμφανίζοταν πολλές φορές στα ανακτηθέντα κείμενα, και σε κάθε εμφάνισή της θα έχει δοθεί ένας εν γένει διαφορετικός βαθμός βεβαιότητας. Υπάρχουν τρεις διαφορετικοί τρόποι με τους οποίους το σύστημα επιλέγει την τελική του απάντηση, ανάλογα με το πώς θα λάβει (αν λάβει) υπόψη του τις πολλαπλές εμφανίσεις των υποψηφίων απαντήσεων:

- **SVM-NG (No Grouping):** Οι πολλαπλές εμφανίσεις της ίδιας υποψήφιας απάντησης δεν ομαδοποιούνται. Το σύστημα επιστρέφει απλά την υποψήφια απάντηση με το μεγαλύτερο βαθμό βεβαιότητας.
- **SVM-G-HM (Grouping – Highest Mean):** Οι πολλαπλές εμφανίσεις της ίδιας υποψήφιας απάντησης ομαδοποιούνται, δηλαδή θεωρούνται μόνο μία υποψήφια απάντηση. Σε κάθε ομάδα αντιστοιχίζεται ο μέσος όρος των βαθμών βεβαιότητας των εμφανίσεων που της ανήκουν. Κατά τα άλλα, το σύστημα επιστρέφει πάλι την υποψήφια απάντηση με το μέγιστο βαθμό βεβαιότητας.
- **SVM-G-HS (Grouping – Highest Sum):** Όπως η προηγούμενη μέθοδος, αλλά σε κάθε ομάδα αντιστοιχίζεται το άθροισμα των βαθμών βεβαιότητας των εμφανίσεων που της ανήκουν.

Στη διάρκεια των πειραμάτων, συγκρίναμε και με τις εξής απλοϊκές (baseline) μεθόδους επιλογής απάντησης:

- **R (Random):** Το σύστημα επιλέγει μια τυχαία απάντηση από το σύνολο των υποψηφίων απαντήσεων. Δεν ομαδοποιούνται οι πολλαπλές εμφανίσεις της ίδιας υποψήφιας απάντησης.
- **HF-NG (Highest Frequency – No Grouping):** Το σύστημα επιλέγει την υποψήφια απάντηση με τις περισσότερες εμφανίσεις μεταξύ των υποψηφίων απαντήσεων.
- **HF-G (Highest Frequency – Grouping):** Το σύστημα επιλέγει την υποψήφια απάντηση με τις περισσότερες εμφανίσεις μεταξύ των υποψηφίων απαντήσεων. Εφαρμόζονται οι κανόνες ομαδοποίησης των υποψηφίων απαντήσεων, οπότε εμφανίσεις πολύ παρόμοιων υποψηφίων απαντήσεων θεωρούνται εμφανίσεις της ίδιας υποψήφιας απάντησης.

Η αξιολόγηση του συστήματος γίνεται ως προς το ποσοστό των ερωτήσεων που απάντησε σωστά. Στην περίπτωση που επιτρέπεται να επιστραφούν πέντε ερωτήσεις ανά ερώτηση, θεωρούμε ότι το σύστημα απάντησε σωστά μια ερώτηση αν τουλάχιστον μία από τις πέντε απαντήσεις του ήταν σωστή.

4 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Στη διάρκεια των πειραμάτων, χρησιμοποιήθηκε μια συλλογή 370 ερωτήσεων. Η συλλογή δημιουργήθηκε στη διάρκεια αυτής της εργασίας χρησιμοποιώντας ως αφετηρία τις ερωτήσεις της εργασίας [8]. Ο παρακάτω πίνακας δείχνει πόσες ερωτήσεις περιλαμβάνει η συλλογή ανά κατηγορία.

Κατηγορία ερωτήσεων	Πλήθος ερωτήσεων
Ερωτήσεις προσώπων	132
Ερωτήσεις οργανισμών	93
Ερωτήσεις χρόνου	145
Όλες οι ερωτήσεις	370

Για κάθε μία ερώτηση, ανακτήθηκαν και αποθηκεύθηκαν το πολύ δέκα κείμενα από κάθε μία από τις δύο εφημερίδες («Τα Νέα» και «Το Βήμα»), όπως περιγράφηκε στο κεφάλαιο 3. Προέκυψε έτσι μια συλλογή συνολικά 3832 κειμένων. Ο παρακάτω πίνακας δείχνει το μέσο αριθμό κειμένων ανά ερώτηση κάθε μίας από τις τρεις κατηγορίες.

Κατηγορία ερωτήσεων	Μέσος αριθμός κειμένων ανά ερώτηση
Ερωτήσεις προσώπων	13
Ερωτήσεις οργανισμών	10
Ερωτήσεις χρόνου	8
Όλες οι ερωτήσεις	10

Σε κάθε κείμενο της συλλογής, εντοπίστηκαν οι υποψήφιες απαντήσεις, όπως περιγράφηκε στο κεφάλαιο 3. Περισσότερες πληροφορίες για τον ακριβή τρόπο εντοπισμού των υποψηφίων απαντήσεων δίνονται στην ενότητα 4.3. Προέκυψε έτσι μια συλλογή συνολικά 58276 υποψηφίων απαντήσεων. Ο παρακάτω πίνακας δείχνει το μέσο αριθμό των ορθών και λανθασμένων υποψηφίων απαντήσεων ανά ερώτηση κάθε μίας από τις τρεις κατηγορίες. Παρατηρούμε ότι οι λανθασμένες υποψήφιες απαντήσεις είναι πολύ περισσότερες από τις ορθές, έχουμε δηλαδή ένα πρόβλημα διαχωρισμού με ανομοιοβαρείς (imbalanced) κατηγορίες. Αυτό αποτελεί πρόβλημα για τις ΜΔΥ και τους περισσότερους αλγορίθμους επιβλεπόμενης μάθησης, γιατί ενέχει τον κίνδυνο ο ταξινομητής να μάθει να κατατάσσει όλες τις περιπτώσεις στην πιο συχνή κατηγορία. Για το λόγο αυτό δοκιμάσαμε να δίνουμε στις ορθές υποψήφιες απαντήσεις, που είναι η λιγότερο συχνή κατηγορία, μεγαλύτερα βάρη απ' ό,τι στις λανθασμένες. Η προσέγγιση αυτή, όμως, δε βελτίωσε τα αποτελέσματα.

<u>Κατηγορία ερωτήσεων</u>	<u>Μέσος αριθμός ορθών υποψηφίων απαντήσεων ανά ερώτηση</u>	<u>Μέσος αριθμός λανθασμένων υποψηφίων απαντήσεων ανά ερώτηση</u>
<i>Ερωτήσεις προσώπων</i>	18	217
<i>Ερωτήσεις οργανισμών</i>	16	139
<i>Ερωτήσεις χρόνου</i>	3	84
<i>Όλες οι ερωτήσεις</i>	12	146

4.2 ΔΙΑΣΤΑΥΡΩΜΕΝΗ ΕΠΙΚΥΡΩΣΗ

Με την τεχνική της διασταυρωμένης επικύρωσης (cross-validation), τα δεδομένα, στην περίπτωσή μας οι ερωτήσεις της συλλογής και τα αντίστοιχα κείμενα και υποψήφιες απαντήσεις, χωρίζονται σε n ίσα μέρη και τα πειράματα επαναλαμβάνονται n φορές. Σε κάθε επανάληψη χρησιμοποιείται ένα διαφορετικό μέρος των δεδομένων για την αξιολόγηση του συστήματος, ενώ τα υπόλοιπα n μέρη χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Τα τελικά αποτελέσματα είναι ο μέσος όρος των αποτελεσμάτων από τις n επαναλήψεις. Με τη μέθοδο αυτή χρησιμοποιούνται τελικά όλα τα δεδομένα τόσο για εκπαίδευση όσο και για αξιολόγηση, ενώ δεν υπάρχει περίπτωση το σύστημα να αξιολογηθεί σε δεδομένα τα οποία έχουν χρησιμοποιηθεί ταυτόχρονα για την εκπαίδευσή του. Μια συνηθισμένη τιμή του n είναι 10 (10-fold cross-validation).

Λόγω του σχετικά μικρού μεγέθους των δεδομένων, χρησιμοποιήσαμε μια παραλλαγή της διασταυρωμένης επικύρωσης, γνωστή ως leave-one-out cross-validation, όπου τα δεδομένα χωρίζονται σε τόσα μέρη όσα και οι περιπτώσεις που περιέχουν, στην περίπτωσή μας, τόσα μέρη όσα είναι και οι ερωτήσεις. Σε κάθε επανάληψη, μία ερώτηση και τα αντίστοιχα κείμενα και υποψήφιες απαντήσεις χρησιμοποιούνται ως δεδομένα αξιολόγησης, ενώ οι υπόλοιπες ερωτήσεις, κείμενα και υποψήφιες απαντήσεις χρησιμοποιούνται ως δεδομένα εκπαίδευσης.

Στην περίπτωση όπου χρησιμοποιούνταν τρεις ΜΔΥ, μία για κάθε κατηγορία ερωτήσεων, αξιολογήσαμε ξεχωριστά κάθε μία ΜΔΥ, εκτελώντας μια διασταυρωμένη επικύρωση της παραπάνω μορφής στις ερωτήσεις της αντίστοιχης κατηγορίας, τα αντίστοιχα κείμενα και τις αντίστοιχες υποψήφιες απαντήσεις. Στην περίπτωση όπου υπήρχε μόνο μία ΜΔΥ για τις ερωτήσεις και των τριών κατηγοριών, επειδή για κάθε κατηγορία ερωτήσεων είχαμε συλλέξει διαφορετικό πλήθος ερωτήσεων, επιλέξαμε 90 ερωτήσεις από κάθε κατηγορία, έτσι ώστε να έχουμε συνολικά 270 ερωτήσεις και να δοθεί ίση βαρύτητα στις ερωτήσεις των τριών κατηγοριών.

4.3 ΠΕΙΡΑΜΑΤΑ

Τα πειράματα χωρίζονται σε δύο ομάδες. Η πρώτη ομάδα περιέχει τα πειράματα που έγιναν στην περίπτωση όπου χρησιμοποιείται μία διαφορετική

ΜΔΥ για κάθε κατηγορία ερωτήσεων. Η δεύτερη ομάδα περιέχει τα πειράματα που έγιναν στην περίπτωση όπου χρησιμοποιήθηκε μία μόνο ΜΔΥ και για τις τρεις κατηγορίες ερωτήσεων.

4.3.1 Πειράματα με τρεις ΜΔΥ, μία για κάθε κατηγορία ερωτήσεων

Σε αυτή την ομάδα πειραμάτων χρησιμοποιήθηκε μία διαφορετική ΜΔΥ για κάθε μία από τις τρεις κατηγορίες ερωτήσεων. Η κάθε ΜΔΥ αξιολογήθηκε σε ξεχωριστά πειράματα.

4.3.1.1 Πειράματα για ερωτήσεις προσώπων

Η συλλογή μας περιείχε 132 ερωτήσεις προσώπων. Από αυτές, οι 22 χρησιμοποιήθηκαν για τη ρύθμιση των παραμέτρων (parameter tuning) της ΜΔΥ. Οι υπόλοιπες 110 ερωτήσεις χρησιμοποιήθηκαν για τη διεξαγωγή των πειραμάτων με leave-one-out cross-validation.

Οι υποψήφιες απαντήσεις για κάθε ερώτηση προέκυψαν από τα κείμενα που είχαν συλλεχθεί για την ερώτηση αυτή και ήταν όλες οι ακολουθίες λέξεων που είχαν χαρακτηριστεί από το σύστημα αναγνώρισης ονομάτων οντοτήτων ως ονόματα προσώπων με μέγιστο (μεταξύ όλων των λέξεων του ονόματος) βαθμό βεβαιότητας μεγαλύτερο ή ίσο του 80%. Από αυτές εξαιρούνταν οι εξής:

- Οι υποψήφιες απαντήσεις οι οποίες είχαν μήκος (μετρούμενο σε χαρακτήρες) μικρότερο του 3.
- Οι υποψήφιες απαντήσεις οι οποίες είχαν μήκος (μετρούμενο σε χαρακτήρες) ίσο με 3 και τελείωναν σε «.».
- Οι λέξεις «και» και «κι». Συγκεκριμένα, αν μια υποψήφια απάντηση περιλαμβανε κάποια από αυτές τις λέξεις (π.χ. αν το «Καραμανλής και Παπανδρέου» είχε λανθασμένα θεωρηθεί ένα όνομα), τότε διαχωριζόταν σε δύο υποψήφιες απαντήσεις, οι οποίες ήταν το τμήμα της υποψήφιας απάντησης πριν από το «και» και το τμήμα της υποψήφιας απάντησης μετά από το «και».

Η ομαδοποίηση των υποψηφίων απαντήσεων έγινε ως εξής:

- Μια υποψήφια απάντηση που αποτελείται από μία μόνο λέξη θεωρείται όμοια με μια υποψήφια απάντηση που αποτελείται από περισσότερες λέξεις, στην περίπτωση που κάποια από τις λέξεις της δεύτερης υποψήφιας απάντησης είναι η ίδια με την πρώτη υποψήφια απάντηση. Για παράδειγμα, η υποψήφια απάντηση «Καραμανλής» ομαδοποιείται με τις υποψήφιες απαντήσεις «Κ. Καραμανλή» και «Κώστας Καραμανλής» και η υποψήφια απάντηση «Κώστας» ομαδοποιείται με την υποψήφια απάντηση «Κώστα Καραμανλή».
- Δύο υποψήφιες απαντήσεις που αποτελούνται από το ίδιο πλήθος λέξεων θεωρούνται όμοιες αν η πρώτη λέξη της μιας ξεκινά με την πρώτη λέξη της άλλης (αγνοώντας την πιθανή ύπαρξη τελείας) και όλες οι υπόλοιπες λέξεις τους είναι οι ίδιες. Για παράδειγμα, η υποψήφια απάντηση «Κ. Καραμανλή» ομαδοποιείται με τις υποψήφιες απαντήσεις «Κώστας Καραμανλής» και «Κων/νου Καραμανλή».

- Μια υποψήφια απάντηση θεωρείται όμοια με μια υποψήφια απάντηση με μεγαλύτερο πλήθος λέξεων, στην περίπτωση που η τελευταία λέξη της πρώτης υποψήφιας απάντησης είναι η ίδια με κάποια λέξη της δεύτερης υποψήφιας απάντησης και όλες οι υπόλοιπες λέξεις της πρώτης υποψήφιας απάντησης είτε είναι οι ίδιες με κάποιες λέξεις της δεύτερης υποψήφιας απάντησης ή κάποιες λέξεις της δεύτερης υποψήφιας απάντησης ξεκινούν με αυτές (αγνοώντας την πιθανή ύπαρξη τελείας). Για παράδειγμα, η υποψήφια απάντηση «Άννα Ψαρούδα Μπενάκη» ομαδοποιείται με τις υποψήφιες απαντήσεις «Άννας Μπενάκη» και «Α. Μπενάκη».
- Για τη σύγκριση των λέξεων χρησιμοποιείται σε όλες τις περιπτώσεις αποκοπή καταλήξεων και μετατροπή όλων των χαρακτήρων των συγκρινόμενων λέξεων στην ίδια γλώσσα (ελληνικά ή αγγλικά). Συγκεκριμένα, σε κάθε υποψήφια απάντηση, τα γράμματα του ελληνικού αλφαριθμητού τα οποία είναι τα ίδια με κάποια αγγλικά γράμματα μετατρέπονται σε αυτά τα αγγλικά γράμματα.

Τα ποσοστά επιτυχίας (ποσοστά ερωτήσεων που απαντήθηκαν σωστά) αυτού του πειράματος παρουσιάζονται στον παρακάτω πίνακα.

Πλήθος απαντήσεων	1 απάντηση	5 απαντήσεις
Μέτρα επίδοσης		
R	0.11	0.32
HF-NG	0.27	0.58
HF-G	0.24	0.51
SVM-NG	0.23	0.44
SVM-G-HM	0.27	0.57
SVM-G-HS	0.32	0.65

Παρατηρούμε ότι τα καλύτερα αποτελέσματα επιτυγχάνονται με τη χρήση της ΜΔΥ σε συνδυασμό με την ομαδοποίηση των υποψηφίων απαντήσεων και την απεικόνιση κάθε ομάδας στο άθροισμα των βαθμών βεβαιότητας των μελών της (μέθοδος SVM-G-HS). Η απλούστερη μέθοδος που επιλέγει απλά τη συχνότερη υποψήφια απάντηση αφού εφαρμόσει ομαδοποίηση στις υποψήφιες απαντήσεις (μέθοδος HF-G) επιτυγχάνει αισθητά χειρότερα αποτελέσματα. Παρατηρούμε, επίσης, ότι τα ποσοστά επιτυχίας για τη μέθοδο HF-NG είναι υψηλότερα απ' ό.τι για την HF-G, ενώ για τις μεθόδους SVM-NG και SVM-G ισχύει το αντίστροφο, δηλαδή η ομαδοποίηση ωφελεί μόνο σε συνδυασμό με τη χρήση της ΜΔΥ. Αξιοσημείωτο είναι ότι η μέθοδος που επιλέγει απλά τη συχνότερη υποψήφια απάντηση χωρίς ομαδοποίηση (μέθοδος HF-NG) επιτυγχάνει αρκετά υψηλά αποτελέσματα. Οπως ήταν αναμενόμενο, τα ποσοστά επιτυχίας είναι υψηλότερα όταν επιτρέπονται πέντε απαντήσεις ανά ερώτηση.

4.3.1.2 Πειράματα για ερωτήσεις οργανισμών

Η συλλογή μας περιείχε 93 ερωτήσεις οργανισμών. Από αυτές οι 13 χρησιμοποιήθηκαν για τη ρύθμιση των παραμέτρων της ΜΔΥ, ενώ οι

υπόλοιπες 80 χρησιμοποιήθηκαν για τη διεξαγωγή των πειραμάτων με leave-one-out cross-validation.

Οι υποψήφιες απαντήσεις για κάθε ερώτηση προέκυψαν από τα κείμενα που είχαν συλλεχθεί για την ερώτηση και ήταν όλες οι ακολουθίες λέξεων που είχαν χαρακτηριστεί από το σύστημα αναγνώρισης ονομάτων οντοτήτων ως ονόματα οργανισμών με μέγιστο (μεταξύ των λέξεων του ονόματος) βαθμό βεβαιότητας μεγαλύτερο ή ίσο του 80%.

Η ομαδοποίηση των υποψηφίων απαντήσεων έγινε ως εξής:

- Δυο υποψήφιες απαντήσεις που αποτελούνται από το ίδιο πλήθος λέξεων θεωρούνται όμοιες, στην περίπτωση που οι αντίστοιχες λέξεις τους είναι οι ίδιες.
- Μια υποψήφια απάντηση θεωρείται όμοια με μια υποψήφια απάντηση που αποτελείται από περισσότερες λέξεις, στην περίπτωση που όλες οι λέξεις της πρώτης υποψήφιας απάντησης περιέχονται στη δεύτερη.
- Μια υποψήφια απάντηση που αποτελείται από τα αρχικά κάποιου οργανισμού θεωρείται όμοια με μια άλλη υποψήφια απάντηση, στην περίπτωση που τα πρώτα γράμματα των λέξεων της δεύτερης υποψήφιας απάντησης είναι τα αρχικά της πρώτης υποψήφιας απάντησης. Για παράδειγμα, η υποψήφια απάντηση «Κομμουνιστικό Κόμμα Ελλάδας» ομαδοποιείται με τις υποψήφιες απαντήσεις «ΚΚΕ» και «Κ.Κ.Ε.» και η υποψήφια απάντηση «Λαϊκός Ορθόδοξος Συναγερμός» ομαδοποιείται με τις υποψήφιες απαντήσεις «ΛΑΟΣ» και «ΛΑ.Ο.Σ.».
- Για τη σύγκριση των λέξεων χρησιμοποιείται σε όλες τις περιπτώσεις αποκοπή καταλήξεων και μετατροπή όλων των χαρακτήρων των συγκρινόμενων λέξεων στην ίδια γλώσσα (ελληνικά ή αγγλικά), όπως στην περίπτωση των ερωτήσεων προσώπων.

Τα ποσοστά επιτυχίας (ποσοστά ερωτήσεων που απαντήθηκαν σωστά) αυτού του πειράματος παρουσιάζονται στον παρακάτω πίνακα.

Πλήθος απαντήσεων	<i>1 απάντηση</i>	<i>5 απαντήσεις</i>
Μέτρα επίδοσης		
<i>R</i>	0.11	0.36
<i>HF-NG</i>	0.26	0.53
<i>HF-G</i>	0.28	0.45
<i>SVM-NG</i>	0.08	0.26
<i>SVM-G-HM</i>	0.13	0.41
<i>SVM-G-HS</i>	0.31	0.58

Παρατηρούμε ότι τα καλύτερα αποτελέσματα επιτυγχάνονται με τη μέθοδο SVM-G-HS, όπως και στην περίπτωση των ερωτήσεων προσώπων, ενώ αρκετά υψηλά αποτελέσματα επιτυγχάνονται και με τη μέθοδο HF-NG στην περίπτωση των πέντε απαντήσεων και με τη μέθοδο HF-G στην περίπτωση της μίας απάντησης. Και πάλι η ομαδοποίηση ωφελεί μόνο σε συνδυασμό με τη χρήση της ΜΔΥ.

4.3.1.3 Πειράματα για χρονικές ερωτήσεις

Η συλλογή μας περιείχε 145 χρονικές ερωτήσεις. Από αυτές, οι 25 χρησιμοποιήθηκαν για τη ρύθμιση των παραμέτρων της ΜΔΥ, ενώ οι υπόλοιπες 120 χρησιμοποιήθηκαν για τη διεξαγωγή των πειραμάτων με leave-one-out cross-validation.

Οι υποψήφιες απαντήσεις για κάθε ερώτηση προέκυψαν από τα κείμενα που είχαν συλλεχθεί για την ερώτηση και ήταν όλες οι ακολουθίες λέξεων που είχαν χαρακτηριστεί από το σύστημα αναγνώρισης ονομάτων οντοτήτων ως χρονικές εκφράσεις. Από αυτές εξαιρούνταν οι εξής:

- Οι υποψήφιες απαντήσεις οι οποίες ήταν ημέρες της εβδομάδας (χωρίς συνοδευτική ημερομηνία), π.χ. «Δευτέρα», «Παρασκευής» κ.ο.κ.
- Οι υποψήφιες απαντήσεις οι οποίες ήταν ώρες (χωρίς συνοδευτική ημερομηνία), π.χ. «23:00», «10.00» κ.ο.κ.
- Οι υποψήφιες απαντήσεις οι οποίες ήταν μελλοντικές ημερομηνίες.

Η ομαδοποίηση των υποψηφίων απαντήσεων έγινε ως εξής:

- Δυο υποψήφιες απαντήσεις που αποτελούνται από το ίδιο πλήθος λέξεων θεωρούνται όμοιες, στην περίπτωση που οι αντίστοιχες λέξεις τους είναι οι ίδιες.
- Μια υποψήφια απάντηση που αποτελείται μόνο από έτος θεωρείται όμοια με μια υποψήφια απάντηση που αποτελείται από μήνα και έτος ή από ημερομηνία, στην περίπτωση που τα δύο έτη είναι τα ίδια. Για παράδειγμα, η υποψήφια απάντηση «2000» ομαδοποιείται με τις υποψήφιες απαντήσεις «Ιούνιο 2000» και «25 Οκτωβρίου 2000».
- Μια υποψήφια απάντηση που αποτελείται από μήνα και έτος θεωρείται όμοια με μια υποψήφια απάντηση που αποτελείται από ημερομηνία, στην περίπτωση που οι μήνες και τα έτη των δύο υποψηφίων απαντήσεων είναι τα ίδια. Για παράδειγμα, η υποψήφια απάντηση «Ιούλιος 2005» ομαδοποιείται με τις υποψήφιες απαντήσεις «10 Ιουλίου 2005» και «25 Ιουλίου 2005».
- Για τη σύγκριση των λέξεων χρησιμοποιείται σε όλες τις περιπτώσεις αποκοπή καταλήξεων και μετατροπή όλων των χαρακτήρων των συγκρινόμενων λέξεων στην ίδια γλώσσα (ελληνικά ή αγγλικά), όπως στις προηγούμενες κατηγορίες ερωτήσεων.

Τα ποσοστά επιτυχίας αυτού του πειράματος παρουσιάζονται στον παρακάτω πίνακα.

Πλήθος απαντήσεων	1 απάντηση	5 απαντήσεις
Μέτρα επίδοσης		
R	0.10	0.28
HF-NG	0.16	0.38
HF-G	0.18	0.43
SVM-NG	0.21	0.43
SVM-G-HM	0.23	0.49
SVM-G-HS	0.24	0.49

Τα καλύτερα αποτελέσματα επιτυγχάνονται πάλι με τη χρήση της ΜΔΥ σε συνδυασμό με την ομαδοποίηση των υποψηφίων απαντήσεων και τη απεικόνιση κάθε ομάδας στο άθροισμα των βαθμών βεβαιότητας των μελών της (μέθοδος SVM-G-HS). Σε αυτή την περίπτωση, η ομαδοποίηση φαίνεται να ωφελεί και όταν επιλέγεται απλά η συχνότερη υποψήφια απάντηση. Τα ποσοστά επιτυχίας είναι πάλι υψηλότερα όταν επιτρέπονται πέντε απαντήσεις ανά ερώτηση.

4.3.1.4 Παρατηρήσεις για τα αποτελέσματα των πειραμάτων

Για την κατηγορία των ερωτήσεων προσώπων, μια ερώτηση μπορεί τις περισσότερες φορές να αναχθεί στη μορφή ερωτηματική έκφραση – ρήμα – κατηγορούμενο ή αντικείμενο (π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδας;» ή «Ποιος ανακάλυψε την πενικιλίνη;»), οπότε η απάντηση στην ερώτηση αυτή είναι ένα υποκείμενο (π.χ. «Ο Καραμανλής είναι ο πρωθυπουργός της Ελλάδας»). Μια που οι μηχανές αναζήτησης των δύο εφημερίδων επιστρέφουν κείμενα που περιέχουν κατά κανόνα τις λέξεις-κλειδιά της ερώτησης, τα επιστρεφόμενα κείμενα περιέχουν κατά κανόνα το κατηγορούμενο ή το αντικείμενο της ερώτησης. Γενικά είναι σπάνιο μια πρόταση να έχει κατηγορούμενο ή αντικείμενο χωρίς υποκείμενο. Έτσι, σε όλα σχεδόν τα επιστρεφόμενα κείμενα εμφανίζοταν, μεταξύ άλλων, και το ζητούμενο υποκείμενο. Το ποσοστό επιτυχίας του συστήματος σε αυτή την κατηγορία ερωτήσεων είναι μεγαλύτερο από ό,τι στις άλλες κατηγορίες, γιατί το σύστημα έπρεπε απλά να ξεχωρίσει ποια από τα υποκείμενα των διαφόρων προτάσεων κάθε κειμένου αντιστοιχούσαν στο ζητούμενο υποκείμενο της ερώτησης. Το ποσοστό επιτυχίας, μάλιστα, θα ήταν ακόμα υψηλότερο αν δεν απαιτούσαμε το σύστημα να βρει την απάντηση που ισχύει κατά τη χρονική στιγμή της ερώτησης. Για παράδειγμα, το χρόνο που γράφεται αυτό το κείμενο, αν η ερώτηση είναι «Ποιος είναι ο πρωθυπουργός της Ελλάδας;», θεωρούμε λανθασμένη την απάντηση «Σημίτης» και σωστή την απάντηση «Καραμανλής».

Για την κατηγορία των ερωτήσεων οργανισμών, μια ερώτηση μπορεί τις περισσότερες φορές να αναχθεί στη μορφή ερωτηματική έκφραση – ρήμα – υποκείμενο ή αντικείμενο (π.χ. «Τι ίδρυσαν οι Τσακάλωφ, Σκουφάς και Ξάνθος;» ή «Σε ποιο υπουργείο είναι επικεφαλής ο Αλογοσκούφης;» ή «Ποιος οργανισμός διαχειρίζεται τα τυχερά παιχνίδια;»), και η απάντηση είναι είτε ένα αντικείμενο (π.χ. «τη Φιλική Εταιρία») είτε ένας επιρρηματικός προσδιορισμός (π.χ. «στο Υπουργείο Οικονομίας») ή ένα υποκείμενο (π.χ. «ο ΟΠΑΠ»). Λόγω αυτής της μεγαλύτερης ποικιλίας, δεν είναι τόσο εύκολο για τις μηχανές αναζήτησης των δύο εφημερίδων να συσχετίσουν το δοσμένο υποκείμενο ή αντικείμενο με το ζητούμενο αντικείμενο, υποκείμενο ή επιρρηματικό προσδιορισμό. Αυτό έχει ως αποτέλεσμα να επιστρέφονται από τις μηχανές αναζήτησης και κείμενα τα οποία περιέχουν τις λέξεις-κλειδιά, αλλά δεν περιέχουν τη ζητούμενη απάντηση. Για το λόγο αυτό είναι μικρότερο και το ποσοστό επιτυχίας του συστήματος στην κατηγορία αυτή.

Το ίδιο ισχύει και στην κατηγορία των ερωτήσεων χρόνου, αλλά σε πολύ μεγαλύτερο βαθμό. Οι μηχανές αναζήτησης των δύο εφημερίδων επιστρέφουν πολλά κείμενα τα οποία περιέχουν τις λέξεις-κλειδιά της ερώτησης, αλλά δεν περιέχουν τη ζητούμενη χρονική έκφραση. Αντίθετα, μπορεί να περιέχουν άλλες χρονικές εκφράσεις, οι οποίες μπερδεύουν το σύστημα, ή ακόμα μπορεί και να μην περιέχουν καθόλου χρονικές εκφράσεις. Το πρόβλημα αυτό οφείλεται, κυρίως, στο ότι δεν μπορούμε να υποδείξουμε στις μηχανές αναζήτησης ότι θέλουμε να μας επιστρέψουν κείμενα που να περιέχουν χρονικές εκφράσεις. Για το λόγο αυτό, το ποσοστό επιτυχίας του συστήματος για την κατηγορία αυτή είναι το μικρότερο.

Ένα άλλο σημαντικό πρόβλημα είναι πως η χρήση των ΜΔΥ στο στάδιο της αξιολόγησης και επιλογής απαντήσεων φαίνεται να επιφέρει πολύ μικρή βελτίωση των αποτελεσμάτων, ιδιαίτερα στις κατηγορίες των ερωτήσεων προσώπων και οργανισμών, έναντι απλούστερων μεθόδων που απλά επιστρέφουν τη συχνότερη υποψήφια απάντηση. Αυτό φαίνεται να υποδηλώνει πως οι ΜΔΥ δεν καταφέρνουν να μάθουν ικανοποιητικά μοντέλα διαχωρισμού των ορθών από τις λανθασμένες υποψήφιες απαντήσεις, κάτι που μπορεί να οφείλεται στο ότι οι ιδιότητες δεν παρέχουν επαρκείς πληροφορίες. Στη διάρκεια της εργασίας εκτελέσθηκαν πρόσθετα πειράματα, στα οποία εξετάσαμε τις καμπύλες μάθησης των ΜΔΥ, δηλαδή το πόσο βελτιώνονται τα αποτελέσματά τους όσο αυξάνεται ο αριθμός των ερωτήσεων εκπαίδευσης. Δεν δείχνουμε εδώ τα αποτελέσματα αυτών των πειραμάτων, γιατί τα διαστήματα εμπιστοσύνης των μετρήσεών τους δείχνουν πως οι μετρήσεις δεν είναι αξιόπιστες.⁴ Παρ' όλα αυτά, οι καμπύλες μάθησης, που συχνά είναι οριζόντιες ήδη από πολύ μικρούς αριθμούς ερωτήσεων εκπαίδευσης, δείχνουν και αυτές ότι οι ΜΔΥ αντιμετωπίζουν σοβαρά προβλήματα στο στάδιο της εκπαίδευσης.

Τέλος, θα πρέπει να σημειωθεί ότι τα αποτελέσματα του συστήματος αυτής της εργασίας επηρεάζονται από τα ποσοστό ορθότητας του συστήματος αναγνώρισης ονομάτων οντοτήτων που χρησιμοποιούμε.

Στους πιο πρόσφατους διαγωνισμούς TREC που αφορούν τον τομέα των συστημάτων ερωταποκρίσεων (Question Answering Track⁵), το σύνολο των ερωτήσεων αξιολόγησης αποτελείται από ερωτήσεις πολλών διαφορετικών κατηγοριών. Ανάμεσά τους περιλαμβάνονται ερωτήσεις των οποίων οι απαντήσεις είναι ονόματα προσώπων, οργανισμών ή χρονικές εκφράσεις, αλλά και ερωτήσεις πολλών άλλων κατηγοριών. Τα καλύτερα αποτελέσματα των τριών τελευταίων διαγωνισμών παρουσιάζονται στον παρακάτω πίνακα [15, 16, 17].

Διαγωνισμοί TREC (QA Track)	Καλύτερα αποτελέσματα
2004	0.77
2005	0.713

⁴ Τα αποτελέσματα αυτών των πειραμάτων περιλαμβάνονται στο συνοδευτικό CD της εργασίας.

⁵ Βλ. <http://trec.nist.gov/>.

2006	0.578
------	-------

Τα αποτελέσματα των διαγωνισμών αυτών δεν είναι άμεσα συγκρίσιμα με τα δικά μας, γιατί προκύπτουν ως μέσοι όροι των αποτελεσμάτων για όλες τις κατηγορίες ερωτήσεων και, όπως προαναφέρθηκε, οι κατηγορίες ερωτήσεων των διαγωνισμών TREC είναι περισσότερες από τις τρεις κατηγορίες της παρούσας εργασίας.

4.3.2 Πειράματα με μόνο μία ΜΔΥ και για τις τρεις κατηγορίες ερωτήσεων

Σε αυτά τα πειράματα χρησιμοποιήθηκε μία κοινή ΜΔΥ και για τις τρεις κατηγορίες ερωτήσεων, με τρεις επιπλέον ιδιότητες, μία για κάθε κατηγορία ερώτησης. Χρησιμοποιήσαμε 90 ερωτήσεις προσώπων, 90 ερωτήσεις οργανισμών και 90 χρονικές ερωτήσεις. Κατά τα άλλα χρησιμοποιήθηκαν οι ίδιες ρυθμίσεις, όπως και στα προηγούμενα πειράματα.

Τα ποσοστά επιτυχίας αυτού του πειράματος παρουσιάζονται στον παρακάτω πίνακα. Στις παρενθέσεις παρουσιάζονται οι μέσοι όροι των αποτελεσμάτων από τη χρήση τριών διαφορετικών ΜΔΥ.

Πλήθος απαντήσεων	<i>I</i> απάντηση	5 απαντήσεις
Μέτρα επίδοσης		
<i>R</i>	0.09 (0.11)	0.32 (0.32)
<i>HF-NG</i>	0.24 (0.23)	0.50 (0.50)
<i>HF-G</i>	0.22 (0.23)	0.45 (0.46)
<i>SVM-NG</i>	0.20 (0.17)	0.36 (0.38)
<i>SVM-G-HM</i>	0.22 (0.21)	0.45 (0.49)
<i>SVM-G-HS</i>	0.29 (0.29)	0.56 (0.57)

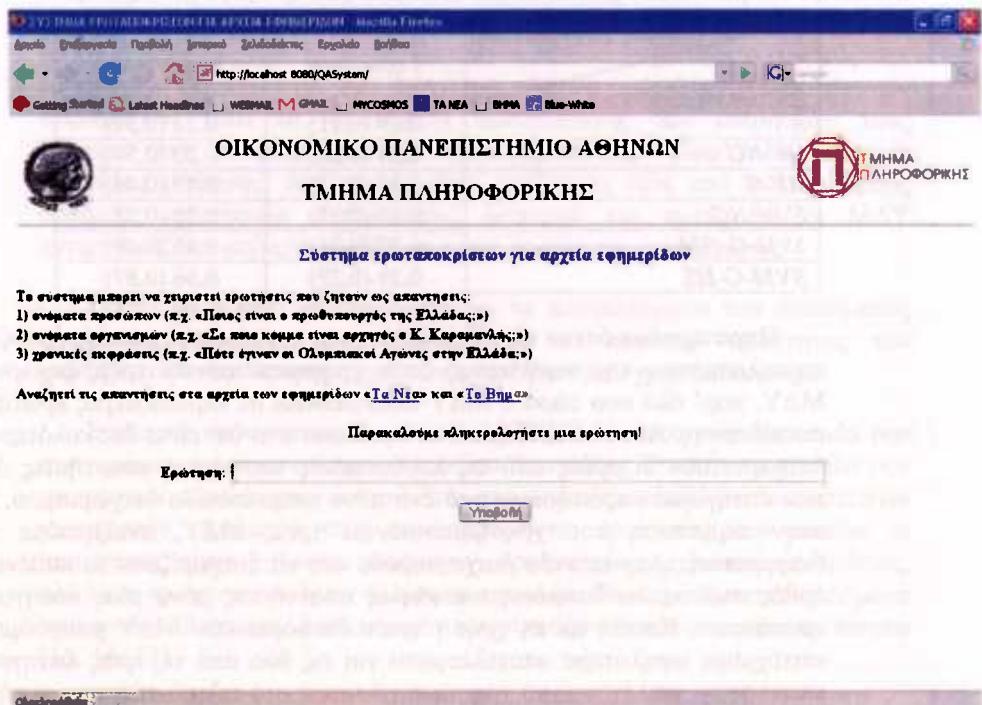
Παρατηρούμε ότι τα αποτελέσματα είναι κοντά στους μέσους όρους των αποτελεσμάτων της περίπτωσης όπου χρησιμοποιούνται τρεις ξεχωριστές ΜΔΥ, παρ' όλο που τώρα η ΜΔΥ εκπαιδεύεται σε περισσότερες ερωτήσεις εκπαίδευσης. Αυτό είναι πιθανό να οφείλεται στο ότι είναι δυσκολότερο να διαχωριστούν οι ορθές από τις λανθασμένες υποψήφιες απαντήσεις όλων των κατηγοριών ερωτήσεων από ένα μόνο υπερ-επίπεδο διαχωρισμού, ενώ στην περίπτωση που χρησιμοποιούνται τρεις ΜΔΥ, αναζητούμε τρία διαφορετικά υπερ-επίπεδα διαχωρισμού, που να διαχωρίζουν το καθένα τις ορθές από τις λανθασμένες υποψήφιες απαντήσεις μόνο μίας κατηγορίας ερωτήσεων. Επειδή με τη χρήση τριών διαφορετικών ΜΔΥ μπορούμε να επιτύχουμε υψηλότερα αποτελέσματα για τις δύο από τις τρεις κατηγορίες ερωτήσεων, επιλέγουμε να χρησιμοποιήσουμε στο τελικό σύστημα αυτή την προσέγγιση.



5 ΤΕΛΙΚΗ ΜΟΡΦΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Το τελικό σύστημα της εργασίας είναι δυνατόν να χρησιμοποιηθεί μέσω του Παγκόσμιου Ιστού. Η αρχική ιστοσελίδα του συστήματος παρουσιάζεται στην Εικόνα 1. Ο χρήστης πρέπει να πληκτρολογήσει την ερώτησή του και να επιλέξει «Υποβολή». Στη συνέχεια, το σύστημα βρίσκει την κατηγορία στην οποία ανήκει η ερώτηση του χρήστη (χρησιμοποιώντας το σύστημα της εργασίας [8]) και αναζητεί στα αρχεία των εφημερίδων «Τα Νέα» και «Το Βήμα» κείμενα τα οποία να περιέχουν τους όρους της ερώτησης. Από τα κείμενα αυτά εξάγονται οι υποψήφιες απαντήσεις (χρησιμοποιώντας το σύστημα αναγνώρισης ονομάτων οντοτήτων των εργασιών [4], [5], [6] και [7]), αξιολογούνται με τη μέθοδο SVM-G-HS και το σύστημα επιστρέφει στο χρήστη τις πέντε υποψήφιες απαντήσεις που θεωρεί καλύτερες.

Για την εκπαίδευση του συστήματος χρησιμοποιήθηκε μία διαφορετική ΜΔΥ για κάθε κατηγορία ερώτησης και κάθε ΜΔΥ εκπαιδεύτηκε σε όλα τα διαθέσιμα δεδομένα της κατηγορίας της. Κατά την επιλογή των υποψηφίων απαντήσεων, σε κάθε κατηγορία ερώτησης χρησιμοποιήθηκαν οι περιορισμοί που αναφέρθηκαν στις αντίστοιχες ενότητες του προηγούμενου κεφαλαίου.



Εικόνα 1: Αρχική ιστοσελίδα του τελικού συστήματος.

Στην Εικόνα 2 παρουσιάζεται ένα παράδειγμα ερώτησης του χρήστη, σωστής κατάταξής της από το σύστημα και απόκρισης του συστήματος, ενώ στις Εικόνες 3 και 4 παρουσιάζεται ένα παράδειγμα ερώτησης του χρήστη, λανθασμένης κατηγοριοποίησής της από το σύστημα, διόρθωσης της κατηγοριοποίησης από το χρήστη και απόκρισης του συστήματος.

ΣΥΣΤΗΜΑ ΕΡΩΤΑΙΚΟΡΙΣΕΩΝ ΕΙΑ ΑΡΧΕΙΑ ΕΦΗΜΕΡΙΔΩΝ ΑΠΟΚΡΙΣΗ ΣΥΣΤΗΜΑΤΟΣ Mozilla Firefox

Άρχισο Επεξεργασία Προβολή Ιστόσελο Σελίδαδείτες Εργαλεία Βοήθεια

http://localhost:8080/QASystem/QASystemInterface

Getting Started Latest Headlines WEBMAIL GMAIL MYCOSMOS TANEA BIMA Blue-White

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εισιτηρια ερωταποκρισίων πα αρχεία εφημερίδων

Το συστήμα μπορεί να χειριστεί φραγμές που ζητούν ως απαντήσεις:

- 1) ονομάτα προσώπων (π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδας;»)
- 2) ονομάτα οργανώσεων (π.χ. «Σε ποιο κόμμα είναι αρχηγός ο Κ. Καραμανλής;»)
- 3) χρονικές εκφράσεις (π.χ. «Πότε γένοντας ο Ολυμπιακός Αγώνας στην Ελλάδα;»)

Αναζητείτε τις απαντήσεις στα οργανισμούς στην Έλληνα και στην Βίβλη.

Παρακαλούμε πλέονταρευτέτε μια φράγμα!

Ερωτήσεις: Ποιος είναι ο πρωθυπουργός της Ελλάδας

Αποκρίση στις ερωτήσεις

Το συστήμα δεκτήριει στις φραγμές που ζητούν ως απαντήσεις έναν άνωμα φραγμό.

Αλλιώς δεν είναι σαν το παρακαλούμε προσδιορίστε τι είναι η απάντηση ή αποτέλεσμα:

Το συστήμα δεκτήριει σε μοτίβα τις παρακαλεσμένες κατευθύνσεις εκφράζοντας φραγμές:

Απαντήσεις	Κέριση στα σωστά φράγματα
Κ. Καραμανλή / Κ. Καραμανλής / Κανονισμόνος Καραμανλή / Κανονιστικόνος Καραμανλή / Κανονιστικόνος Καραμανλή / Καραμανλή / Καραμανλή / ΚΑΡΑΜΑΝΛΗ	Σ2 Σ8 Σ1 Σ6 Σ9 Σ3 Σ7 Σ3
Γ. Καραμανλής / Καραμανλής / Καραμανλή / ΚΑΡΑΜΑΝΛΗ	Σ2 Σ1 Σ3 Σ7
Γ. Αλογοσφύρη / Γ. Αλογοσφύρης / Αλογοσφύρη / Αλογοσφύρη	Σ9 Σ8 Σ3
Κ. Σημετής / Κώστα Σημετής / Κώστας Σημετής / Σημετή / Σημετής	Σ4 Σ2 Σ4
Γ. Καραϊδηρη / Καραϊδηρη / Καραϊδηρης	Σ1

Ολοκληρώθηκε

Εικόνα 2: Παράδειγμα ερώτησης χρήστη, σωστής κατηγοριοποίησής της από το σύστημα και απόκρισης συστήματος.

ΣΥΣΤΗΜΑ ΕΡΩΤΑΙΚΟΡΙΣΕΩΝ ΕΙΑ ΑΡΧΕΙΑ ΕΦΗΜΕΡΙΔΩΝ ΑΠΟΚΡΙΣΗ ΣΥΣΤΗΜΑΤΟΣ Mozilla Firefox

Άρχισο Επεξεργασία Προβολή Ιστόσελο Σελίδαδείτες Εργαλεία Βοήθεια

http://localhost:8080/QASystem/QASystemInterface

Getting Started Latest Headlines WEBMAIL GMAIL MYCOSMOS TANEA BIMA Blue-White

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εισιτηρια ερωταποκρισίων πα αρχεία εφημερίδων

Το συστήμα μπορεί να χειριστεί φραγμές που ζητούν ως απαντήσεις:

- 1) ονομάτα προσώπων (π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδας;»)
- 2) ονομάτα οργανώσεων (π.χ. «Σε ποιο κόμμα είναι αρχηγός ο Κ. Καραμανλής;»)
- 3) χρονικές εκφράσεις (π.χ. «Πότε γένοντας ο Ολυμπιακός Αγώνας στην Ελλάδα;»)

Αναζητείτε τις απαντήσεις στα οργανισμούς στην Έλληνα και στην Βίβλη.

Παρακαλούμε πλέονταρευτέτε μια φράγμα!

Ερωτήσεις: Ποιος είναι αρχηγός ο Καραμανλής

Αποκρίση στις ερωτήσεις

Το συστήμα δεκτήριει στις φραγμές που ζητούν ως απαντήσεις έναν άνωμα φραγμό.

Αλλιώς δεν είναι σαν το παρακαλούμε προσδιορίστε τι είναι η απάντηση ή αποτέλεσμα:

Αλτηή ή κατηγορια φραγμής δεν εκποστηρίζεται από το σύστημα!

Ολοκληρώθηκε

Εικόνα 3: Παράδειγμα ερώτησης χρήστη και λανθασμένης κατηγοριοποίησής της από το σύστημα.

The screenshot shows a Firefox browser window with the URL <http://localhost:8080/QASystem/QASystemInterface>. The page title is "ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ". The main content area displays a question in Greek:

Σύστημα ερωταπακισμάτων για αρχεία εφημερίδων

Το σύστημα μαρτυρεί να χειρίζεται ερωτήσεις που ζητούν ή απονήσεις:

- 1) ονοματα προσωπών (π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδος;»)
- 2) ονοματα οργανισμών (π.χ. «Σε ποιο κομμα εντάξεις ο Κ. Καραμανλής;»)
- 3) χρονικές εκφράσεις (π.χ. «Πότε έγινε ο Ολυμπιακοί Αγώνες στην Ελλάδα;»)

Αναζητείται η απάντηση στα ωρδά την εφημερίδη «[To Neos](#)» και «[To Vima](#)».

Παρακαλούμε πληκτρολόγηστε μια ερώτηση!

Ερώτηση: Πού είναι αρχηγός ο Καραμανλής?

Απόκριση στη στήματος:

Το σύστημα θεωρήσει σημείου τις απαρακούσιες απαντήσεις κατά φθινοπώντα στηρίζοντας

Απαντήση	Κείμενο στα οποία βρέθηκε
Παλαικ	V10 V7 V6 V4 V3
N Δ / N Δ / NΔ	N8 V10 N4 V7 N2 N1 V5 V4 N10 V3 V2 V1
ΣΕΛ	N6
Υπουργικό Διμβούλιο	N8 N10 V9 N2
Λασινιά Αμερική / ΕΛΑ / ΛΑΟΣ	N8 V1 N7 V1

Εικόνα 4: Παράδειγμα διόρθωσης της κατηγορίας της ερώτησης από το χρήστη και απόκρισης του συστήματος.

Στην πρώτη στήλη των πινάκων παρουσιάζονται οι απαντήσεις τις οποίες θεώρησε σωστές το σύστημα κατά φθινοπώντα σειρά βεβαιότητας. Στη δεύτερη στήλη παρουσιάζονται τα κείμενα από τα οποία προέκυψε η κάθε απάντηση. Τα ονόματα των κειμένων είναι και σύνδεσμοι προς τα ίδια τα κείμενα.

6 ΣΥΜΠΕΡΑΣΜΑΤΑ, ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

6.1 ΣΥΝΟΨΗ

Στη διάρκεια αυτής της εργασίας αναπτύχθηκε ένα σύστημα ερωταποκρίσεων για αρχεία ελληνικών εφημερίδων. Το σύστημα υποστηρίζει ερωτήσεις προσώπων, ερωτήσεις οργανισμών και χρονικές ερωτήσεις. Χρησιμοποιεί προαιρετικά λογισμικό προηγούμενης εργασίας, το οποίο κατατάσσει τις ερωτήσεις σε κατηγορίες. Εναλλακτικά, η κατηγορία της κάθε ερώτησης μπορεί να προσδιοριστεί από τον ίδιο το χρήστη. Το σύστημα της παρούσας εργασίας χρησιμοποιεί επίσης λογισμικό άλλων προηγούμενων εργασιών, το οποίο εντοπίζει ονόματα προσώπων, οργανισμών και χρονικές εκφράσεις σε ελληνικά κείμενα. Σε κάθε ερώτηση, υποψήφιες απαντήσεις είναι τα ονόματα προσώπων, οργανισμών ή οι χρονικές εκφράσεις, ανάλογα με την κατηγορία της ερώτησης, που εντοπίζονται στα κείμενα τα οποία ανακτώνται από τα αρχεία των εφημερίδων με τους όρους της ερώτησης. Το σύστημα της εργασίας αξιολογεί κάθε μία υποψήφια απάντηση χρησιμοποιώντας μια Μηχανή Διανυσμάτων Υποστήριξης (ΜΔΥ). Τα πειραματικά αποτελέσματα της εργασίας έδειξαν ότι είναι προτιμότερο να χρησιμοποιούνται τρεις διαφορετικές ΜΔΥ, μία για κάθε κατηγορία ερωτήσεων, αντί για μία κοινή ΜΔΥ. Έδειξαν, όμως, και ότι οι ΜΔΥ της εργασίας αντιμετωπίζουν σοβαρά προβλήματα κατά το στάδιο της εκπαίδευσης, με αποτέλεσμα να οδηγούν σε ελαφρά μόνο καλύτερα αποτελέσματα έναντι απλούστερων μεθόδων αξιολόγησης των υποψηφίων απαντήσεων, που δεν χρησιμοποιούν μηχανική μάθηση.

6.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Αν και προσπαθήσαμε, δεν μπορέσαμε να βρούμε επιπλέον ιδιότητες, οι οποίες να βελτιώσουν την επίδοση του συστήματος. Μια ιδέα ήταν να χρησιμοποιήσουμε επιπλέον ιδιότητες που να αντιστοιχούν στα μέτρα ομοιότητας της εργασίας [18], τα οποία έχουν χρησιμοποιηθεί για την εύρεση παραφράσεων, ώστε να εντοπίζουμε φράσεις που είναι παραφράσεις της ερώτησης. Πειράματα που έγιναν στη διάρκεια της εργασίας με τις επιπλέον αυτές ιδιότητες δεν έδειξαν βελτίωση των αποτελεσμάτων. Αντιθέτως, οδήγησαν σε μείωση του ποσοστού ορθότητας σχεδόν κατά 10%. Αυτό ενδέχεται να οφείλεται στο ότι οι υλοποιήσεις των μέτρων ομοιότητας που χρησιμοποιήθηκαν έχουν κατασκευαστεί για τα αγγλικά και δεν είναι σίγουρο αν συμπεριφέρονται σωστά σε ελληνικά κείμενα. Ή α είχε ενδιαφέρον να επανεξεταστεί αυτή η προσέγγιση, αφού ελεγχθούν ή και επανα-υλοποιηθούν τα μέτρα ομοιότητας για ελληνικά κείμενα.

Από τα πειράματα που διεξήχθησαν κατά τη διάρκεια της ανάπτυξης του συστήματος, συμπεράναμε ότι σημαντικότατο ρόλο στην επιτυχία του συστήματος παίζει η γνώση του τρόπου λειτουργίας των μηχανών αναζήτησης των εφημερίδων και των δυνατοτήτων που προσφέρουν κατά τη σύνταξη των ερωτημάτων προς αυτές. Οι σχετικές πληροφορίες που είχαμε για τις μηχανές αναζήτησης των εφημερίδων που χρησιμοποιήσαμε ήταν πολύ περιορισμένες, κάτι που οδήγησε σε σοβαρά προβλήματα, ιδιαίτερα στις χρονικές ερωτήσεις, όπου συχνά δεν μπορούσαμε να ανακτήσουμε κείμενα που να περιέχουν τις ορθές απαντήσεις. Ή α είχε ενδιαφέρον το σύστημα της εργασίας να βελτιωθεί σε

συνεργασία με τους κατασκευαστές ή διαχειριστές των μηχανών αναζήτησης των ιδίων ή άλλων εφημερίδων. Επιπλέον, το σύστημα θα μπορούσε να τροποποιηθεί, έτσι ώστε να μη χρησιμοποιεί μόνο εφημερίδες αλλά και γενικές μηχανές αναζήτησης, όπως το Google, καθώς και την ελληνική έκδοση της ηλεκτρονικής εγκυκλοπαίδειας Wikipedia.⁶

⁶ Βλ. <http://el.wikipedia.org/>



7 ΑΝΑΦΟΡΕΣ

- [1] M. Pasca, «*Open-Domain Question Answering from Large Text Collections*», CSLI, Stanford University, 2003.
- [2] S. M. Harabagiu, S. J. Maiorano και M.A. Pasca, «*Open-Domain Textual Question Answering Techniques*», Natural Language Engineering, 9(3):231–267, 2003.
- [3] H.T. Ng, J.L.P. Kwan και Y. Xia, «*Question Answering Using a Large Text Database: A Machine Learning Approach*». Πρακτικά του συνεδρίου *Empirical Methods in Natural Language Processing* (EMNLP 2001), Carnegie Mellon University, Η.Π.Α., 2001.
- [4] Γ. Λουκαρέλλι, «Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα», διπλωματική εργασία μεταπτυχιακού διπλώματος ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.
- [5] Ξ. Βασιλάκος, «Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα με χρήση μηχανών διανυσμάτων υποστήριξης», πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.
- [6] Γ. Λουκαρέλλι και I. Ανδρουτσόπουλος, «*A Greek Named-Entity Recognizer that Uses Support Vector Machines and Active Learning*». Πρακτικά του 4ου Πανελλήνιου Συνεδρίου Τεχνητής Νοημοσύνης (ΣΕΤΝ 2006), Ηράκλειο Κρήτης, 2006.
- [7] Γ. Λουκαρέλλι, Ξ. Βασιλάκος και I. Ανδρουτσόπουλος, «*Named Entity Recognition in Greek Texts with an Ensemble of Support Vector Machines and Active Learning*», International Journal on Artificial Intelligence Tools, World Scientific.
- [8] X. Brusagiotis, «Αυτόματη κατάταξη ελληνικών ερωτήσεων σε κατηγορίες», πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2007.
- [9] E.M. Voorhees, «*The TREC Question Answering Track*», Natural Language Engineering, 7(4):361–378, 2001.
- [10] W.A. Woods, R.M. Kaplan και B.N. Webber, «*The Lunar Sciences Natural Language Information System: Final Report*», BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, Massachusetts, 1972.
- [11] I. Ανδρουτσόπουλος, G.D. Ritchie και P. Thanisch, «*Natural Language Interfaces to Databases - An Introduction*». Natural Language Engineering, 1(1):29-81, Cambridge University Press, 1995.
- [12] T.M. Mitchell, «*Machine Learning*», McGraw-Hill, 1997.
- [13] Stuart Russell and Peter Norvig, «*Artificial Intelligence: A Modern Approach*», 2^η έκδοση, Prentice Hall, 2002.
- [14] Δ. Γαλάνης, «Αυτόματη κατασκευή παραδειγμάτων εκπαίδευσης για το χειρισμό ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων που χρησιμοποιούν μηχανική μάθηση», πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2004.



- [15] Ellen M. Voorhees. 2005. *Overview of the TREC 2004 question answering track*. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 52-62.
- [16] Ellen M. Voorhees and Hoa T. Dang. 2005. *Overview of the TREC 2005 question answering track*. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.
- [17] Hoa T. Dang, Jimmy Lin and Diane Kelly. 2006. *Overview of the TREC 2006 question answering track*. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*.
- [18] Π. Μαλακασιώτης και I. Ανδρουτσόπουλος, «*Learning Textual Entailment using SVMs and String Similarity Measures*». Πρόκειται να παρουσιαστεί στο ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Πράγα, Τσεχία, 2007.



