ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

θέμα:

«Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων σε Ελληνικά Κείμενα με Χρήση Μηχανών Διανυσμάτων Υποστήριξης»

Βασιλάκος Ξενοφών

Επιβλέπων : Ανδρουτσόπουλος Ιωάννης

Αθήνα, Σεπτέμβριος 2006

Περίληψη

Η παρούσα εργασία ασχολείται με την αναγνώριση και κατάταξη ονομάτων προσώπων και οργανισμών σε ελληνικά κείμενα, με χρήση συνδυασμών Μηχανών Διανυσμάτων Υποστήριξης (ΜΔΥ - Support Vector Machines). Αποτελεί επέκταση και βελτίωση προηγούμενης εργασίας, που επικεντρώθηκε στην αναγνώριση ονομάτων προσώπων και χρονικών εκφράσεων. Η εκπαίδευση και αξιολόγηση του συστήματος της εργασίας έγιναν σε συλλογή ελληνικών κειμένων προερχόμενων από άρθρα του ημερησίου τύπου. Η εκπαίδευση του συστήματος περιελάμβανε πειράματα παθητικής και ενεργητικής μάθησης, με δύο περάσματα των κειμένων και χρήση διαφορετικών ΜΔΥ σε κάθε πέρασμα. Ειδικότερα, στην ενεργητική μάθηση δοκιμάστηκαν 4 μέθοδοι επιλογής παραδειγμάτων εκπαίδευσης. Πέραν αυτού, δοκιμάστηκαν δύο διαφορετικοί τρόποι εκπαίδευσης των ΜΔΥ του δεύτερου περάσματος. Σε γενικές γραμμές, τα πειράματα έδειξαν ότι η ενεργητική μάθηση είναι προτιμότερη από την παθητική, αφού οδηγεί σε υψηλότερα ποσοστά ορθότητας με τον ίδιο αριθμό παραδειγμάτων εκπαίδευσης, αποτέλεσμα που συμφωνεί με τις διαπιστώσεις της προηγούμενης εργασίας. Έδειξαν, επίσης, ότι η χρήση του δεύτερου περάσματος οδηγεί και σε σημαντική βελτίωση της ταχύτητας του συστήματος, όταν χρησιμοποιείται μεγάλος αριθμός παραδειγμάτων εκπαίδευσης.

Ευχαριστίες

Οφείλω να ευχαριστήσω το Λέκτορα του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών κ. **Ιωάννη Ανδρουτσόπουλο**, για την επίβλεψή του κατά τη διάρκεια της εκπόνησης της παρούσας εργασίας. Επίσης ευχαριστώ τον υποψήφιο διδάκτορα του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών κ. **Γεώργιο Λουκαρέλλι** για την ευγενική παραχώρηση του συστήματος που είχε αναπτύξει, τις ιδιαίτερα χρήσιμες συμβουλές του, τις ιδέες και τις απόψεις του.

Περιεχόμενα

1. Εισαγωγή

    1.1. Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων (NERC)

    1.2. Σκοπός παρούσας εργασίας / σύστημα που επεκτάθηκε

    1.3. Μηχανές Διανυσμάτων Υποστήριξης (SVMs) (σύντομη αναφορά)

    1.4. Διάρθρωση παρόντος κειμένου

2. Προηγούμενες εργασίες

    2.1. Η εργασία του Λουκαρέλλι

    2.2. Πρόσθετη σημαντική βιβλιογραφία

3. Επεκτάσεις και Διαφοροποιήσεις συστήματος

    3.1. Χρήση δύο μηχανών διανυσμάτων υποστήριξης

    3.2. Επιπλέον ιδιότητες διανυσμάτων

    3.3. Ενεργητική Μάθηση

4. Διεξαγωγή Πειραμάτων

4.1. Συλλογές κειμένων (δεδομένα εκπαίδευσης και ελέγχου)

4.2. Συντονισμός / Επιλογή ιδιοτήτων

4.3. Πρώτο Πέρασμα

4.3.1. Συντονισμός ΜΔΥ (SVMs tuning)
4.3.2. Επιλογή ιδιοτήτων
4.3.3. Παθητική μάθηση
4.3.4. Ενεργητική Μάθηση

4.4. Δεύτερο Πέρασμα

4.4.1. Συντονισμός ΜΔΥ (SVMs tuning)
4.4.2. Επιλογή ιδιοτήτων
4.4.3. Παθητική μάθηση
4.4.4. Ενεργητική Μάθηση , 1η προσέγγιση
4.4.5. Ενεργητική Μάθηση , 2η προσέγγιση

4.5.Σύγκριση και σχολιασμός

4.5.1. Πρώτο Πέρασμα
4.5.2. Δεύτερο Πέρασμα
4.5.2.1. 1η προσέγγιση
4.5.2.2. 2η προσέγγιση
4.5.3. Χρόνος διεξαγωγής πειραμάτων

5. Προτάσεις για βελτίωση του συστήματος

6. Αναφορές σε άλλες εργασίες

ΠΑΡΑΡΤΗΜΑ

I. Επεξήγηση συντομογραφιών

II. Επεξήγηση υπολογισμού διαστημάτων εμπιστοσύνης

1. Εισαγωγή

1.1 Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων

Η **Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων** (*Named Entity Recognition and Categorization* – ***NERC***) έχει ως στόχο της τον εντοπισμό ονομάτων οντοτήτων που εμφανίζονται μέσα σε κείμενα. Παραδείγματα τέτοιων ονομάτων οντοτήτων είναι τα ονόματα προσώπων (π.χ. *Ιωάννης-Παύλος Β'*, *Ι. Παπαδόπουλος*), οργανισμών (π.χ. *ΟΤΕ, Ε.Ε.*), τοπωνυμίων (*Λεωφ. Καραμανλή 10, Αττική*), προϊόντων κ.ά. Η Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων μπορεί να φανεί χρήσιμη ως *προκαταρκτικό στάδιο σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας*. Τέτοια συστήματα είναι, για παράδειγμα, τα Συστήματα Εξαγωγής Πληροφοριών από κείμενα, τα Συστήματα Ερωταποκρίσεων για συλλογές κειμένων κλπ. Στο παρελθόν έχει καταβληθεί σημαντική προσπάθεια με αξιοσημείωτα αποτελέσματα πάνω στην κατασκευή συστημάτων NERC, κυρίως όμως για αγγλικά κείμενα (βλ. π.χ. [6, 8]).

Τα συστήματα NERC αξιολογούνται συνήθως ως προς την ακρίβεια (precision) και την ανάκλησή τους (recall) στις κατηγορίες ονομάτων οντοτήτων που υποστηρίζουν. Λέγοντας **«ακρίβεια»** μιας κατηγορίας ονομάτων οντοτήτων εννοούμε το ποσοστό των εκφράσεων που κατετάγησαν *ορθώς* από το σύστημα σε αυτή την κατηγορία, επί του συνόλου των εκφράσεων που κατετάγησαν από το σύστημα στην κατηγορία. Αντίθετα, ως **«ανάκληση»** μιας κατηγορίας ορίζεται το ποσοστό των εκφράσεων που κατετάγησαν ορθώς από το σύστημα σε αυτή την κατηγορία, επί του συνόλου των εκφράσεων της συλλογής κειμένων αξιολόγησης που ανήκουν πραγματικά σε αυτή την κατηγορία.

Από τα προηγούμενα καθίσταται σαφές ότι δεν αρκεί ένα σύστημα NERC να είναι μόνο ακριβές, δηλαδή όσες εκφράσεις σημειώνει ως ονόματα μιας κατηγορίας να ανήκουν πράγματι σε αυτή την κατηγορία. Πρέπει και να εντοπίζει όσο το δυνατόν περισσότερα ονόματα κάθε κατηγορίας.

1.2 Σκοπός παρούσας εργασίας / σύστημα που επεκτάθηκε.

Η παρούσα εργασία αποσκοπεί στην επέκταση και βελτίωση των δυνατοτήτων του ήδη ελεύθερα διαθέσιμου **Συστήματος NERC για ελληνικά κείμενα** που έχει αναπτυχθεί από τον κ. Λουκαρέλλι Γεώργιο [1], [15]. Το προϋπάρχον σύστημα υποστήριζε μόνο χρονικές εκφράσεις και ονόματα προσώπων. Για την αναγνώριση των χρονικών εκφράσεων χρησιμοποιούσε πρότυπα κανονικών εκφράσεων (*regular expression patterns*), ενώ για την αναγνώριση ονομάτων προσώπων χρησιμοποιούσε δύο **Μηχανές Διανυσμάτων Υποστήριξης** (**ΜΔΥ**, *Support Vector Machines*, *SVMs*), που χρησιμοποιούνταν σε δύο διαδοχικά περάσματα των κειμένων (μία ΜΔΥ ανά πέρασμα). Το σύστημα υποστήριζε, ακόμη, τόσο **παθητική μάθηση** (*passive learning*, **PL**) όσο και **ενεργητική μάθηση** (*active learning*, **AL**). Στην ενεργητική μάθηση το ίδιο το σύστημα προτείνει παραδείγματα εκπαίδευσης, τα οποία κατατάσσονται στη συνέχεια χειρωνακτικά από έναν άνθρωπο, και αυτό οδηγεί σε ταχύτερη βελτίωση των επιδόσεων του συστήματος κατά την εκπαίδευση και σε μείωση των παραδειγμάτων που απαιτείται να καταταγούν χειρωνακτικά.

Η παρούσα εργασία διατήρησε τις μεθόδους εντοπισμού χρονικών εκφράσεων της προηγούμενης, αλλά εισήγαγε περισσότερες ΜΔΥ, προκειμένου να εντοπίζονται τόσο ονόματα προσώπων όσο και οργανισμών. Πιο συγκεκριμένα, σε κάθε πέρασμα χρησιμοποιούνται τώρα δύο ΜΔΥ, μία για τον εντοπισμό ονομάτων προσώπων και μία για τον εντοπισμό ονομάτων οργανισμών, συνολικά 4 ΜΔΥ. Οι ΜΔΥ κάθε περάσματος εκπαιδεύονται ταυτόχρονα, κάτι που επιτρέπει στην ενεργητική μάθηση να επιλέγει παραδείγματα εκπαίδευσης βάσει των «προτιμήσεων» και των δύο ΜΔΥ του αντίστοιχου περάσματος. Επιπλέον, το νέο σύστημα υποστηρίζει περισσότερα μέτρα επιλογής παραδειγμάτων εκπαίδευσης κατά την ενεργητική μάθηση, η αποτελεσματικότητα των οποίων διερευνήθηκε πειραματικά. Έγιναν ακόμα οι ακόλουθες βελτιώσεις:

- Χρήση περισσότερων ιδιοτήτων (*attributes*) στη διανυσματική αναπαράσταση των προς κατάταξη εκφράσεων.
- Ανάπτυξη «φιλικής» γραφικής διεπαφής χρήστη, που υποστηρίζει τόσο το στάδιο εκπαίδευσης όσο και χρήσης του συστήματος.
- Δυνατότητα διεξαγωγής πολλαπλών πειραμάτων PL χωρίς να απαιτείται η παρουσία του χρήστη.
- Παροχή βοηθητικών λειτουργιών, όπως αυτόματη αποθήκευση αποτελεσμάτων και στατιστικών, επεξεργασία αρχείων αποτελεσμάτων /

στατιστικών, αυτόματη αλλαγή παραμέτρων από πείραμα σε πείραμα, αλλαγή διαδρομών δίσκου για τα χρησιμοποιούμενα αρχεία κ.ά.

Το παρόν σύστημα εκπαιδεύτηκε και αξιολογήθηκε σε μια συλλογή ελληνικών κειμένων προερχόμενων από τον ημερήσιο τύπο. Το θεματολόγιο της συλλογής περιελάμβανε ευρύ φάσμα πολιτικών, αθλητικών, πολιτιστικών και άλλων άρθρων. Αυτό το γεγονός καθιστά από μόνο του το εγχείρημα της εκπαίδευσης δύσκολο. Ενδεικτικά αναφέρουμε ότι τα αποτελέσματα F-measure (συνδυασμός ακρίβειας και ανάκλησης, βλ. παρακάτω) για τη κατηγορία προσώπων στην εργασία του Λουκαρέλλι προσέγγιζαν το 87,5% πάνω σε αυτή τη συλλογή κειμένων. Αντίθετα, το ίδιο σύστημα είχε πετύχει F-measure 93,94% πάνω σε μια δεύτερη συλλογή από αμιγώς οικονομικά κείμενα.

1.3 Μηχανές Διανυσμάτων Υποστήριξης

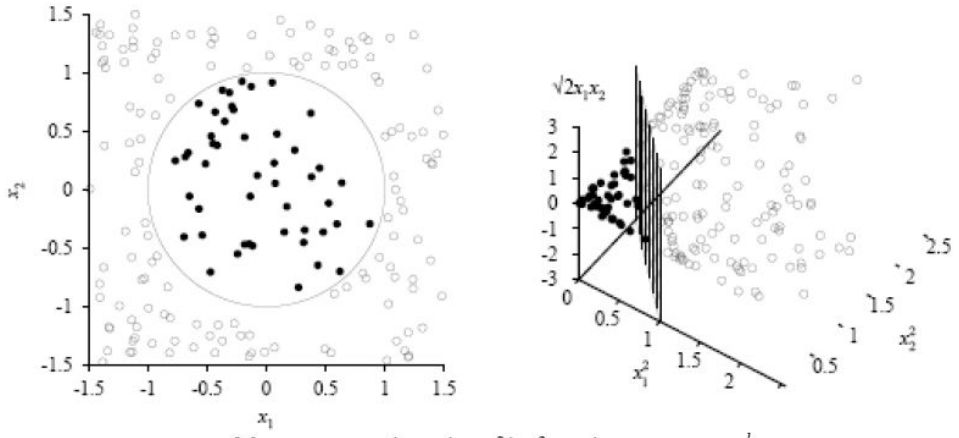
Οι **Μηχανές Διανυσμάτων Υποστήριξης** *(***ΜΔΥ***, Support Vector Machines)* [4][12][13] ανήκουν στους *αλγόριθμους Επιβλεπόμενης Μηχανικής Μάθησης* με αξιοσημείωτη επιτυχία σε προβλήματα κατάταξης. Όπως και οι περισσότεροι αλγόριθμοι μηχανικής μάθησης, παριστάνουν τα προς κατάταξη αντικείμενα ως διανύσματα ιδιοτήτων. Στην περίπτωσή μας, τα προς κατάταξη αντικείμενα είναι λεκτικές μονάδες (tokens) και οι ιδιότητες παρέχουν πληροφορίες όπως αν η προς κατάταξη λεκτική μονάδα ξεκινά με κεφαλαίο, αν έχει στα αριστερά της «κ.» ή στα δεξιά της «Α.Ε.» κλπ.

Κάθε ΜΔΥ χρησιμοποιείται συνήθως για το διαχωρισμό δύο κατηγοριών. Στην περίπτωσή μας, υπάρχει μία ΜΔΥ για κάθε υποστηριζόμενο είδος ονομάτων, και η ΜΔΥ εκπαιδεύεται στο να διαχωρίζει τις λεκτικές μονάδες που αποτελούν μέρη ονομάτων του συγκεκριμένου είδους (θετική κατηγορία) από τις υπόλοιπες (αρνητική κατηγορία). Κατά το στάδιο της εκπαίδευσης, η ΜΔΥ βρίσκει το βέλτιστο υπερεπίπεδο που διαχωρίζει τα διανύσματα εκπαίδευσης (**ΔΕκ**) των δύο κατηγοριών, δηλαδή το υπερεπίπεδο που τα διαχωρίζει με το μέγιστο περιθώριο. Το ίδιο υπερεπίπεδο χρησιμοποιείται στη συνέχεια για το διαχωρισμό νέων αντικειμένων των οποίων δεν είναι γνωστές οι κατηγορίες. Οι ΜΔΥ προβάλλουν συχνά τα διανύσματα σε ένα νέο χώρο περισσότερων διαστάσεων, προκειμένου να γίνει ευκολότερος ο διαχωρισμός των δύο κατηγοριών, που ενδέχεται να μην είναι γραμμικά διαχωρίσιμες στον αρχικό χώρο. . Η μετάβαση στο νέο διανυσματικό χώρο γίνεται με τη χρήση μίας συνάρτησης μετασχηματισμού.

Ένα απλό παράδειγμα αλλαγής διανυσματικού χώρου είναι το ακόλουθο, όπου με βάση τη συνάρτηση μετασχηματισμού $\mathbf{F(x) = <x_1^2, x_2^2, \sqrt{2}\cdot x_1\cdot x_2>}$ γίνεται μετάβαση από το χώρο των δύο διαστάσεων σε εκείνο των τριών διαστάσεων. Εκεί μπορεί να βρεθεί γραμμικός διαχωριστής, αντίθετα από τον αρχικό χώρο όπου τα δεδομένα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα.[1] Περισσότερες εισαγωγικές πληροφορίες για τις ΜΔΥ δίνονται στην εργασία του Λουκαρέλλι [1].

---

[1] Το παράδειγμα και το σχήμα προέρχονται από το βιβλίο «Artificial Intelligence – A Modern Approach» των S. Russel και P. Norvig, Prentice Hall.



Μετασχηματισμός από τις δύο διαστάσεις στις τρεις [1]

(1.4-1) Παράδειγμα μετασχηματισμού
από δύο σε τρεις διαστάσεις

1.4 Διάρθρωση παρόντος κειμένου

Στο ***κεφάλαιο 2*** γίνεται λόγος για τη σημαντικότερη βιβλιογραφία πάνω στην οποία στηρίζεται η παρούσα προσπάθεια. Αρχικά γίνεται μία περιληπτική αναφορά στην εργασία του κ. Λουκαρέλλι, εργασία πάνω στην οποία στηρίζεται σε μεγάλο βαθμό η παρούσα. (βλ και κεφάλαιο 6)

Στο ***κεφάλαιο 3*** περιγράφονται με αναλυτικό τρόπο οι επιμέρους επεκτάσεις του συστήματος σε σχέση με το προϋπάρχον. Παρατίθενται οι στόχοι που τέθηκαν και τα οφέλη που προέκυψαν από τις επεκτάσεις του συστήματος. Ειδικά για τις επεκτάσεις, πέραν του κεφαλαίου 3, ο αναγνώστης παραπέμπεται και στο ***Εγχειρίδιο Χρήσης*** του συστήματος που αναπτύχθηκε για τους σκοπούς της παρούσα εργασία.

Στο ***κεφάλαιο 4*** παρέχονται πρώτα πληροφορίες για τα δεδομένα εκπαίδευσης και ελέγχου. Ακολουθεί αναλυτική περιγραφή του τρόπου διεξαγωγής των πειραμάτων και λεπτομερής παρουσίαση και σχολιασμός των αποτελεσμάτων τους.

Στο ***κεφάλαιο 5*** προτείνονται ιδέες για τη βελτίωση του συστήματος και την αποφυγή λαθών, βάσει των εμπειριών που αποκομίσθηκαν από την παρούσα εργασία.

Τέλος, στο ***κεφάλαιο 6*** δίνονται όλες οι αναφορές τις παρούσας εργασίας σε άλλες.

Οι συντομογραφίες που χρησιμοποιούνται σε αυτή την εργασία παρατίθενται στο Παράρτημα, στο πεδίο Ι. Επίσης, το παράρτημα περιλαμβάνει και μία σύντομη επεξήγηση των διαστημάτων εμπιστοσύνης των διαγραμμάτων ανάκλησης και ακρίβειας (ΙΙ).

2. Προηγούμενες εργασίες

2.1 Η εργασία του Λουκαρέλλι

Όπως προαναφέρθηκε, η παρούσα εργασία αποτελεί επέκταση και βελτίωση της εργασίας του Λουκαρέλλι [1], [15]. Τόσο το κείμενο εκείνης της εργασίας, όσο και η βιβλιογραφία στην οποία αναφέρεται ενέπνευσαν τη παρούσα σε θέματα επιλογής ιδιοτήτων, αποφυγής λαθών και τρόπου διεξαγωγής των πειραμάτων.

Η πλειοψηφία των *ιδιοτήτων (attributes)* της παρούσας εργασίας προέρχεται από το σύστημα του Λουκαρέλλι. Σε αυτές προστέθηκαν και νέες, που προέκυψαν κυρίως από την εξέταση γειτονικών λεκτικών μονάδων σε μεγαλύτερη απόσταση από την υπό κατάταξη λεκτική μονάδα, όπως εξηγείται αναλυτικότερα παρακάτω..

Η παρούσα εργασία χρησιμοποίησε τα επισημειωμένα κείμενα εκπαίδευσης και ελέγχου της εργασίας του Λουκαρέλλι, καθώς επίσης και μεγάλο μέρος του λογισμικού της εργασίας του. Συγκεκριμένα αφομοιώθηκαν και / ή επεκτάθηκαν τα ακόλουθα τμήματα του λογισμικού της εργασίας του Λουκαρέλλι. Με «*» είναι σημειωμένα τα τμήματα που δεν τροποποιήθηκαν καθόλου, ενώ με «#» είναι σημειωμένα τα τμήματα που υπέστησαν επεκτάσεις ή βελτιώσεις:

- ο κώδικας διαχωρισμού των κειμένων σε λεκτικές μονάδες*,
- ο διαχωριστής των κειμένων σε περιόδους*,
- ο κώδικας αναγνώρισης των χρονικών εκφράσεων*,
- ο κώδικας εκπαίδευσης και ελέγχου του συστήματος για παθητική μάθηση#,
- το τμήμα κώδικα που αφορά την εξαγωγή τιμών ιδιοτήτων για τις λεκτικές μονάδες των κειμένων#,
- ο κώδικας που επιλέγει τα παραδείγματα εκπαίδευσης κατά την διεξαγωγή των πειραμάτων ενεργητικής μάθησης#,
- ο κώδικας συντονισμού (tuning) των παραμέτρων των ΜΔΥ*,
- το σύστημα αρχειοθέτησης των δεδομένων που χρησιμοποιούνται ή παράγονται από το σύστημα#,
- οι ασφαλείς κανόνες και οι λίστες τετριμμένων λέξεων (safe rules, stop-word Lists)*.

Όπως και στην εργασία του Λουκαρέλλι, εξακολουθεί να χρησιμοποιείται η ελεύθερα διαθέσιμη υλοποίηση των ΜΔΥ LibSVM [4] και το JNI (Java Native Interface) [14].

2.2 Πρόσθετη σημαντική βιβλιογραφία

Πέραν της βιβλιογραφίας που περιλαμβάνεται στην εργασία του Λουκαρέλλι, χρησιμοποιήθηκαν και οι ακόλουθες εργασίες.

- *«Multi-Criteria-based Active Learning for Named Entity Recognition»*, **Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, Chew-Lim Tan [11].** Η συγκεκριμένη δημοσίευση προτείνει, μεταξύ άλλων, διάφορους τρόπους επιλογής παραδειγμάτων εκπαίδευσης κατά την ενεργητική μάθηση. Με βάση τις ιδέες αυτής της δημοσίευσης, υλοποιήθηκαν και αξιολογήθηκαν πειραματικά στη διάρκεια της παρούσας εργασίας 4 διαφορετικά μέτρα επιλογής παραδειγμάτων εκπαίδευσης.

Τα συμπεράσματα, ωστόσο, στα οποία καταλήξαμε έρχονται σε αντίθεση με εκείνα της προαναφερθείσας δημοσίευσης. Συνοπτικά αναφέρουμε ότι τα πειραματικά μας αποτελέσματα δείχνουν πως η πιο απλή μέθοδος, που επιλέγει τα παραδείγματα των οποίων τα διανύσματα βρίσκονται πιο κοντά στο τρέχον υπερ-επίπεδο διαχωρισμού της ΜΔΥ, είναι και η καλύτερη. Αυτό ισχύει τόσο από πλευράς επιδόσεων (ακρίβεια, ανάκληση), όσο και από πλευράς απαιτήσεων μνήμης και χρόνου διεξαγωγής των πειραμάτων. Περισσότερες πληροφορίες παρέχονται σε επόμενες ενότητες.

- *«Active Learning with Support Vector Machines»,* **Ανδρέας Βλάχος [2]**

Η απόσταση ενός υποψηφίου παραδείγματος εκπαίδευσης από το τρέχον υπερ-επίπεδο διαχωρισμού της ΜΔΥ (ή τα άλλα, πιο περίπλοκα μέτρα που προτείνει η προηγούμενη εργασία) αποτελεί μια ένδειξη του πόσο χρήσιμο είναι το υποψήφιο παράδειγμα για την εκπαίδευση της ΜΔΥ. Στην περίπτωσή μας, όμως, έχουμε δύο ΜΔΥ ανά πέρασμα, οι οποίες εκπαιδεύονται παράλληλα, οπότε τίθεται το ερώτημα πώς θα συνδυάζονται οι ενδείξεις χρησιμότητας που προκύπτουν από τις δύο ΜΔΥ για κάθε υποψήφιο παράδειγμα, δηλαδή (στην απλούστερη περίπτωση) πώς θα συνδυάζονται οι αποστάσεις από τα υπερεπίπεδα διαχωρισμού. Ο Βλάχος μελέτησε, μεταξύ άλλων, διάφορους τρόπους συνδυασμού των αποστάσεων. Βάσει των πειραματικών αποτελεσμάτων που παρουσιάζει, επιλέξαμε να χρησιμοποιούμε τη μικρότερη από τις δύο αποστάσεις. Περισσότερες πληροφορίες παρέχονται και πάλι σε επόμενες ενότητες.

- *«Fast Uncertainty Sampling for Labeling Large e-mail Corpora»,* **Richard Segal, Ted Markowitz, William Arnold. [3]**

Η ιδέα που παρουσιάζεται στη δημοσίευση αυτή είναι ότι κατά τη σταδιακή εκπαίδευση ενός ταξινομητή με ενεργητική μάθηση, η χρησιμότητα των υποψηφίων παραδειγμάτων εκπαίδευσης (η απόστασή τους από το υπερεπίπεδο στην απλούστερη περίπτωση για μια ΜΔΥ) δεν αλλάζει σε μεγάλο βαθμό καθώς προστίθενται νέα παραδείγματα εκπαίδευσης. Άρα είναι σπατάλη χρόνου να επανυπολογίζονται οι αποστάσεις όλων των υποψηφίων παραδειγμάτων εκπαίδευσης μετά από την επιλογή κάποιων παραδειγμάτων εκπαίδευσης και την επανεκπαίδευση του ταξινομητή. Σύμφωνα με τα συμπεράσματα της δημοσίευσης, αρκεί ο επανυπολογισμός της χρησιμότητας να γίνεται για τα $M^*\log_2 N$ υποψήφια παραδείγματα που είχαν προηγουμένως τις υψηλότερες τιμές χρησιμότητας. Στον προηγούμενο τύπο, **N** είναι το πλήθος των υποψηφίων διανυσμάτων εκπαίδευσης και **M** ο αριθμός των παραδειγμάτων εκπαίδευσης που προστίθενται στα δεδομένα εκπαίδευσης σε κάθε επανάληψη της ενεργητικής μάθησης (το μέγεθος της «**δέσμης**»). Παρότι δεν έγινε τελικά χρήση αυτού του τρόπου, το σύστημα παρέχει αυτή την δυνατότητα για μελλοντική χρήση.

Σχετικές με την παρούσα εργασία είναι, επίσης, οι δημοσιεύσεις [9] και [10], οι οποίες εξετάζουν τη χρήση ενεργητικής μάθησης στην αναγνώριση αγγλικών ονομάτων οντοτήτων.

3. Επεκτάσεις και Διαφοροποιήσεις Συστήματος

3.1 Χρήση δύο Μηχανών Διανυσμάτων Υποστήριξης

Κύριο χαρακτηριστικό του συστήματος που αναπτύχθηκε είναι ότι χρησιμοποιεί παράλληλα δύο ΜΔΥ σε κάθε πέρασμα. Η κάθε μία από τις δύο ΜΔΥ εκπαιδεύεται στο να ταξινομεί τις λεκτικές μονάδες (που παριστάνονται ως διανύσματα ιδιοτήτων) ως προς μία κατηγορία ονομάτων, δηλαδή να αποφασίζει αν η κάθε λεκτική μονάδα αποτελεί μέρος ονόματος της συγκεκριμένης κατηγορίας ονομάτων ή όχι. Η πρώτη κατηγορία ονομάτων είναι εκείνη των προσώπων και η άλλη είναι εκείνη των οργανισμών. Οι δύο αυτές ΜΔΥ λειτουργούν ως ένα σύστημα σε κάθε πέρασμα, αφού εκπαιδεύονται ταυτόχρονα πάνω σε (σχεδόν ίδια – βλ. παρακάτω) παραδείγματα λεκτικών μονάδων, ενώ κατά την κατάταξη νέων λεκτικών μονάδων ερωτώνται και οι δύο ΜΔΥ.

- ***ΑΠΟΦΑΣΕΙΣ ΣΥΣΤΗΜΑΤΟΣ***

Όταν το σύστημα ερωτάται ως προς τη κατηγορία όπου ανήκει μία λεκτική μονάδα, τότε ακολουθεί την επόμενη σειρά ενεργειών:

- Υπολογίζονται τα διανύσματα ιδιοτήτων για την ΜΔΥ της κατηγορίας των προσώπων και αντίστοιχα για εκείνη των οργανισμών.

- Ερωτώνται οι δύο ΜΔΥ αν ανήκει η λεκτική μονάδα στις αντίστοιχες κατηγορίες ονομάτων. Διευκρινίζουμε ότι όπως και στην εργασία του Λουκαρέλλι, χρησιμοποιούνται επίσης«ασφαλείς κανόνες» και «λίστες τετριμμένων λέξεων» (stop-words), διαφορετικές για κάθε κατηγορία ονομάτων. Αν οι ασφαλείς κανόνες της κατηγορίας των προσώπων κατατάξουν τη λεκτική μονάδα ως μη-όνομα ή αν η λεκτική μονάδα περιέχεται στη λίστα τετριμμένων λέξεων της κατηγορίας των προσώπων, τότε η αντίστοιχη ΜΔΥ δεν ερωτάται καν. Ομοίως για την κατηγορία των οργανισμών.

- Οι κάθε μία ΜΔΥ αποκρίνεται με ένα βαθμό βεβαιότητας, κανονικοποιημένο στο [-1, +1], που δείχνει τη βεβαιότητά της ότι η λεκτική μονάδα ανήκει ή όχι στην αντίστοιχη κατηγορία ονομάτων.
    - Εάν και οι δύο βαθμοί βεβαιότητας είναι αρνητικοί, τότε η λεκτική μονάδα δεν θεωρείται ότι δεν αποτελεί μέρος ούτε ονόματος προσώπου ούτε ονόματος οργανισμού.
    - Διαφορετικά, η λεκτική μονάδα κατατάσσεται στην κατηγορία ονομάτων με τη μεγαλύτερη βεβαιότητα.

Εάν χρησιμοποιείται και δεύτερο πέρασμα, τότε η κατάταξη γίνεται βάσει των βαθμών βεβαιότητας των δύο ΜΔΥ του δεύτερου περάσματος, οι οποίες χρησιμοποιούν επιπλέον ιδιότητες που προκύπτουν από τους βαθμούς βεβαιότητας των ΜΔΥ του πρώτου περάσματος. Ο τρόπος λειτουργίας του δεύτερου περάσματος περιγράφεται λεπτομερέστερα σε επόμενες ενότητες.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΧΡΗΣΗΣ ΔΥΟ ΜΔΥ ΑΝΑ ΠΕΡΑΣΜΑ

Η ταυτόχρονη χρήση δύο ΜΔΥ σε κάθε πέρασμα, μίας για κάθε είδος ονομάτων, παρέχει μια σειρά πλεονεκτημάτων, σε σχέση με τη χρήση ενός μόνο ταξινομητή που θα κατέτασσε κάθε λεκτική μονάδα ως μέρος ονόματος προσώπου, οργανισμού ή τίποτα από τα δύο (συνολικά κατάταξη σε τρεις κατηγορίες):

- Οι ΜΔΥ υποστηρίζουν φυσικότερα το διαχωρισμό μεταξύ δύο κατηγοριών (στην περίπτωσή μας θετικές και αρνητικές περιπτώσεις για κάθε είδος ονομάτων).

- Η κάθε ΜΔΥ μπορεί να έχει το δικό της σύνολο ιδιοτήτων, το οποίο να είναι προσαρμοσμένο για τη κατηγορία ονομάτων στην οποία ειδικεύεται η συγκεκριμένη ΜΔΥ. Άλλα συστήματα τα οποία κατατάσσουν τις λεκτικές μονάδες σε περισσότερες από μία κατηγορίες χρησιμοποιώντας ένα μόνο ταξινομητή έχουν το μειονέκτημα ότι χρησιμοποιούν το ίδιο σύνολο ιδιοτήτων για όλες τις κατηγορίες ονομάτων.

Όπως προαναφέρθηκε, οι δύο ΜΔΥ του κάθε περάσματος εκπαιδεύονται σε σχεδόν τα ίδια παραδείγματα λεκτικών μονάδων. Γράφουμε «σχεδόν», επειδή λεκτικές μονάδες εκπαίδευσης που κατατάσσονται ως μη-πρόσωπα από τους ασφαλείς κανόνες ονομάτων προσώπων δεν δίνονται ως παραδείγματα εκπαίδευσης στη ΜΔΥ των ονομάτων προσώπων και ομοίως για τους ασφαλείς κανόνες των ονομάτων οργανισμών και τις λίστες τετριμμένων λέξεων. Με αυτόν τον τρόπο, επειδή οι ασφαλείς κανόνες και οι λίστες τετριμμένων λέξεων είναι εν

γένει διαφορετικές για τις δύο κατηγορίες ονομάτων, τα παραδείγματα λεκτικών μονάδων που δίνονται στις δύο ΜΔΥ κατά την εκπαίδευσή τους δεν είναι ακριβώς ίδια. Αν εξαιρέσουμε, όμως, τις λεκτικές μονάδες που απορρίπτονται από τους ασφαλείς κανόνες και τις λίστες τετριμμένων λέξεων, οι δύο ΜΔΥ συναντούν κατά την εκπαίδευσή τους τα ίδια παραδείγματα λεκτικών μονάδων.

Οι ασφαλείς κανόνες και οι λίστες τετριμμένων λέξεων χρησιμοποιούνται επειδή σε κάθε κατηγορία ονομάτων, τα αρνητικά παραδείγματα (λεκτικές μονάδες που δεν αποτελούν μέρη ονομάτων της αντίστοιχης κατηγορίας) είναι πολύ περισσότερα από τα θετικά, και αυτή η δυσαναλογία οδηγεί τους περισσότερους αλγορίθμους μάθησης να μάθουν να κατατάσσουν όλες τις περιπτώσεις ως αρνητικές. Με τη χρήση των ασφαλών κανόνων και των λιστών τετριμμένων λέξεων η δυσαναλογία αυτή μειώνεται δραστικά (βλ. [1], [15] για περισσότερες λεπτομέρειες).

3.2 Επιπλέον ιδιότητες διανυσμάτων

Πέραν των ιδιοτήτων που κληρονομήθηκαν από το σύστημα του Λουκαρέλλι [1], [15], το σύστημα της παρούσας εργασίας περιλαμβάνει και επιπλέον ιδιότητες. Αυτές οι ιδιότητες προέρχονται είτε από τη διεύρυνση του παραθύρου των γειτονικών λεκτικών μονάδων που εξετάζονται γύρω από την προς κατάταξη λεκτική μονάδα (το παράθυρο έχει μέγεθος 7 αντί 3 που είχε πριν, δηλαδή τώρα εξετάζονται εκτός από την προς κατάταξη λεκτική μονάδα και οι 3 προηγούμενες και οι 3 επόμενες λεκτικές μονάδες), είτε πρόκειται για εντελώς καινούργιες ιδιότητες. Κατά το **πρώτο πέρασμα (FP)** χρησιμοποιήθηκαν 165 και 170 ιδιότητες για πρόσωπα και οργανισμούς αντίστοιχα, ενώ στο **δεύτερο πέρασμα (SP)** χρησιμοποιήθηκαν 207 ιδιότητες για πρόσωπα και 217 ιδιότητες για οργανισμούς.

Ακολουθούν πίνακες με τις ιδιότητες που χρησιμοποιήθηκαν στο FP και SP. Το μέγεθος της τελείας που υπάρχει σε κάθε κελί αντιστοιχεί στο **πληροφοριακό κέρδος (IG - Information Gain)** της συγκεκριμένης ιδιότητας, δηλαδή ιδιότητες με υψηλότερο πληροφοριακό κέρδος παριστάνονται με μεγαλύτερες τελείες. *Το πληροφοριακό κέρδος μιας ιδιότητας είναι η αναμενόμενη μείωση της αβεβαιότητας (εντροπίας) ως προς την απόφαση που έχει να λάβει ο αντίστοιχος ταξινομητής, την οποία επιφέρει η γνώση τις τιμής της ιδιότητας.* Επομένως, ιδιότητες με υψηλότερο πληροφοριακό κέρδος είναι μάλλον πιο χρήσιμες στον ταξινομητή (βλ. [1] για τους σχετικούς μαθηματικούς τύπους). Τονίζουμε ότι το IG της κάθε ιδιότητας δεν αποδίδει κατά τρόπο απόλυτο τη χρησιμότητα της ιδιότητας αυτής για τη ΜΔΥ, αφού η ΜΔΥ συνδυάζει όλες τις τιμές ιδιοτήτων για να κατατάξει ένα υποψήφιο παράδειγμα. Η αξιολόγηση των ιδιοτήτων με βάση το IG αποτέλεσε ένα κριτήριο επιλογής του βέλτιστου συνόλου ιδιοτήτων για κάθε ΜΔΥ (βλ 4.4.2 και 4.5.2 για λεπτομέρειες).

Στις γραμμές του παρακάτω πίνακα αναφέρονται τα ονόματα των ιδιοτήτων του FP, ενώ στις στήλες η λεκτική μονάδα του παραθύρου για την οποία παρέχει πληροφορίες η ιδιότητα. Περισσότερες εξηγήσεις για τις ιδιότητες αυτές δίνονται στην εργασία [1],[15].

		1st pass SVM for person names							1st pass SVM for organization names							
no.	feature descriptions	t_{-3}	t_{-2}	t_{-1}	t_0	t_1	t_2	t_3	t_{-3}	t_{-2}	t_{-1}	t_0	t_1	t_2	t_3	
1–7	comma?	•	•	●	●	•	•	•	•	•	•	•	•	•	•	B
8–14	full stop?	•	•	●	●	●	•	•	•	•	•	•	•	•	•	B
15–21	dash?	•	•	•	•	•	•	•	•	•	•	•	•	•	•	B
22–28	slash?	•	0	•	0	•	•	0	•	•	•	•	•	•	0	B
29–35	number?	•	•	•	0	•	•	•	•	•	•	•	•	•	•	B
36–42	Greek characters?	•	●	•	•	•	•	•	●	●	●	•	•	•	•	B
43–49	Latin characters?	•	•	•	•	•	•	•	•	•	•	●	•	•	•	B
50–56	first character capital?	•	•	•	●	●	•	•	•	•	•	•	•	●	•	B
57–63	all characters capital?	•	●	●	●	•	•	•	•	•	•	•	•	•	•	B
64–70	length in characters	●	●	●	●	●	•	•	●	●	●	●	0	0	0	n
71–77	common surname prefix?	•	•	•	●	•	•	•	•	•	•	•	•	•	•	B
78–84	common surname suffix?	•	•	•	●	●	•	•	•	•	•	•	•	•	•	B
85–91	common first name?	●	●	●	●	•	0	•	●	●	●	●	•	•	0	n
92–98	common last character?	•	•	•	•	•	•	•	●	●	●	•	•	•	•	B
99–105	common sing. adj. ending?	•	•	•	●	•	•	•	•	•	•	●	•	•	•	B
106–112	plural noun/adj. ending?	•	•	•	•	•	•	•	•	•	•	•	•	•	•	B
113–119	common sing. gen. ending?	•	•	•	•	•	•	•	•	•	•	•	•	•	•	B
120–126	ends in final sigma?	•	•	•	•	•	•	•	•	●	●	•	•	•	•	B
127–133	distance from "A.E." etc.	•	•	•	•	0	0	0	•	•	•	•	•	•	•	n
134–140	starts with "ministry" etc.?	•	•	•	•	•	0	•	•	•	•	•	•	•	•	B
141–147	is last token of sentence?	•	•	●	•	•	•	•	•	•	●	•	•	•	•	B
148–154	part of article's title?	•	•	•	•	•	•	•	•	•	•	•	•	•	•	B
155–161	distance from start of name				●							●				B
162	directly preceded by Mr(s)?				●							●				B
163	preceded by plural Mr(s)?				●							•				B
164	in P_{1-2}^{t-1} list?			●							●					n
165	in P_{3-4}^{t-1} list?			•							●					n
166	in $P_{>4}^{t-1}$ list?			●							•					n
167	prev. tokens in $P_{1-2}^{t-7,\ldots,t-1}$				●							●				n
168	prev. tokens in $P_{3-4}^{t-7,\ldots,t-1}$				•							●				n
169	prev. tokens in $P_{>4}^{t-7,\ldots,t-1}$				•							•				n
170	in R_{1-2}^{t-1} list?			●							●					n
171	in R_{3-4}^{t-1} list?			●							●					n
172	in $R_{>4}^{t-1}$ list?			•							•					n
173	prev. tokens in $R_{1-2}^{t-7,\ldots,t-1}$				•							●				n
174	prev. tokens in $R_{3-4}^{t-7,\ldots,t-1}$				•							●				n
175	prev. tokens in $R_{>4}^{t-7,\ldots,t-1}$				0							●				n
	B: Boolean, n: numeric, 0 : $IG = 0$, • : $0 < IG \leq 0.01$, ● : $0.01 < IG \leq 0.1$, ● : $IG > 0.1$															

(3.2-1.1) Πίνακας αξιολόγησης ιδιοτήτων FP

Ακολουθεί ο πίνακας ιδιοτήτων για το δεύτερο πέρασμα (SP). Αυτός ο πίνακας περιλαμβάνει μόνο τις επιπλέον ιδιότητες του δεύτερου περάσματος σε σύγκριση με το πρώτο. (Κατά το δεύτερο πέρασμα χρησιμοποιούνται επίσης όλες οι ιδιότητες του πρώτου περάσματος). Οι επιπλέον ιδιότητες εξαρτώνται από τις αποφάσεις των ταξινομητών του πρώτου περάσματος (π.χ. υπάρχει επιπλέον ιδιότητα που δείχνει το βαθμό βεβαιότητας του πρώτου περάσματος ότι η προηγούμενη λεκτική μονάδα είναι μέρος ονόματος προσώπου). Για την εκτίμηση του πληροφοριακού κέρδους των επιπλέον ιδιοτήτων χρησιμοποιήθηκαν οι αποφάσεις των ταξινομητών του FP, εκπαιδευμένων σε περίπου 10900 διανύσματα εκπαίδευσης (Δεκ). Από τις [1] ,[15] υιοθετήθηκε η ιδέα των λιστών «σίγουρων αποφάσεων» (sure lists). Υπάρχουν, δηλαδή, δύο λίστες, μία για ονόματα οργανισμών και μία για ονόματα προσώπων, που περιέχουν λεκτικές μονάδες τις οποίες οι αντίστοιχοι ταξινομητές (ΜΔΥ) του πρώτου περάσματος είχαν κατατάξει με μεγάλη βεβαιότητα ως πρόσωπα ή οργανισμούς, αντίστοιχα, οπουδήποτε μέσα στο ίδιο κείμενο.

		2nd pass SVM for person names							2nd pass SVM for organization names							
no.	feature descriptions	t_{-3}	t_{-2}	t_{-1}	t_0	t_1	t_2	t_3	t_{-3}	t_{-2}	t_{-1}	t_0	t_1	t_2	t_3	
1–7	1st pass person confidence	0	•	•	●	•	•	•	•	•	•	•	•	•	•	n
8–14	1st pass org. confidence	•	•	•	•	•	•	•	•	•	•	•	•	•	•	n
15–21	in S_p?	0	•	0	•	•	•	•	0	0	•	•	•	0	0	B
22–28	in S_o?	0	•	•	0	•	•	0	•	•	•	•	•	•	•	B
29–35	distance from prev. person	•	•	0	•	•	•	•	•	•	•	•	•	•	•	n
36–42	distance from prev. org.	0	0	0	0	•	•	•	•	•	•	•	•	•	•	n
43–49	distance from next person	•	•	•	•	•	•	•	•	•	•	0	0	0	0	n
50–56	distance from next org.	0	0	0	•	•	•	•	•	•	•	•	•	•	0	n
B: Boolean, n: numeric, $0 : 0.001 < IG = 0$, $• : 0.001 < IG \leq 0.01$, $• : 0.01 < IG \leq 0.1$, $● : IG > 0.1$																

(3.2-1.2) Πίνακας επιπλέον ιδιοτήτων για SP

3.3 Ενεργητική Μάθηση

Για το σκοπό της AL έγιναν **δοκιμές** με 4 διαφορετικά μέτρα επιλογής παραδειγμάτων εκπαίδευσης. Σε κάθε επανάληψη της AL, τα μέτρα αυτά χρησιμοποιούνται για να επιλεγούν από μια δεξαμενή υποψηφίων παραδειγμάτων εκπαίδευσης εκείνα τα παραδείγματα που θεωρούνται χρησιμότερα. Τα επιλεγέντα παραδείγματα προστίθενται στη συνέχεια στο σύνολο των δεδομένων εκπαίδευσης του ταξινομητή, ο ταξινομητής επανεκπαιδεύεται στο νέο, διευρυμένο σύνολο δεδομένων και η διαδικασία AL επαναλαμβάνεται. Στις δοκιμές που έγιναν, σε κάθε επανάληψη επιλέγονταν 100 παραδείγματα εκπαίδευσης, τα οποία ονομάζουμε «δέσμη» (batch). Παρακάτω περιγράφονται τα μέτρα επιλογής που δοκιμάστηκαν. Όπως προαναφέρθηκε, τα μέτρα αυτά βασίζονται σε ιδέες της ομάδας των Dan Shen, κ.ά. [11].

1. Πληροφοριακό Όφελος

Το **Πληροφοριακό Όφελος (Informativeness)** ενός υποψήφιου διανύσματος εκπαίδευσης (ΔΕκ) είναι ένα μέγεθος που ορίζεται με βάση τη βεβαιότητα που θα απέδιδε ο ταξινομητής στην απόφασή του για το συγκεκριμένο διάνυσμα, εάν το τελευταίο ήταν διάνυσμα ελέγχου (ΔΕλ). *Όσο πιο βέβαιος είναι ο ταξινομητής για την απόφασή του, τόσο μικρότερο είναι το πληροφοριακό όφελος*.

Στην περίπτωση των ΜΔΥ, η βεβαιότητα της απόφασης μιας ΜΔΥ για ένα υποψήφιο διάνυσμα εκπαίδευσης αντιστοιχεί χονδρικά στην απόσταση του διανύσματος από το τρέχον υπερεπίπεδο διαχωρισμού της ΜΔΥ. *Όσο πιο κοντά στο υπερεπίπεδο βρίσκεται το υποψήφιο διάνυσμα εκπαίδευσης, τόσο μεγαλύτερο είναι το πληροφοριακό όφελός του*, επειδή θεωρείται χρησιμότερο για την εκπαίδευση της ΜΔΥ.

Πιο συγκεκριμένα, το Πληροφοριακό Όφελος κάθε υποψηφίου διανύσματος υπολογίζεται από τον παρακάτω τύπο (3.3.1), όπου η τιμή της κανονικοποιημένης πιθανότητας για μία απόφαση ενός ταξινομητή είναι ίση με τη πιθανότητα του να είναι θετικό το διάνυσμα ως προς την υπό εξέταση κατηγορία ονομάτων $P_{pos}(\vec{v})$, μείον τη πιθανότητα να είναι αρνητικό $P_{neg}(\vec{v})$. Οι τιμές των κανονικοποιημένων πιθανοτήτων ανήκουν στο διάστημα [-1,1], με το –1 να αποδίδεται στα σίγουρα

αρνητικά διανύσματα και το 1 στα σίγουρα θετικά αντίστοιχα. Όσο η τιμή της κανονικοποιημένης πιθανότητας τείνει προς το μηδέν, τόσο πιο διφορούμενη είναι η απόφαση της ΜΔΥ.

$$\mathbf{Info}(\vec{v}) = 1 - \mathbf{Dist}(\mathbf{SVM_{hyp}}, \vec{v}),$$
$$\mathbf{Dist}\,(\mathbf{SVM_{hyp}}, \vec{v}) = |\,\mathbf{P_{pos}}(\vec{v}) - \mathbf{P_{neg}}(\vec{v})\,|$$

(3.3-1) Τύπος υπολογισμού Πληροφοριακού Οφέλους

- Στα παραπάνω $\mathbf{Dist(SVM_{hyp},\vec{v})}$ είναι η απόσταση του υποψηφίου διανύσματος εκπαίδευσης $\vec{v}$ από το υπερεπίπεδο της ΜΔΥ $\mathbf{SVM_{hyp}}$, $\mathbf{P_{pos}}$ είναι η εκτίμηση της ΜΔΥ για την πιθανότητα να είναι θετικό το $\vec{v}$, δηλαδή να πρόκειται για όνομα προσώπου ή οργανισμού, αντίστοιχα,
- $\mathbf{P_{neg}}$ είναι η εκτίμηση της ΜΔΥ για την πιθανότητα να είναι αρνητικό το $\vec{v}$.

Στην περίπτωσή μας, όπου υπάρχουν δύο ΜΔΥ σε κάθε πέρασμα, προκύπτουν δύο τιμές Πληροφοριακού Οφέλους για κάθε υποψήφιο παράδειγμα εκπαίδευσης, μία από κάθε ΜΔΥ. Βάσει των συμπερασμάτων που περιγράφονται στην εργασία Βλάχου [2] , χρησιμοποιούμε τη μέγιστη από τις δύο τιμές.

2. Πληροφοριακό Όφελος και Τοπική Ανομοιομορφία

Η βασική ιδέα που ενέπνευσε αυτό το συνδυασμό μεθόδων είναι ότι *πέραν από το να επιλέγουμε παραδείγματα για τα οποία η ΜΔΥ είναι αβέβαιη, πρέπει ταυτόχρονα τα παραδείγματα που επιλέγονται σε κάθε επανάληψη (τα μέλη της κάθε «δέσμης») να είναι κατά το δυνατόν ανόμοια μεταξύ τους*. Τα υποψήφια ΔΕκ πρώτα διατάσσονται κατά φθίνουσα σειρά πληροφοριακού οφέλους. Στη συνέχεια, αρχίζουμε να γεμίζουμε τη δέσμη ξεκινώντας από τα υποψήφια ΔΕκ με το υψηλότερο πληροφοριακό όφελος. Αν, όμως, ένα υποψήφιο ΔΕκ μοιάζει πολύ με αυτά που έχουν ήδη εισαχθεί στη δέσμη, τότε απορρίπτεται και συνεχίζουμε με τα ΔΕκ χαμηλότερου πληροφοριακού οφέλους, μέχρι να γεμίσει η δέσμη.

Ως μέτρο ομοιότητας επιχειρήθηκε αρχικά να χρησιμοποιηθεί η συνάρτηση του *πυρήνα (kernel) K(* $\vec{v}_1$, $\vec{v}_2$*)* της ΜΔΥ. Η προσέγγιση, όμως, αυτή παρουσιάζει

δυσκολίες, διότι η συγκεκριμένη συνάρτηση χρησιμοποιεί ως παράγοντες παραμέτρους της ΜΔΥ [4],[1],[15] των οποίων οι τιμές αλλάζουν κατά τον επανασυντονισμό (retuning) της ΜΔΥ και είναι διαφορετικές για τις δύο ΜΔΥΧάριν απλότητας, προτιμήθηκε τελικά ως μέτρο ομοιότητας το *εσωτερικό γινόμενο (cosine similarity) cos(*$\vec{v}_1$*,* $\vec{v}_2$*)*, το οποίο απαιτεί μόνο τις ιδιότητες των δυο διανυσμάτων.

Ένα άλλο ερώτημα είναι ποιο θα είναι το κατώφλι ανομοιότητας, πέραν του οποίου θα απορρίπτονται από τη δέσμη τα υποψήφια ΔΕκ. Αρχικά χρησιμοποιήθηκε ως κατώφλι η *μέση ομοιότητα* μεταξύ όλων των ζευγών υποψηφίων διανυσμάτων εκπαίδευσης. Τελικώς, όμως, επιλέχθηκε μια τιμή μικρότερη από τη μέση (λιγότερο «αυστηρή»), διότι η προηγούμενη τιμή επέφερε απόρριψη πολύ μεγάλου αριθμού διανυσμάτων, με αποτέλεσμα να χρειάζεται συχνά να καταφεύγουμε στην πρώτη, απλούστερη μέθοδο επιλογής (πληροφοριακό όφελος μόνο) προκειμένου να γεμίζουμε τις δέσμες.

Καλούμε την απόρριψη υποψηφίων ΔΕκ που μοιάζουν πολύ με άλλα ΔΕκ της τρέχουσας δέσμης *«Τοπική Ανομοιομορφία» (local diversity)*, κατ' αντιδιαστολή προς το κριτήριο της *«Ολικής Ανομοιομορφίας» (global diversity)* που περιγράφεται σε επόμενη ενότητα.

3. Πληροφοριακό Όφελος, Αντιπροσωπευτικότητα και Τοπική Ανομοιομορφία

Η μέθοδος αυτή είναι ίδια με την προηγούμενη, με τη διαφορά ότι τα υποψήφια διανύσματα εκπαίδευσης διατάσσονται αρχικά κατά φθίνουσα σειρά μιας πιο σύνθετης βαθμολογίας, που προκύπτει από τον παρακάτω τύπο, αντί να διατάσσονται απλά κατά σειρά πληροφοριακού οφέλους. Η βαθμολογία (score) των διανυσμάτων είναι ένα συνδυασμός Πληροφοριακού Οφέλους και Αντιπροσωπευτικότητας. (Η τιμή λ που χρησιμοποιήσαμε ήταν η ίδια με εκείνη που χρησιμοποίησαν οι Shen κ.ά. [11].)

Ως ***Αντιπροσωπευτικότητα repres(*** $\vec{v}$***)*** *ενός υποψηφίου διανύσματος εκπαίδευσης* $\vec{v}$ *ορίζεται η μέση ομοιότητα του διανύσματος με όλα τα άλλα υποψήφια διανύσματα εκπαίδευσης.*

score = λ * info($\vec{v}$) + (1-λ) * repres($\vec{v}$),
λ = 0,6

(3.3-2) Τύπος υπολογισμού βαθμολογίας υποψηφίου ΔΕκ

Ο συνδυασμός αυτών των μεθόδων υποκρύπτει μία αντίθεση: *Από τη μία προτιμούμε διανύσματα εκπαίδευσης με μεγάλη Αντιπροσωπευτικότητα (δηλαδή διανύσματα που μοιάζουν με πολλά άλλα υποψήφια διανύσματα εκπαίδευσης) και από την άλλη απορρίπτουμε όσα διανύσματα είναι πολύ όμοια με τα υπόλοιπα υποψήφια ΔΕκ που έχουν ήδη προστεθεί στη δέσμη.* (Όπως και στην προηγούμενη μέθοδο, κατά το γέμισμα της δέσμης απορρίπτονται διανύσματα που μοιάζουν πολύ με άλλα διανύσματα που έχουν ήδη προστεθεί στη δέσμη, δηλαδή χρησιμοποιείται και το κριτήριο της Τοπικής Ανομοιομορφίας.) Το σκεπτικό είναι ότι θέλουμε μεν κάθε διάνυσμα της δέσμης να αντιπροσωπεύει (να μοιάζει με) πολλά άλλα υποψήφια διανύσματα εκπαίδευσης (να είναι μια συχνή περίπτωση), αλλά δεν θέλουμε όλα τα διανύσματα της δέσμης να αντιπροσωπεύουν την ίδια περίπου περίπτωση. Στις προκαταρκτικές δοκιμές ενεργητικής μάθησης που πραγματοποιήσαμε, όμως, παρατηρήσαμε ότι ο παραπάνω τύπος (score) προτιμούσε διανύσματα εκπαίδευσης που ήταν πολύ παρόμοια μεταξύ τους, τα οποία και απορρίπτονταν στη συνέχεια από το κριτήριο της Τοπικής Ανομοιομορφίας. Δηλαδή τα κριτήρια της Τοπικής Ανομοιομορφίας και της Αντιπροσωπευτικότητας συγκρούονταν. Πλέον αυτού, οι προκαταρκτικές δοκιμές ενεργητικής μάθησης που πραγματοποιήσαμε με το σύνθετο κριτήριο επιλογής παραδειγμάτων εκπαίδευσης αυτής της ενότητας οδήγησαν σε αποτελέσματα κατά πολύ χειρότερα ακόμα και από εκείνα της παθητικής μάθησης. Λόγω των παραπάνω, το κριτήριο επιλογής αυτής της ενότητας εγκαταλείφθηκε.

4. Πληροφοριακό όφελος, Τοπική Ανομοιομορφία και Ολική Ανομοιομορφία

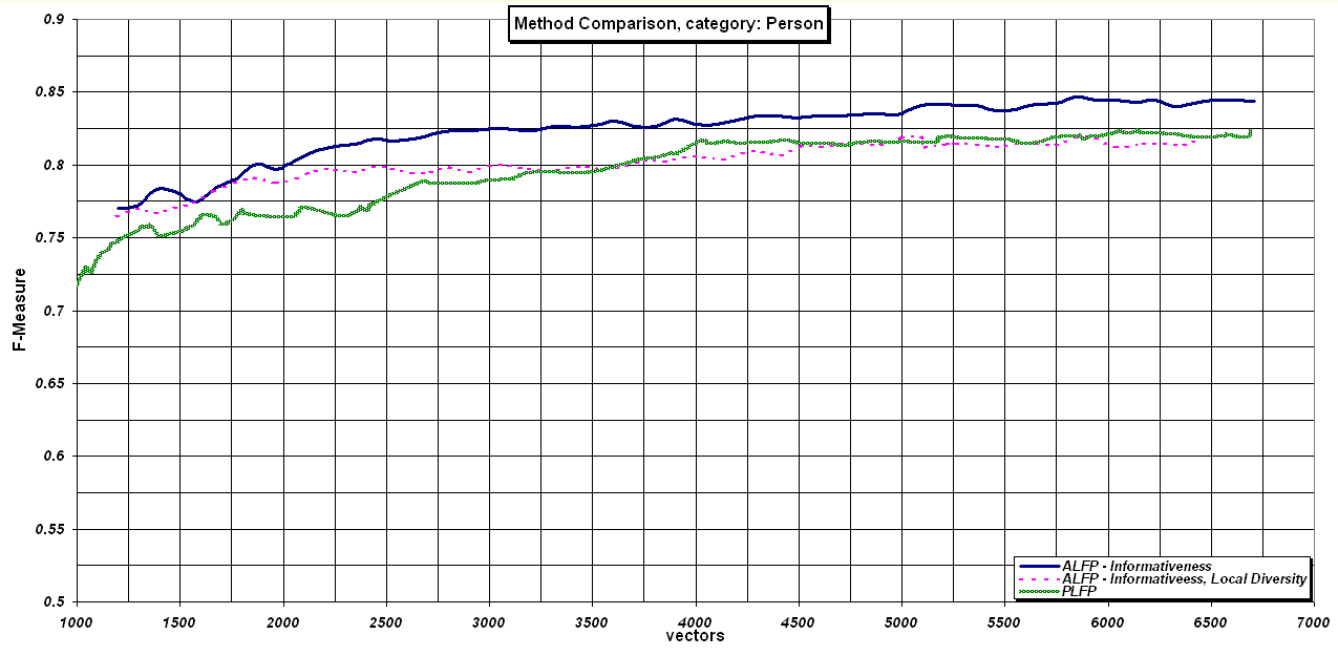
Στην περίπτωση αυτή, γεμίζαμε κάθε δέσμη όπως στο 2ο τρόπο, αλλά απορρίπταμε από τη δέσμη όχι μόνο τα υποψήφια ΔΕκ που έμοιαζαν πολύ με άλλα ΔΕκ της ίδιας δέσμης (τοπική ανομοιομορφία), αλλά και τα υποψήφια ΔΕκ που έμοιαζαν πολύ με άλλα ΔΕκ οποιασδήποτε προηγούμενης δέσμης (ολική ανομοιομορφία). Στις προκαταρκτικές δοκιμές ενεργητικής μάθησης, όμως, παρατηρήσαμε πως όσο και αν κάναμε πιο ελαστικό το κατώφλι ομοιότητας της Ολικής Ανομοιομορφίας, απορρίπτονταν σχεδόν όλα τα εναπομείναντα υποψήφια ΔΕκ, με αποτέλεσμα να αναγκαζόμαστε και πάλι να καταφεύγουμε συχνά σε απλούστερες μεθόδους επιλογής ΔΕκ προκειμένου να γεμίζουμε τις δέσμες. Λόγω αυτού του προβλήματος, και ο τρόπος επιλογής ΔΕκ αυτής της ενότητας εγκαταλείφθηκε.

Τα επόμενα διαγράμματα (3.3-1) και (3.3-2) παρουσιάζουν ορισμένα από τα αποτελέσματα των προκαταρκτικών δοκιμών, τα οποία δείχνουν ότι ο 1ος τρόπος επιλογής ΔΕκ (πληροφοριακό όφελος μόνο), αποφέρει καλύτερα αποτελέσματα από το 2ο (πληροφοριακό όφελος και τοπική ανομοιομορφία). Το F-measure είναι ένας συνδυασμός ακρίβειας (precision) και ανάκλησης (recall), που ορίζεται ως εξής:

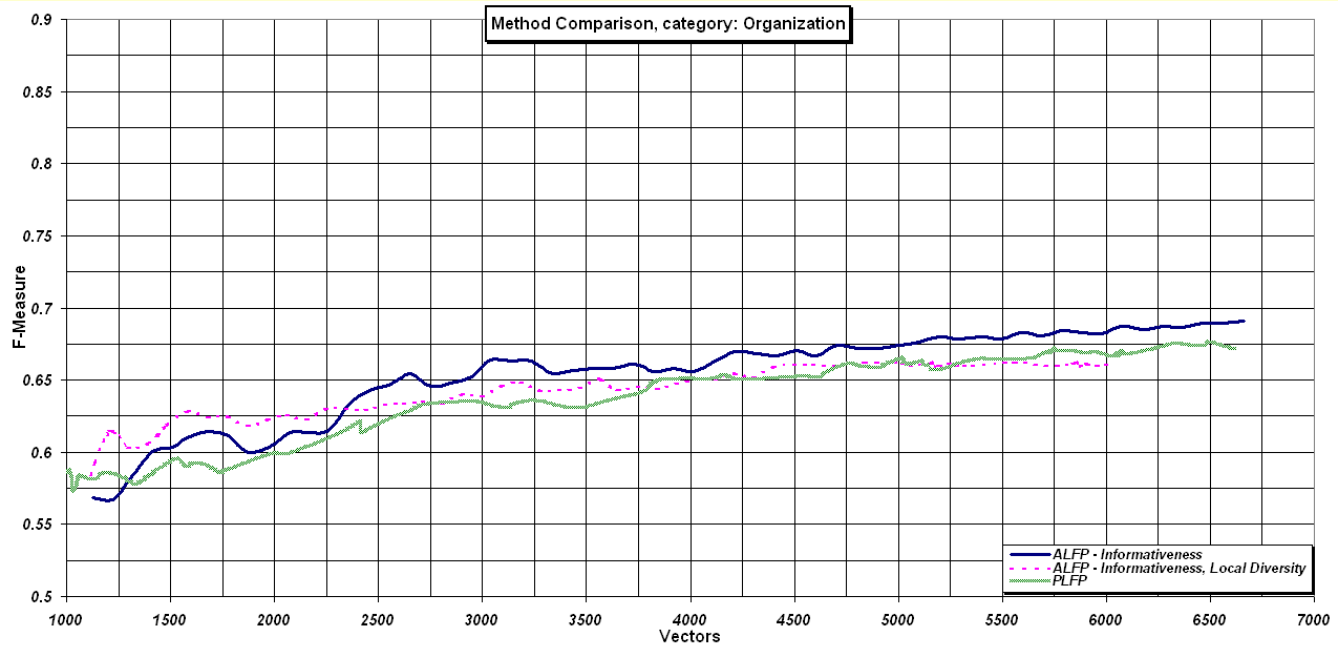
$$\text{F-measure} = \frac{(\beta^2 + 1) \cdot \text{ακρίβεια} \cdot \text{ανάκληση}}{\beta^2 \cdot \text{ακρίβεια} + \text{ανάκληση}}$$

Οφείλουμε να αναφέρουμε ότι στις προκαταρκτικές αυτές δοκιμές η «δεξαμενή» υποψηφίων ΔΕκ από την οποία επιλέγονταν τα παραδείγματα εκπαίδευσης ήταν μικρότερη από ό,τι στα πειράματα του επόμενου κεφαλαίου. Τα διανύσματα ελέγχου (ΔΕλ) ήταν επίσης λιγότερα. (Η δεξαμενή των προκαταρκτικών δοκιμών περιείχε περίπου 6680 υποψήφια ΔΕκ και τα ΔΕλ ήταν περίπου 81500.) Παρ' όλα αυτά δεν υπήρξε στις προκαταρκτικές δοκιμές καμία ένδειξη ότι η προσθήκη του κριτηρίου της τοπικής ανομοιομορφίας ενδέχεται να βελτιώνει τα αποτελέσματα σε σχέση με τη χρήση μόνο του κριτηρίου του πληροφοριακού οφέλους. Επίσης, όπως προαναφέρθηκε, οι άλλες δύο μέθοδοι επιλογής ΔΕκ, που χρησιμοποιούν επιπλέον τα κριτήρια της αντιπροσωπευτικότητας και της ολικής ανομοιομορφίας, οδήγησαν σε ακόμη χειρότερα αποτελέσματα, ενώ έχουν και το πρόσθετο μειονέκτημα ότι αυξάνουν το υπολογιστικό κόστος της επιλογής ΔΕκ, λόγω των επιπλέον συγκρίσεων διανυσμάτων που απαιτούν.

Βάσει των αποτελεσμάτων των προκαταρκτικών δοκιμών, στα πειράματα ενεργητικής μάθησης του επόμενου κεφαλαίου επιλέγουμε ΔΕκ βάσει του κριτηρίου του πληροφοριακού οφέλους και μόνο.



(3.3-3.1) Διάγραμμα Σύγκρισης f-measure μεθόδων AL
(Κατηγορία: **Πρόσωπα**)



(3.3-3.2) Διάγραμμα Σύγκρισης f-measure μεθόδων AL
(Κατηγορία: **Οργανισμοί**)

4. Διεξαγωγή Πειραμάτων

4.1 Συλλογές κειμένων

Οι συλλογές των κειμένων που χρησιμοποιήθηκαν στα πειράματα προέρχονται από τον **ημερήσιο τύπο** και πιο συγκεκριμένα από άρθρα των εφημερίδων « Βήμα» και «Τα Νέα». Και οι δύο συλλογές κειμένων περιλαμβάνουν άρθρα πολιτικού και κοινωνικού περιεχομένου, οικονομικά άρθρα, άρθρα σχετικά με πολιτισμικές εκδηλώσεις, αθλητικά άρθρα κ.ά. Όπως προαναφέρθηκε, αυτή η ποικιλία κάνει δυσκολότερο το πρόβλημα της αναγνώρισης και κατάταξης ονομάτων.

Στη διάρκεια της εργασίας του Λουκαρέλλι είχαν επισημειωθεί τα ονόματα προσώπων και οργανισμών 400 άρθρων των δύο συλλογών, 200 άρθρα από «Το Βήμα» και 200 από «Τα Νέα» (βλ. [1] για πληροφορίες σχετικές με τον τρόπο επισημείωσης). Το μέσο μέγεθος κάθε άρθρου είναι περίπου 6,3ΚΒ. Τα επισημειωμένα αυτά άρθρα διαχωρίστηκαν σε δύο τμήματα. Το ένα τμήμα χρησιμοποιήθηκε για την εκπαίδευση του συστήματος κατά την παθητική μάθηση, την αξιολόγηση των ιδιοτήτων και το συντονισμό των ΜΔΥ (συλλογή εκπαίδευσης, 198 άρθρα), και το άλλο για τον έλεγχο της απόδοσης του συστήματος (συλλογή ελέγχου, 202 άρθρα).

Τα υπόλοιπα (μη επισημειωμένα) άρθρα αποτέλεσαν τη «δεξαμενή» των μη επισημειωμένων παραδειγμάτων της AL. Η δεξαμενή χωρίστηκε σε 10, περίπου ίσα μεταξύ τους, μέρη. Σε κάθε επανάληψη της AL επιλέγονταν ΔΕκ από το επόμενο (κυκλικά) μέρος, ώστε να μειωθεί το υπολογιστικό κόστος της αξιολόγησης των υποψηφίων ΔΕκ (βλ. [1] για περισσότερες λεπτομέρειες). Το συνολικό μέγεθος της «δεξαμενής» κειμένων είναι 5000 άρθρα.

4.2 Συντονισμός / Επιλογή ιδιοτήτων

Ο συντονισμός (tuning) των παραμέτρων της κάθε ΜΔΥ αποτελεί μια διαδικασία βάσει της οποίας επιλέγονται οι τιμές των παραμέτρων C και γ της ΜΔΥ. Η παράμετρος γ είναι παράμετρος του πυρήνα RBF που χρησιμοποιούμε, ενώ η C ορίζει τη συνολική βαρύτητα που αποδίδεται στα παραδείγματα εκπαίδευσης που δεν καταφέρνει η ΜΔΥ να κατατάξει σωστά και εκτός του περιθωρίου (βλ. [1] για περισσότερες λεπτομέρειες). Για το συντονισμό των παραμέτρων των ΜΔΥ χρησιμοποιήσαμε τη σχετική μέθοδο αναζήτησης πλέγματος (grid search) και 5-πλής διασταυρωμένης επικύρωσης (5-fold cross-validation) που παρέχει η υλοποίηση LibSVM [4]. Για το συντονισμό της ΜΔΥ των προσώπων χρησιμοποιήθηκαν 1100 διανύσματα προσώπων και μη-προσώπων, ενώ για το συντονισμό της ΜΔΥ των οργανισμών χρησιμοποιήθηκαν 1060 διανύσματα οργανισμών και μη-οργανισμών. Όλα αυτά τα διανύσματα προέκυψαν από 14 κείμενα του συνόλου εκπαίδευσης, εξαιρώντας λεκτικές μονάδες που περιλαμβάνονταν στις λίστες τετριμμένων λέξεων και λεκτικές μονάδες που τις κατέτασσαν ως μη-πρόσωπα ή μη-οργανισμούς, αντίστοιχα, οι ασφαλείς κανόνες.

Η *επιλογή των ιδιοτήτων* γίνεται με κριτήριο τη μεγιστοποίηση του F-measure των ταξινομητών. Οι ταξινομητές εκπαιδεύονται πάνω σε διανύσματα ιδιοτήτων αποτελούμενα από υποσύνολα του πλήρους συνόλου των διαθέσιμων ιδιοτήτων και επιλέγεται το υποσύνολο που οδήγησε στα καλύτερα αποτελέσματα. Πιο συγκεκριμένα, για κάθε κατηγορία ονομάτων, ταξινομούνται όλες οι διαθέσιμες ιδιότητες (βλ. πίνακες 3.2-1.1 και 3.2-1.2) κατά φθίνουσα σειρά πληροφοριακού κέρδος (IG).[2] Στη συνέχεια δημιουργούνται υποσύνολα ιδιοτήτων, όπου το κάθε υποσύνολο προκύπτει από το προηγούμενο με αφαίρεση των n χειρότερων (βάσει IG) ιδιοτήτων και κρατείται τελικά το υποσύνολο για το οποίο η ΜΔΥ της συγκεκριμένης κατηγορίας ονομάτων είχε τα καλύτερα αποτελέσματα.

Στη διάρκεια της επιλογής ιδιοτήτων χρησιμοποιήθηκε μόνο το σύνολο των κειμένων εκπαίδευσης (βλ. ενότητα 4.1) και όχι το σύνολο ελέγχου, ώστε η επιλογή ιδιοτήτων να μην επηρεάζεται από το σύνολο ελέγχου. Πιο συγκεκριμένα, το σύνολο εκπαίδευσης χωρίστηκε σε δυο μέρη:

(set_α) τα κείμενα αυτού του υποσυνόλου (103 κείμενα) χρησιμοποιήθηκαν για την εκπαίδευση των ΜΔΥ με κάθε υποσύνολο ιδιοτήτων [3], και

(set_β) τα κείμενα αυτού του υποσυνόλου (99 κείμενα) χρησιμοποιήθηκαν για τον υπολογισμό του F-measure στη διάρκεια της επιλογής ιδιοτήτων.

---

[2] Τα πληροφοριακά κέρδη υπολογίστηκαν χρησιμοποιώντας το ελεύθερα διαθέσιμο σύστημα μηχανικής μάθησης WEKA [5].

[3] Αυτό το υποσύνολο περιείχε και τα κείμενα πάνω στα οποία έγινε ο συντονισμός (tuning).

4.3 Πρώτο πέρασμα

Στην ενότητα αυτή περιγράφονται τα πειράματα που διεξήχθησαν με ένα μόνο πέρασμα. Πρόκειται, δηλαδή, για πειράματα όπου χρησιμοποιούνται μόνο δύο ΜΔΥ, μία για τα ονόματα προσώπων και μία για τα ονόματα οργανισμών.

4.3.1. Συντονισμός παραμέτρων ΜΔΥ

Η διαδικασία συντονισμού της ενότητας 4.2 εκτελέστηκε δύο φορές πάνω στα αρχικά 1100 διανύσματα εκπαίδευσης των πειραμάτων παθητικής μάθησης που περιγράφονται στη συνέχεια. Τα αποτελέσματα της πρώτης εκτέλεσης χρησιμοποιήθηκαν ως είσοδος στη δεύτερη εκτέλεση, που είχε σκοπό να εντοπίσει με μεγαλύτερη ακρίβεια τις καλύτερες τιμές των παραμέτρων.

Στην πρώτη εκτέλεση χρησιμοποιήθηκαν όλοι οι συνδυασμοί παραμετρικών τιμών C από 2^0, 2^1, … έως 2^5, και γ από 2^{-5}, 2^{-4}, … έως 2^5. Η επιλογή του καλύτερου συνδυασμού έγινε βάσει αποτελεσμάτων 5-πλής διασταυρωμένης επικύρωσης.

Για της τιμές C = C_1 και γ = γ_1 που επεστράφησαν από τη πρώτη εκτέλεση, διερευνήθηκαν οι τιμές C' = $\log_2(C_1)$ και γ' = $\log_2(\gamma_1)$. Ο αλγόριθμος εκτελέστηκε και πάλι με C = $C^{C'-1}$, $C^{C'-0.75}$, … έως $C^{C'+1}$ και γ = $\gamma^{\gamma'-1}$, $\gamma^{\gamma'-0.75}$, … έως $\gamma^{\gamma'+1}$. Το βήμα μεταβολής στον εκθέτη του 2 ήταν 0,25 ώστε να υπολογιστεί με μεγαλύτερη ακρίβεια το βέλτιστο ζεύγος C και γ.

Οι τιμές των παραμέτρων που επελέγησαν από την παραπάνω διαδικασία περιλαμβάνονται στο πίνακα που έπεται:

	Πρόσωπα	Οργανισμοί
C	8 (2^3)	2.3784142300054 42 ($2^{1.2543}$)
γ	0.0078125 (2^{-7})	0.018581361171917516 ($2^{-5.76974}$)

(4.3.1-1) Παραμετρικές τιμές ΜΔΥ FP

4.3.2. Επιλογή ιδιοτήτων

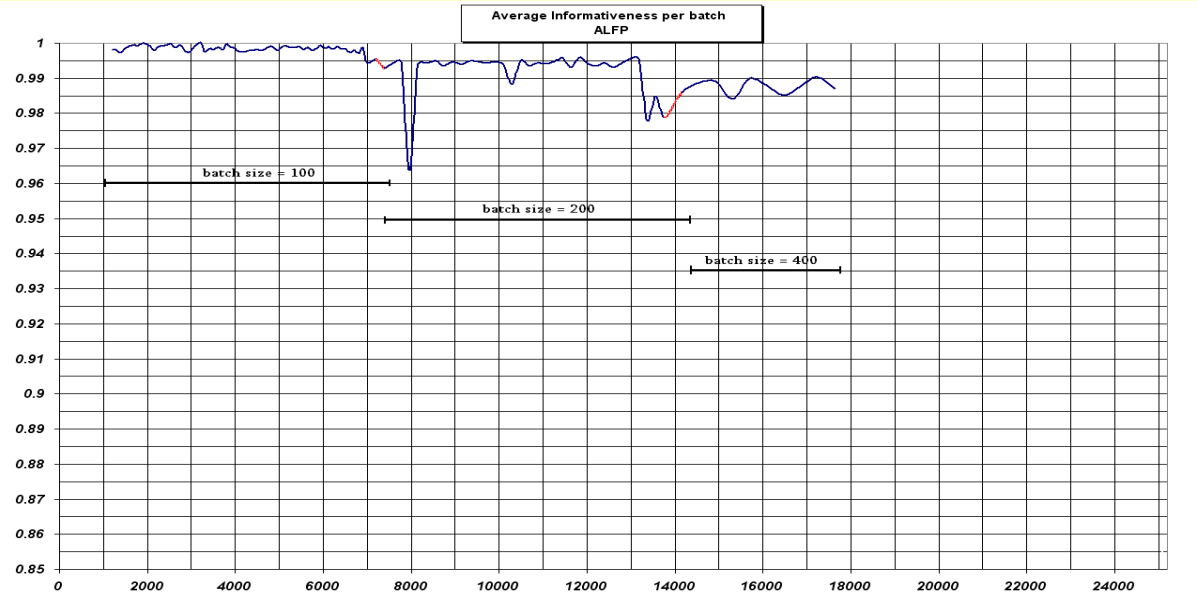
Έγιναν διαδοχικές επανεκπαιδεύσεις πάνω στο σύνολο δεδομένων set_α και έλεγχοι πάνω στο set_β με χρήση διαφορετικού υποσυνόλου ιδιοτήτων κάθε φορά. Μετά την πρώτη εκπαίδευση, όσες ιδιότητες παρουσίασαν μηδενικό IG σε κάθε κατηγορία ονομάτων οντοτήτων αφαιρέθηκαν άμεσα από τις υποψήφιες ιδιότητες αυτής της κατηγορίας. Τα υποψήφια υποσύνολα ιδιοτήτων της κάθε κατηγορίας προέκυψαν από την αφαίρεση των 20, 45, 60, 80 και 100 λιγότερο επωφελών βάσει IG ιδιοτήτων. Τελικώς κρατήθηκαν 170 ιδιότητες για τη κατηγορία των Οργανισμών και 165 ιδιότητες για εκείνη των Προσώπων.

4.3.3. Παθητική μάθηση

Στην περίπτωση της **παθητικής μάθησης (PL)** έγιναν διαδοχικά πειράματα με διαφορετικό κάθε φορά πλήθος ΔΕκ που προέρχονταν από τη συλλογή εκπαίδευσης της ενότητας 4.1. Σε κάθε πείραμα προσθέταμε στα δεδομένα εκπαίδευσης των ΜΔΥ τα αντίστοιχα ΔΕκ ενός επιπλέον χειρωνακτικά επισημειωμένου άρθρου της συλλογής εκπαίδευσης, εξαιρώντας διανύσματα που αντιστοιχούσαν σε τετριμμένες λέξεις ή λέξεις που τις κατέτασσαν ως μη-πρόσωπα ή μη-οργανισμούς οι αντίστοιχοι ασφαλείς κανόνες . Το σύστημα επανεκπαιδευόταν και επανελεγχόταν με σκοπό να μελετηθεί η βελτίωση των αποτελεσμάτων του.

4.3.4 Ενεργητική Μάθηση

Στην περίπτωση της **ενεργητικής μάθησης (AL)**, σε κάθε επανάληψη του πειράματος το ίδιο το σύστημα προτείνει στον άνθρωπο-εκπαιδευτή του τα παραδείγματα εκπαίδευσης που πρέπει να καταταγούν χειρωνακτικά και να προστεθούν στα δεδομένα εκπαίδευσης. Όπως προαναφέρθηκε, στα πειράματα ενεργητική μάθησης τα παραδείγματα εκπαίδευσης επιλέγονταν κατά δέσμες από μια δεξαμενή μη επισημειωμένων κειμένων χρησιμοποιώντας ως μέτρο επιλογής το πληροφοριακό όφελος. Οι δύο ΜΔΥ εκπαιδεύτηκαν αρχικά στα πρώτα περίπου 1100 διανύσματα εκπαίδευσης της παθητικής μάθησης.



(4.3.4-1) Μέσο Πληροφοριακό Όφελος Δεσμών FP

Η έντονη πτώση της καμπύλης στα περίπου 8000 και 13100 διανύσματα εκπαίδευσης αποδίδεται σε τοπικά φαινόμενο. Η καμπύλη επανέρχεται άμεσα σε συνήθη επίπεδα τιμών.

Οι δέσμες των διανυσμάτων είχαν αρχικά μέγεθος 100 έως τα 7200 ΔΕκ, μετά είχαν μέγεθος 200 έως τα 14200 ΔΕκ και κατόπιν μέγεθος 400. Αναγκαστήκαμε να μειώσουμε το μέγεθος των δεσμών διότι ο χρόνος εκτέλεσης των πειραμάτων είχε καταλήξει να είναι πολύ μεγάλος. (Μετά την επιλογή κάθε δέσμης, πρέπει να υπολογιστεί εκ νέου το πληροφοριακό όφελος των εναπομεινάντων υποψηφίων παραδειγμάτων εκπαίδευσης της δεξαμενής. Μεγαλώνοντας το μέγεθος των δεσμών, απαιτούνται λιγότεροι επανυπολογισμοί πληροφοριακού οφέλους.)Το παραπάνω διάγραμμα δείχνει το μέσο πληροφοριακό όφελος ανά δέσμη. Σημειώνονται με κόκκινο τα σημεία στα οποία μεγάλωσε το μέγεθος των δεσμών.

Όπως είναι φυσικό, η αύξηση του μεγέθους της δέσμης οδηγεί σε μικρή πτώση του μέσου πληροφοριακού οφέλους ανά δέσμη, αφού μετά την αύξηση οι δέσμες περιλαμβάνουν και διανύσματα για τα οποία το σύστημα είναι λιγότερο αβέβαιο. Τονίζεται ότι η «ορατότητα» της μείωσης του μέσου πληροφοριακού κέρδους οφείλεται εν πολλοίς και στη κλίμακα του διαγράμματος. Η πραγματική μείωση είναι πολύ μικρή, της τάξεως του 0.4% έως 0.6%.

Σε ότι αφορά τα διανύσματα που επέλεγε το σύστημα, παρατηρήθηκε καθ' όλη τη διάρκεια των πειραμάτων (ακόμα και κατά το SP) ότι το σύστημα είχε τη τάση να αμφιβάλλει για τις λεκτικές μονάδες που ξεκινούσαν με κεφαλαίο γράμμα. Τέτοια παραδείγματα είναι Αστικός Κώδικας, Κ.Ο.Κ., Η.Π.Α, Αρχιεπίσκοπος, Πρωθυπουργός, Τσάνπιονς Λίγκ κ.α. Το σύστημα επέμενε να ρωτά (με φθίνουσα συχνότητα μετά τα 12000 ΔΕκ) τα παραπάνω παραδείγματα παρότι είχε εκπαιδευτεί σε πολλά παρόμοια διανύσματα. Ίσως το παραπάνω να προδίδει τη δυσκολία του να καταταγούν ορισμένα από τα διανύσματα αυτά σε αποκλειστικά μία μόνο κατηγορία ονομάτων οντοτήτων, ακόμα και από τον άνθρωπο.

4.4 Δεύτερο Πέρασμα

Το στοιχείο που διαφοροποιεί το δεύτερο πέρασμα (SP) από το πρώτο (FP) είναι οι επιπλέον ιδιότητες του SP, οι οποίες υπολογίζονται από τις αποφάσεις των ταξινομητών του FP. Με άλλα λόγια γίνεται εκμετάλλευση της γνώσης που αποκομίστηκε από το πρώτο πέρασμα.

Στα πειράματα SP οι ΜΔΥ του πρώτου περάσματος είναι ήδη εκπαιδευμένες και δεν τροποποιούνται πλέον. Οι ΜΔΥ του FP που χρησιμοποιήθηκαν ήταν εκπαιδευμένες σε περίπου 10900 ΔΕκ επιλεγμένα με ενεργητική μάθηση. Επελέγη αυτός ο αριθμός ΔΕκ για τις ΜΔΥ του FP, επειδή τα πειράματα FP έδειξαν ότι οδηγεί στον καλύτερο συνδυασμό αποτελεσμάτων ακρίβειας (precision) των δύο κατηγοριών ονομάτων (87,89% και 76.5% για πρόσωπα και οργανισμούς αντίστοιχα). Με το **SP γίνεται προσπάθεια κυρίως να βελτιωθεί η ανάκληση (recall)** του συστήματος.

4.4.1. Συντονισμός παραμέτρων ΜΔΥ

Ο συντονισμός (tuning) των παραμέτρων της κάθε ΜΔΥ έγινε ακολουθώντας την ίδια διαδικασία με εκείνη που περιγράφεται στο πρώτο πέρασμα. Χρειάστηκαν ωστόσο κάποιες προσαρμογές στα στάδια που ακολουθούν:

(α) Για την 5-πλή διασταυρωμένη επικύρωση επισημειώθηκαν και χρησιμοποιήθηκαν 12 εντελώς νέα κείμενα.

(β) Ο αλγόριθμος αναζήτησης των καλύτερων τιμών των παραμέτρων εκτελέστηκε τρεις φορές, επειδή σε αντίθεση με το πρώτο πέρασμα έπρεπε να εξεταστεί ένα μεγαλύτερο εύρος τιμών για τις παραμέτρους. Αρχικές προσπάθειες εκτέλεσης του αλγορίθμου ακριβώς όπως κατά το πρώτο πέρασμα (μόνο δύο επαναλήψεις), οδήγησαν σε παραμέτρους που μείωναν το F-measure των αποτελεσμάτων ελέγχου πάνω στο ανεξάρτητο σύνολο ελέγχου (set_β).

Για τη κατηγορία των προσώπων, στη πρώτη εκτέλεση υπολογίστηκαν όλοι οι συνδυασμοί παραμετρικών τιμών C από 2^{-5}, 2^{-3}, 2^{-1}, … έως 2^{5}, και γ από 2^{-10}, 2^{-5}, 2^{0}, … έως 2^{5}. Στα προηγούμενα φαίνεται ότι για κάθε τιμή το βήμα μεταβολής του εκθέτη με βάση το 2, ήταν ίσο με 2 και 5 αντίστοιχα. Το ζεύγος που επελέγη για τα πρόσωπα ήταν C = 32 (2^{5}) και γ = 0.009765625 (2^{-10}).

Στη δεύτερη εκτέλεση υπολογίστηκαν όλοι οι συνδυασμοί παραμετρικών τιμών C από 2^3, 2^4, … έως 2^7, γ από 2^{-15}, 2^{-14}, 2^{-13}, … έως 2^{-5} για τη κατηγορία των προσώπων και C από 2^3, 2^4, … έως 2^7, γ από $2^{-5.5}$, $2^{-4.5}$, $2^{-3.5}$, … έως $2^{4.5}$ για τη κατηγορία των οργανισμών αντίστοιχα. Το ζεύγος που επελέγη για τα πρόσωπα ήταν C = 16 (2^4) και γ = 0.00390625 (2^{-8}) ενώ για τους οργανισμούς C = 8 (2^3) και γ = 0,22 ($2^{-5.5}$).

Στη τρίτη εκτέλεση υπολογίστηκαν όλοι οι συνδυασμοί παραμετρικών τιμών C από 2^3, $2^{3.225}$, … έως 2^5, γ από 2^{-9}, $2^{-8.75}$, … έως 2^{-7} για τη κατηγορία των προσώπων και C από 2^2, $2^{2.25}$, … έως 2^4, γ από $2^{-6.5}$, $2^{-6.25}$, … έως $2^{4.5}$ για τη κατηγορία των οργανισμών αντίστοιχα.

Ανάλογη υπήρξε και η σειρά των εκτελέσεων (τρεις το πλήθος και αυτές) και για τη κατηγορία των οργανισμών, Οι τελικές τιμές που επελέγησαν από την παραπάνω διαδικασία περιλαμβάνονται στο πίνακα που έπεται:

	Πρόσωπα	Οργανισμοί
C	9.5136569200217 ($2^{3.25}$)	4.6751089 ($2^{2.225}$)
γ	0.00276213 ($2^{-8.5}$)	0.0078125 (2^{-7})

(4.4.1-1) Παραμετρικές τιμές ΜΔΥ SP

4.4.2. Επιλογή ιδιοτήτων

Κατά το SP έγινε χρήση όλων των ιδιοτήτων του FP και επιπλέον ιδιοτήτων, των οποίων οι τιμές υπολογίζονταν βάσει των αποφάσεων των ταξινομητών του FP. Αυτές οι επιπλέον ιδιότητες εξετάστηκαν ως προς το πληροφοριακό τους κέρδος πάνω σε διανύσματα δεδομένων των κειμένων του set_α .

Πιο συγκεκριμένα, από το set_α αφαιρέθηκαν τα κείμενα τα οποία είχαν χρησιμοποιηθεί για τον συντονισμό των ΜΔΥ του FP, διότι τα συγκεκριμένα κείμενα χρησιμοποιήθηκαν και για την αρχική εκπαίδευση των ταξινομητών πριν την

έναρξη της AL του FP. Στη θέση τους προστέθηκαν τα 12 νέα κείμενα[4] που αναφέρονται στο συντονισμό των ΜΔΥ κατά το SP.

Σε αντίθεση με το FP, όπου έγιναν διαδοχικές επανεκπαιδεύσεις πάνω στο σύνολο set_α και έλεγχοι στο set_β με διαφορετικά υποσύνολα ιδιοτήτων, κατά το SP κρατήθηκαν όλες οι επιπλέον ιδιότητες για τις οποίες το πληροφοριακό κέρδος ήταν μη μηδενικό, κυρίως επειδή είχαν όλες μεγαλύτερο πληροφοριακό κέρδος από τις ιδιότητες που χρησιμοποιήθηκαν στο πρώτο πέρασμα. Καταλήξαμε έτσι να χρησιμοποιήσουμε 217 (170 + 47) ιδιότητες για τη κατηγορία των οργανισμών και 207 (165 + 42) ιδιότητες για εκείνη των προσώπων.

4.4.3. Παθητική μάθηση

Κατά το SP δεν εκτελέστηκαν πειράματα παθητικής μάθησης, μια που τα αποτελέσματα του FP είχαν ήδη δείξει την ανωτερότητα της ενεργητικής μάθησης.

4.4.4. Ενεργητική Μάθηση, 1η Προσέγγιση

Στην περίπτωση αυτή (**Δεύτερο Πέρασμα 1η Προσέγγιση - AL SP(1)**), οι ΜΔΥ των προσώπων και των οργανισμών του SP εκπαιδεύτηκαν αρχικά στα 12 κείμενα που είχαν χρησιμοποιηθεί για το συντονισμό των παραμέτρων τους (περίπου 1200 διανύσματα εκπαίδευσης για τη ΜΔΥ προσώπων και περίπου 1110 για τη ΜΔΥ οργανισμών).

Οι δέσμες είχαν από την αρχή μέγεθος ίσο με 400.

4.4.5. Ενεργητική Μάθηση, 2η Προσέγγιση

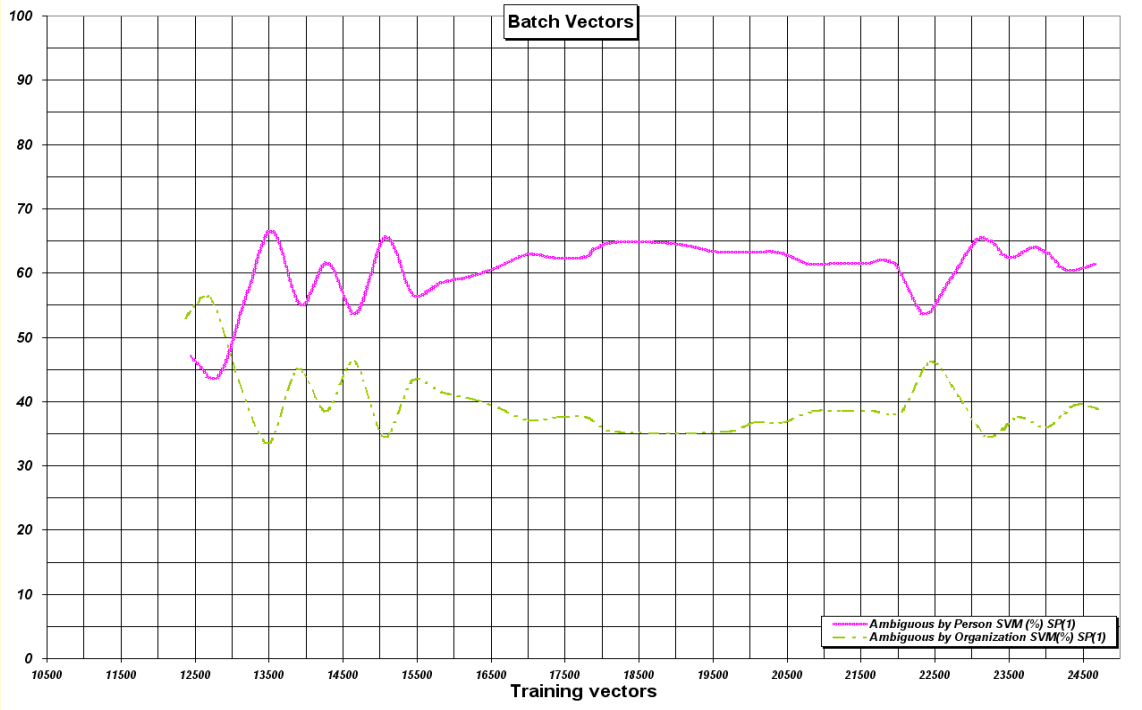
Στην περίπτωση αυτή (**Δεύτερο Πέρασμα 2η Προσέγγιση - AL SP(2)**), οι ΜΔΥ του SP εκπαιδεύτηκαν αρχικά σε όλα τα ΔΕκ που είχαν χρησιμοποιηθεί για την εκπαίδευση των ΜΔΥ του FP, αφού προηγουμένως αυτά επανα-υπολογίστηκαν, ώστε να περιέχουν και τις επιπλέον ιδιότητες του SP. Αυτό συνιστά το μοναδικό

[4] Τα ίδια 12 κείμενα όπου, όπως αναλύεται στο 4.4.5, χρησιμοποιήθηκαν για τη δημιουργία αρχικού ταξινομητή SP με το δεύτερο τρόπο (AL SP(2)).

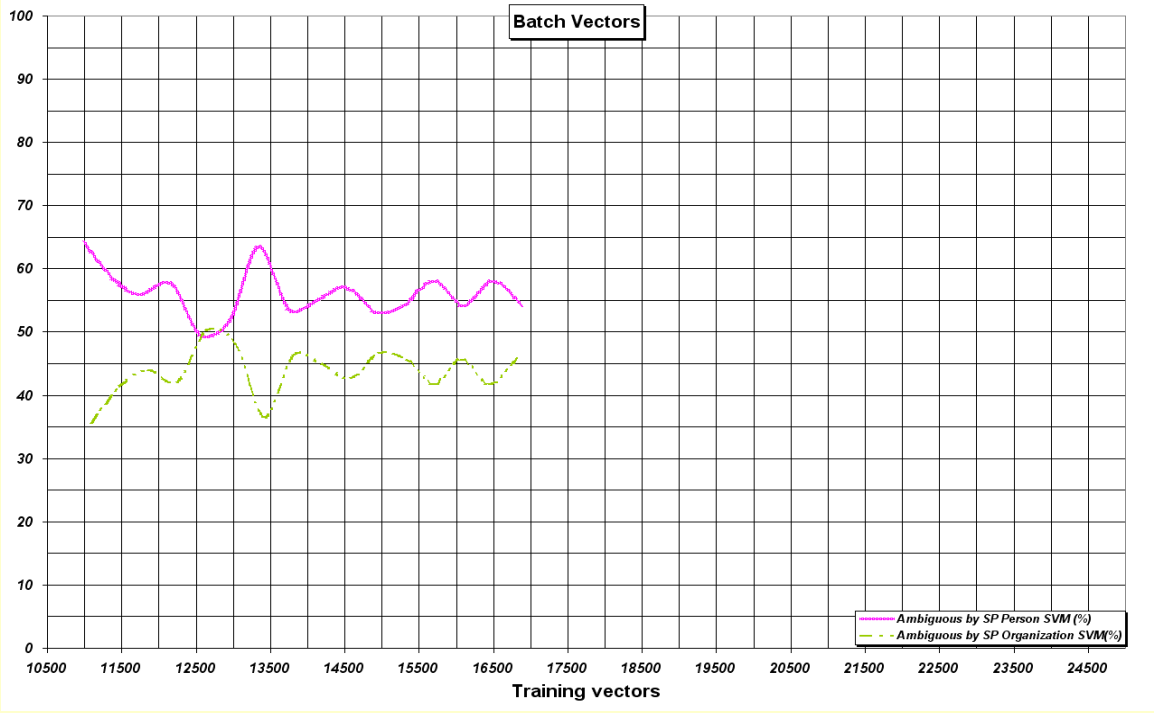
στοιχείο που διαφοροποιεί την AL SP(2) από την AL SP (1), καθότι τόσο οι παράμετροι των ΜΔΥ, όσο και τα σύνολα ιδιοτήτων ανά ΜΔΥ, παραμένουν ίδια. Στόχος αυτής της προσέγγισης είναι οι ΜΔΥ του SP να είναι εξαρχής εκπαιδευμένες σε πολλά ΔΕκ. Έτσι όμως ελλοχεύει ο κίνδυνος να ωθηθούν οι ΜΔΥ τους SP στο να εμπιστεύεται περισσότερο από ό,τι πραγματικά θα έπρεπε τις επιπλέον ιδιότητες του SP (που αντανακλούν τις αποφάσεις του FP): *οι ΜΔΥ του FP είχαν συναντήσει και κατά την εκπαίδευσή τους τα ΔΕκ στα οποία εκπαιδεύονται αρχικά οι ΜΔΥ του SP και ενδέχεται οι αποφάσεις τους για αυτά τα ΔΕκ να έχουν μεγαλύτερη αξιοπιστία από ό,τι θα είχαν οι αποφάσεις τους για εντελώς νέες περιπτώσεις.*

Ως ένα μέτρο για την αποφυγή του προηγούμενου κινδύνου αποφασίστηκε να χρησιμοποιηθούν οι ίδιες ιδιότητες με εκείνες που επιλέχθηκαν για την AL SP (1). Ένα δεύτερο μέτρο ήταν εκείνο της διατήρησης των ίδιων τιμών των παραμέτρων C και γ με εκείνες που είχαν υπολογιστεί για την AL SP (1).

Επιλογή διανυσμάτων δέσμης



(4.5-1.1) – AL SP(1)



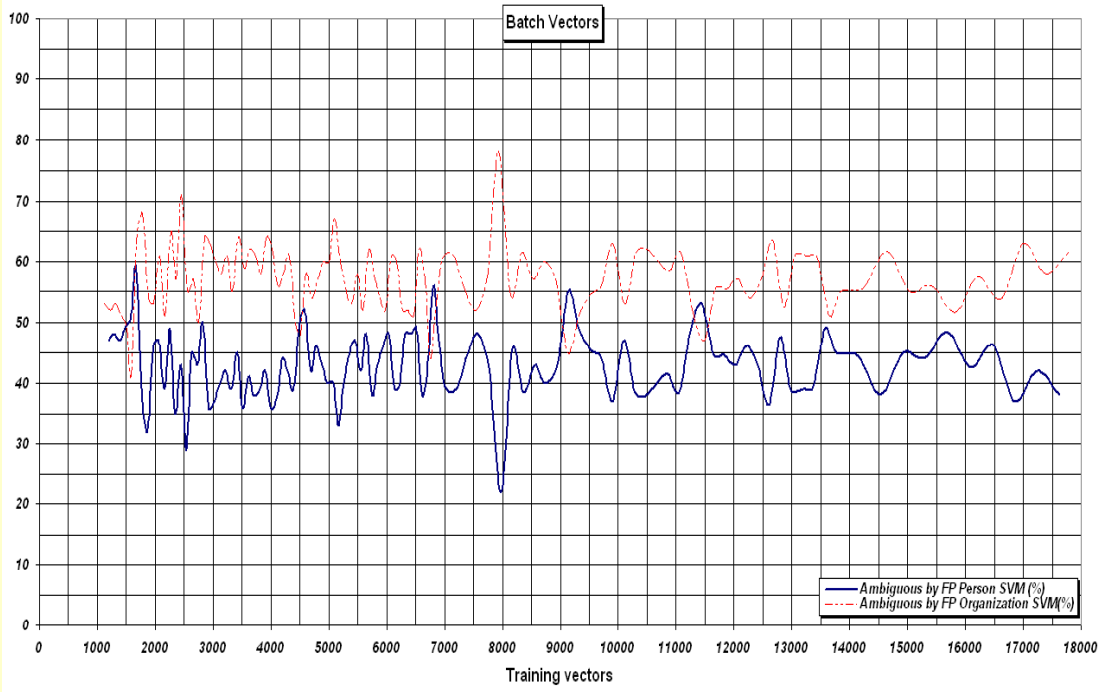
(4.5-1.2) – AL SP(2)

4.5 Σύγκριση και σχολιασμός

Συνοψίζοντας, εκτελέστηκαν πειράματα παθητικής και ενεργητικής μάθησης, με ένα και δύο περάσματα. Τα αποτελέσματα των πειραμάτων οδηγούν σε μια σειρά συμπερασμάτων ως προς τις επιδόσεις του συστήματος, τον καταβαλλόμενο κόπο χειρωνακτικής επισημείωσης παραδειγμάτων εκπαίδευσης και την ταχύτητα (εκπαίδευσης και κατάταξης) του συστήματος. Τα αποτελέσματα παρουσιάζονται υπό τη μορφή διαγραμμάτων στα οποία παρουσιάζονται ταυτόχρονα για οργανισμούς και πρόσωπα, οι καμπύλες f-measure, ανάκλησης και ακρίβειας. Επίσης παραθέτονται διαγράμματα χρόνου διεξαγωγής των πειραμάτων, καθώς και άλλα συμπληρωματικά διαγράμματα. Με βάση τα διαγράμματα αυτά, γίνεται σύγκριση των αποτελεσμάτων μεταξύ AL και PL, μεταξύ FP και SP, και τέλος ανάμεσα στη δυσκολία εκπαίδευσης του συστήματος στις κατηγορίες των οργανισμών και προσώπων.

Σε κάποια από τα διαγράμματα υπάρχουν ράβδοι λάθους (error bars) που δείχνουν τα διαστήματα εμπιστοσύνης με πιθανότητα 95%. Διαστήματα εμπιστοσύνης δεν έχουν επισημανθεί για το μέγεθος f-measure, καθότι δεν είναι ξεκάθαρο σε ποια ακριβώς τυχαία μεταβλητή αναφέρεται το μέγεθος αυτό. Επίσης, αποφύγαμε τα διαστήματα εμπιστοσύνης σε διαγράμματα που περιέχουν πολλές καμπύλες για λόγους ευκολίας κατανόησης των ήδη «πυκνών» αυτών διαγραμμάτων .

Ως γενικό συμπέρασμα αναφέρουμε οτι η κατηγορία των οργανισμών είναι σαφώς πιο δύσκολη σε σύγκριση με εκείνη των προσώπων. Το προηγούμενο συμπεραίνεται από το γεγονός ότι 1^{ον} επετεύχθη χειρότερο f-measure αποτελεσμάτων για τους οργανισμούς και 2^{ον} ο χρόνος εκπαίδευσης της ΜΔΥ των οργανισμών ήταν εμφανώς μεγαλύτερος. Η αυξημένη δυσκολία των οργανισμών ευθύνεται ενδεχομένως και στο γεγονός ότι κατά την ενεργητική μάθηση του πρώτου περάσματος επιλέγονται περισσότερα διανύσματα βάσει της απόστασής τους από το υπερεπίπεδο της ΜΔΥ των οργανισμών, παρά βάσει της απόστασής τους από το υπερεπίπεδο της ΜΔΥ των προσώπων, όπως φαίνεται στο διάγραμμα 4.5-2. (Υπενθυμίζεται ότι χρησιμοποιούμε τη μικρότερη από τις δύο αποστάσεις κατά την επιλογή ΔΕκ.)



(4.5-2) – Επιλογή διανυσμάτων δέσμης ν FP

Αντίθετα, όμως, στο SP (1η και 2η προσέγγιση) η επιλογή γίνεται κυρίως βάσει της απόστασης από το υπερεπίπεδο της ΜΔΥ των προσώπων. Ίσως αυτός να είναι και ο λόγος για τον οποίο η βελτίωση με το SP των αποτελεσμάτων για τα πρόσωπα είναι υπαρκτή αν και μικρή, σε αντίθεση με τους οργανισμούς, όπου τα αποτελέσματα είναι χειρότερα (βλ διαγράμματα 4.5-6,7,8). Ενδέχεται οι επιπλέον ιδιότητες του SP να κάνουν τη ΜΔΥ των οργανισμών να εμφανίζεται εσφαλμένα περισσότερο βέβαιη για τις αποφάσεις της (μεγαλύτερες αποστάσεις από το υπερεπίπεδο), σε σύγκριση με εκείνη των προσώπων, αλλά οι χρονικοί περιορισμοί της εργασίας δεν επέτρεψαν την περαιτέρω διερεύνηση αυτής της υπόθεσης.

Τα αποτελέσματα του συστήματος επηρεάζονται σαφώς και από το γεγονός της ποικιλίας στη θεματολογία των κειμένων, όπως δείχνουν και τα πειράματα της εργασίας [1], [15]. Επίσης, υπάρχουν περιπτώσεις ονομάτων στα κείμενα που είναι ιδιαίτερα διφορούμενες, δηλαδή περιπτώσεις στις οποίες ακόμα και ένας άνθρωπος δεν μπορεί να αποφασίσει π.χ. αν πρόκειται για όνομα προσώπου, οργανισμού, τοποθεσίας κλπ. Σε αυτές θα πρέπει να προστεθούν και περιπτώσεις ορθογραφικών λαθών. Κατά συνέπεια, **τα λάθη που εμφανίζεται να κάνει το σύστημα δεν είναι πάντα βέβαιο ότι είναι πραγματικά λάθη.**

Ακολουθούν κάποιες **χαρακτηριστικά δύσκολες περιπτώσεις** που εμφανίστηκαν κατά την ενεργητική κυρίως μάθηση:

1) Παραδείγματα όπου είναι δύσκολη η διάκριση μεταξύ οργανισμού και τοποθεσίας:

   - ***ΟΑΚΑ,***
   - ***Βουλή,***
   - ***Θέατρο «Κάτια Δανδουλάκη»***

Είναι δύσκολο ακόμη και από τα συμφραζόμενα να κατανοηθεί το κατά πόσον πρόκειται για οργανισμό ή τοποθεσία. Επιπλέον, ακόμα και αν είναι δυνατόν για τον άνθρωπο να καταλάβει από τα συμφραζόμενα, το σύστημα «βλέπει» μόνο 3 λέξεις γύρω από κάθε λεκτική μονάδα.

2) Παραδείγματα φράσεων όπως:

   - *«Ψάχνουν έναν νέο* ***«Καραμανλή»*** *» ,* *(Πρόσωπο)*
   - *«Σε ρόλο «Αντί-****Τζόλε»*** *»* *(Πρόσωπο)*
   - *ο νέος* ***Αϊνστάιν!*** *(Πρόσωπο)*
   - *Ο* ***«Θρύλος»*** *(Τίποτα)*
   - *Το* ***«τριφύλλι»*** *(Τίποτα)*
   - *Το «Κίνημα Σούπερ Σταρ* ***Πάουλο*** *Σόουζα και ΣΙΑ»* *(Τίποτα)*

σημ: Εντός παρενθέσεων δίνονται οι κατηγορίες που είχαν σημειωθεί χειρωνακτικά. Στις περιπτώσεις αυτές δεν υπάρχει σαφής προσδιορισμός κάποιου προσώπου ή οργανισμού και ακόμα και ένας άνθρωπος δεν μπορεί να πει με σιγουριά αν πρόκειται για πρόσωπο, οργανισμό , ή τίποτα από τα δύο.

3) Παραδείγματα όπως:

   - *Πανεπιστήμιο του* ***Χάρβαρντ (Harvard University)***
   - *Πολεμικό Μουσείο της* ***Αθήνας***
   - *Κεντρική Τράπεζα των* ***ΗΠΑ***
   - *Κοινωνικό Φόρουμ της* ***Γένοβας*** *(GSF)*

Σε περιπτώσεις όπου είναι στο χρήστη γνωστό από πριν ότι το τοπωνύμιο αποτελεί μέρος του ονόματος ενός οργανισμού, ο χρήστης μπορεί να το επισημειώσει με σιγουριά. Στον αντίποδα βρίσκονται περιπτώσεις όπως

*«Οργανώσεις ΑΤΑC **Γαλλίας** και **Ιταλίας**»*, που είναι πολύ δύσκολο να καταλάβει κανείς εάν πρόκειται για τοπικούς προσδιορισμούς ή μέρος του πλήρους ονόματος του οργανισμού. Σε αυτές τις περιπτώσεις τηρήθηκε η τακτική του να επισημειώνονται τα τοπωνύμια ως μέρη των ονομάτων των οργανισμών.

4) *«η πλευρά Μητσοτάκη», «Ο Ν . Χαλκιόπουλος του «κλίματος Άκη» » .*

   Πρόκειται για αναφορά σε Πρόσωπο ή σε μια ομάδα ατόμων και άρα σε κάποιο Οργανισμό . Σε αντίστοιχες περιπτώσεις τα παραδείγματα θεωρήθηκαν Οργανισμοί.

5) ***«Ε.Τ.** ο εξωγήινος», «Ο ΕΤ αποτέλεσε τη πρώτη ιδιαίτερα επιτυχημένη…»* : *Πρόκειται για **πρόσωπο ή τίποτα** (τίτλος ταινίας); Ακόμα χειρότερες είναι οι περιπτώσεις όπου ούτε από τα συμφραζόμενα δεν είναι ευκρινές το σε τι αναφέρεται το «ΕΤ» . Επιλογή μας ήταν να επισημειώνουμε τέτοια παραδείγματα ως τίποτα.*

6) *Οι **«ΗΠΑ»*** και η ***«Σοβιετική Ένωση»*** δεν είναι οργανισμοί. Αντίθετα η «**Ε.Ε.»** είναι Οργανισμός και άρα επισημειώνονται διαφορετικά κάποιες πολύ συχνές περιπτώσεις που κατά τα άλλα είναι παρόμοιες μεταξύ τους.

7) *«ΤουςΤαλιμπάν», «Άρης», «ΠΑΝΑΘΗΝΑΪΚΟΣ», «ΑΘΗΝΑ 2004 ».*

Πρόκειται για περιπτώσεις τυπογραφικών λαθών που επισημειώθηκαν χειρωνακτικά σαν να ήταν τυπογραφικά ορθές. Π.χ. το «Α» στο «ΠΑΝ Α ΘΗΝΑΪΚΟΣ» επισημειώθηκε ως λεκτική μονάδα ονόματος οργανισμού.

8) Φεστιβάλ:

Υπάρχουν πολλές περιπτώσεις όπου αναφέρεται ο όρος «φεστιβάλ» ως χρονικό γεγονός (*«…κατά τη διάρκεια του 1ου **Φεστιβάλ** του Βερολίνου το…»*) και άρα ανήκει στην κατηγορία **τίποτα.** Υπάρχουν όμως και περιπτώσεις όπου θεωρήθηκε τμήμα ονόματος οργανισμού (ακόμα και για λεκτικές μονάδες τοποθεσίας): π.χ. *«Η διεύθυνση του Φεστιβάλ Κινηματογράφου **Θεσσαλονίκης**, προέβη σε ανακοίνωση…».*

9) **Ίντερνετ : «**πρόσβαση στο **Internet»** :

Πρόκειται για κάποιον οργανισμό ή απλώς τίποτα; Στα κείμενα επιλέξαμε να το χαρακτηρίζουμε ως **τίποτα**.

10) Πολλές φορές αναφέρεται η λεκτική μονάδα «**Εθνική**» ως συντομογραφία του
   - *«**Εθνική Ομάδα Ποδοσφαίρου**»* ή του
   - «**Εθνική Τράπεζα**».

   Όταν ο όρος «Εθνική» παρέπεμπε σε εθνική ομάδα, τότε επισημειώθηκε ως **οργανισμός** . Σε φράσεις όπως *«Η **Ελπίδες** μας ηττήθηκαν»*, έγινε πάλι επισημείωση της λεκτικής μονάδος ως **οργανισμός**, διότι πρόκειται για την «**Εθνική Ομάδα Νέων 19 έως 22 ετών**» γνωστή και ως **«Εθνική Ελπίδων».**

Ονόματα όπως «**Θρύλος**», το «**Τριφύλλι**» , «**Δικέφαλος** του Βορά» ή «**Βασίλισσα** της Μαδρίτης» επισημειώθηκαν ως τίποτα διότι αποτελούν ψευδώνυμα και όχι ονόματα ποδοσφαιρικών οργανισμών. Το ίδιο έγινε και με παρατσούκλια προσώπων όπως «ο «**Ξανθός»** .

11) Περιπτώσεις όπως:
   - «Υπόθεση **Κοσκωτά» (γεγονός),**
   - **«**επιτροπή **Γεωργακόπουλου»,**
   - **«**παρέα **Γεωργάτος** και ΣΙΑ**» (ομάδα ατόμων:** μεταφορικός λόγος ως εταιρία**),**
   - **«**Περιβάλλον **Μηχαλολιάκου**» (Οργανισμός), «υπόθεση **Οτσαλάν**»

αποτελούν παραδείγματα λεκτικών μονάδων όπου κάποιοι οργανισμοί ή γεγονότα αποκαλούνται με το όνομα του προσώπου που πρωταγωνιστεί ή προΐσταται αυτών αντίστοιχα. Παράλληλα, όμως, είναι συχνά δύσκολο να διακρίνει κανείς εάν πρόκειται για ευθεία αναφορά στο πρόσωπο αυτό ή τον οργανισμό ή το γεγονός που το πρόσωπο αντιπροσωπεύει. Αποφασίστηκε να επισημειώνεται ως πρόσωπα ή οργανισμοί κατά περίπτωση**.**

12) Περιπτώσεις όπου παραθέτονται **τμήματα πανεπιστημιακών σχολών και Τ.Ε.Ι. χωρισμένα με κόμματα** . Αυτές οι περιπτώσεις συναντώνται σε άρθρα που αφορούν θέματα πανελλαδικών εξετάσεων. Η λέξη «τμήματα» μπορεί να μην εμφανίζεται από την εφαρμογή[5], ανήκει.

[5] Υπάρχει μια πεδίο κειμένου (text field) στη φόρμα εκτέλεσης των πειραμάτων ενεργητικής μάθησης, όπου εμφανίζεται η «γειτονιά» των λέξεων της υπό εξέταση λεκτικής μονάδας. Αυτή η γειτονιά εκτείνεται σε ±16 λέξεις από αυτή για την οποία ερωτάται να κατατάξει ο χρήστης.

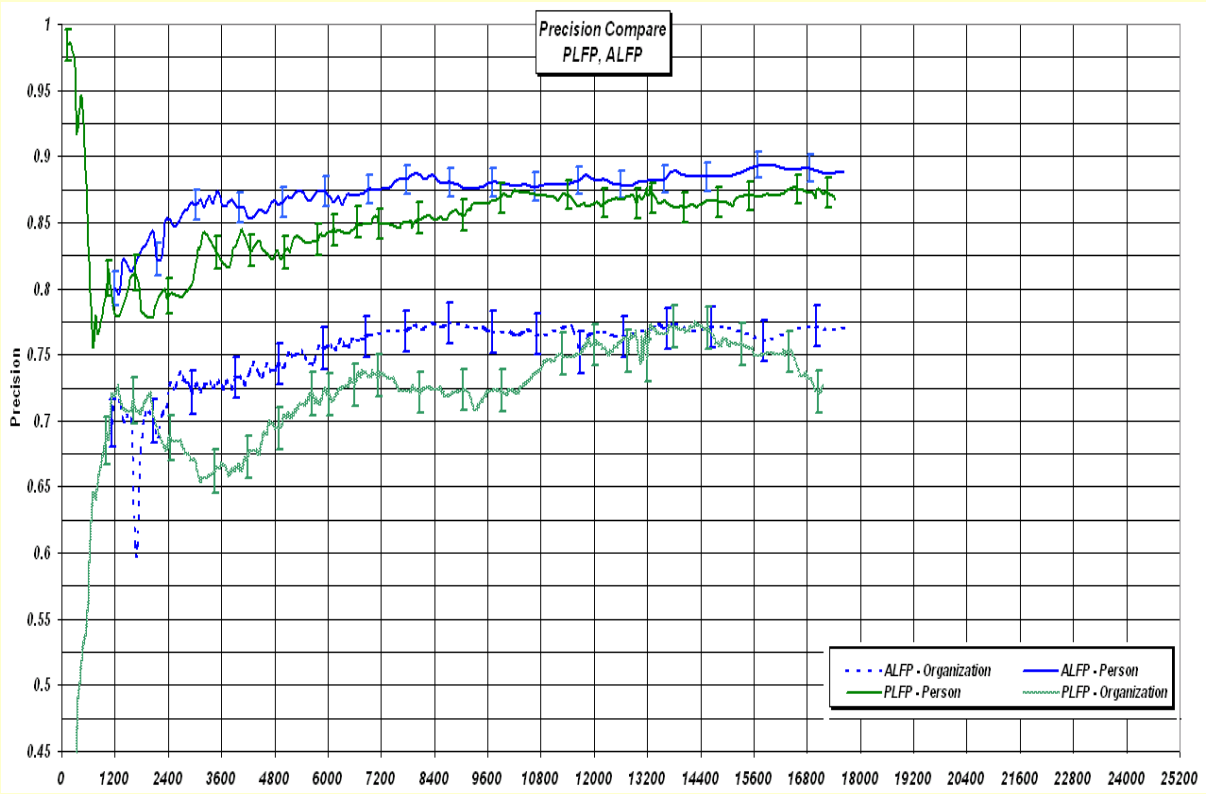
ωστόσο επισημειώθηκαν ως οργανισμοί διότι είναι γνωστό στον χρήστη ότι πρόκειται για οργανισμούς.

13) Ανάλογες περιπτώσεις συναντώνται κατά τη **παράθεση ονομάτων ενδεκάδων ποδοσφαιριστών ή άλλων ονομάτων** (π.χ. μελών ενός Διοικητικού Συμβουλίου ή τοπικών κομματικών οργανώσεων). Τονίζουμε και εδώ την αδυναμία του συστήματος να εξετάζει λεκτικές μονάδες σε απόσταση μεγαλύτερη των τριών από την τρέχουσα λεκτική μονάδα.

14) «…όπως είναι η νόσος του **Πάρκινσον** όπου τα βλαστικά κύτταρα…» .

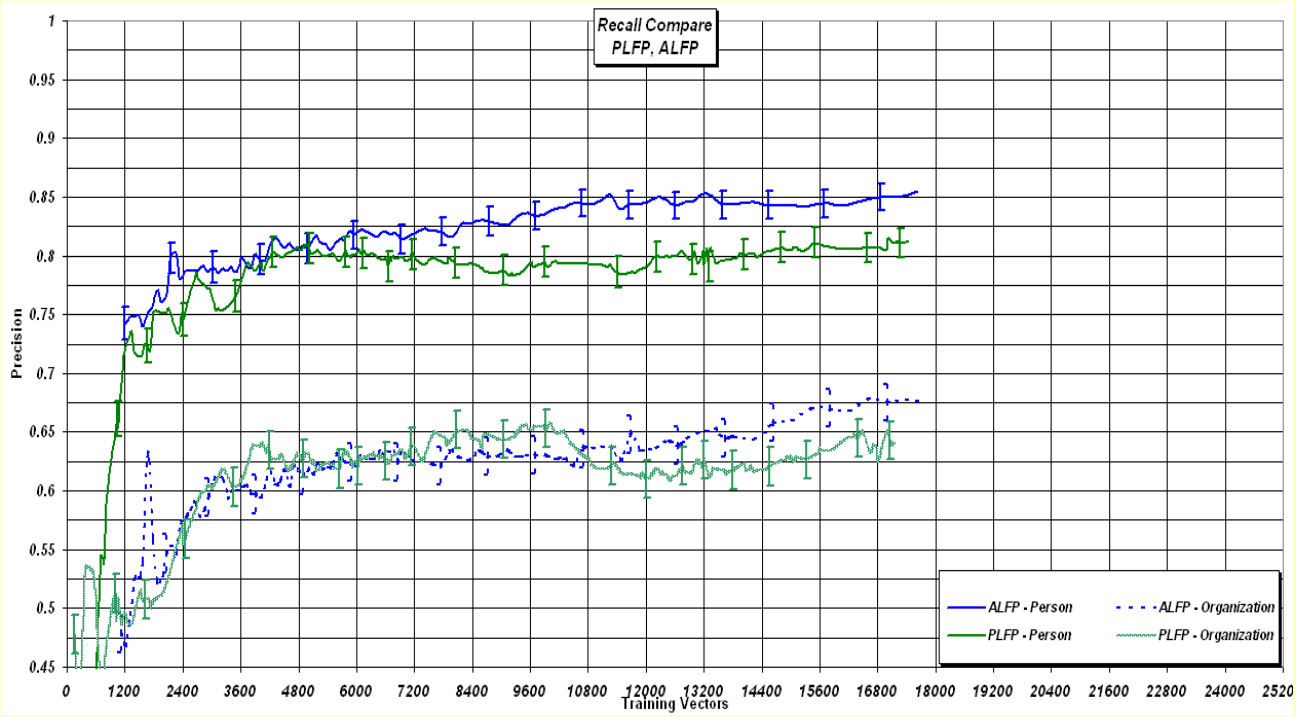
Περιπτώσεις σαν αυτές επισημειώθηκαν ως τίποτα, παρότι το «Πάρκινσον» είναι το όνομα του προσώπου που μελέτησε τη νόσο. Ωστόσο, όμως, η αναφορά του ονόματος του δεν έχει να κάνει με τον ίδιο, αλλά με τη νόσο.

4.5.1 Πρώτο πέρασμα



(4.5-3) Ακρίβεια FP (PLFP, ALFP)

Στο παραπάνω διάγραμμα είναι εμφανές ότι με την ενεργητική μάθηση (AL) επιτυγχάνεται γενικά υψηλότερη ακρίβεια (precision) από ό,τι με την παθητική, για ίδιο αριθμό διανυσμάτων εκπαίδευσης, παρ' όλο που υπάρχουν σε ορισμένα σημεία επικαλύψεις των διαστημάτων εμπιστοσύνης. Αν και η διαφορά μεταξύ AL και PL είναι μεγαλύτερη στην περίπτωση των οργανισμών, η ακρίβεια των οργανισμών στην AL παρουσιάζει εντονότερες διακυμάνσεις από ό,τι στην περίπτωση των προσώπων και πλησιάζει τελικά την ακρίβεια των οργανισμών της PL.



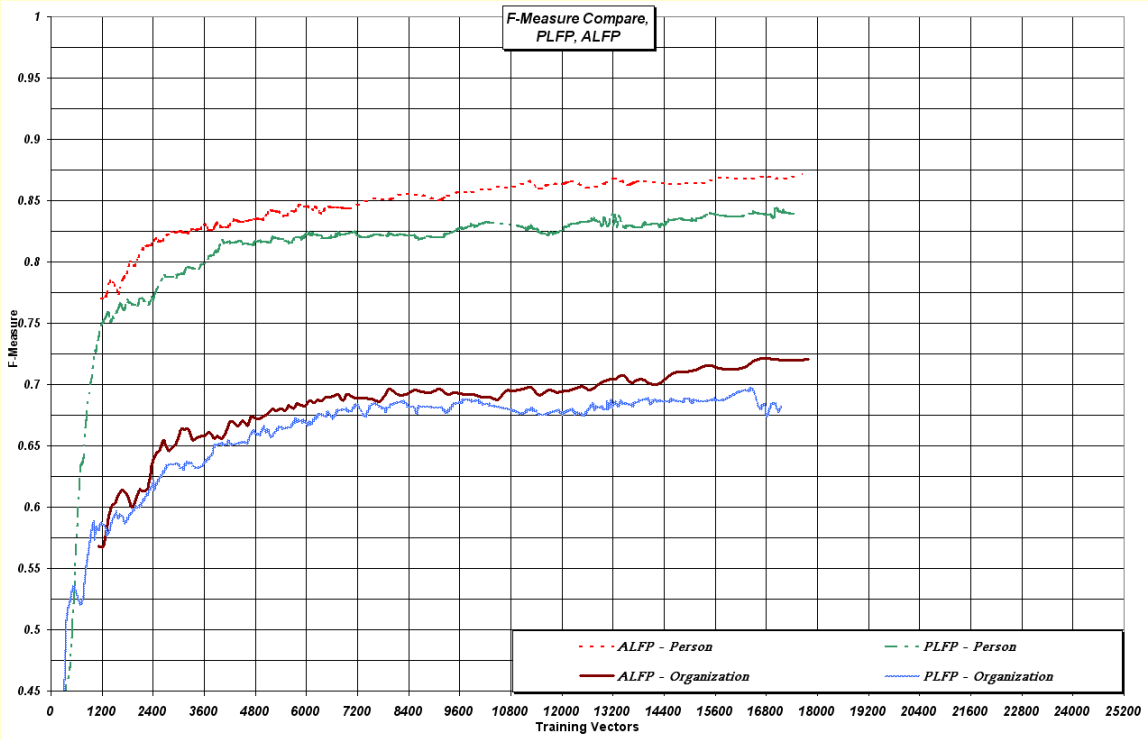
(4.5-4) Ανάκληση αποτελεσμάτων FP (PL FP, AL FP)

Στο διάγραμμα *(4.5-4)* φαίνεται ότι για τη κατηγορία των προσώπων η ανάκληση της AL είναι σαφώς υψηλότερη της ανάκλησης της PL. Αντίθετα, στην περίπτωση των οργανισμών χρειάζεται να φτάσουμε τα 10800 διανύσματα εκπαίδευσης για να παρουσιαστεί εμφανές προβάδισμα της AL.

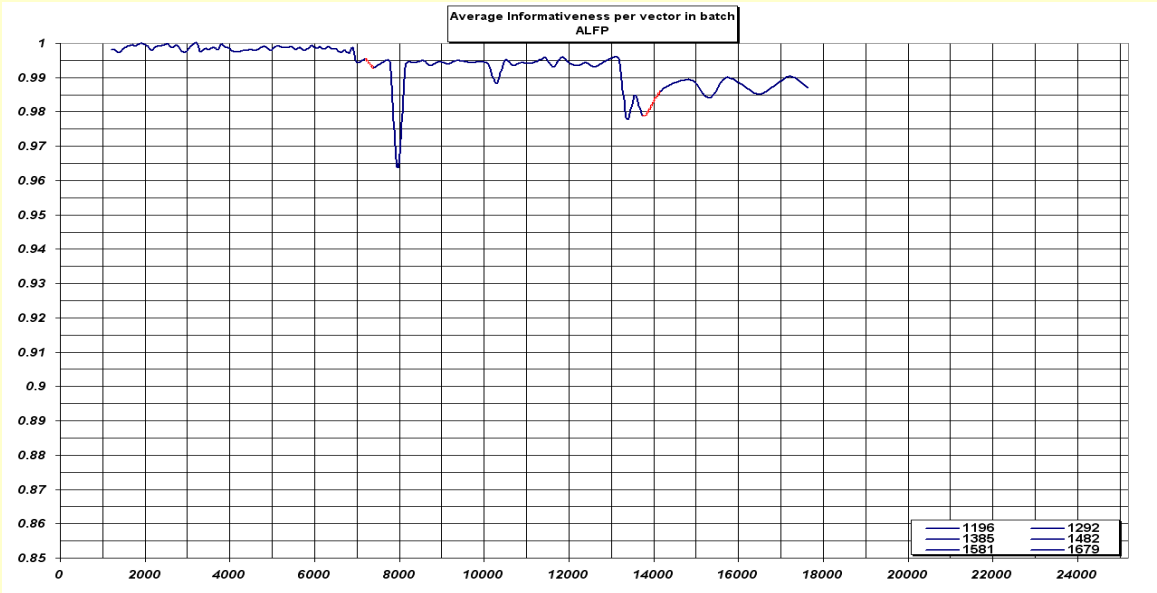
Επισημαίνουμε ότι *τα **μεγέθη της ανάκλησης και της ακρίβειας** έχουν γενικά την τάση να παρουσιάζουν μεταβολές προς την αντίθετη φορά (tradeoff).* Στόχος μας λοιπόν είναι να πετύχουμε το καλύτερο δυνατό συνδυασμό των δύο αυτών μεγεθών, βελτιώνοντας το ένα μέγεθος με όσο το δυνατόν μικρότερη μείωση για το άλλο. Στο ακόλουθο διάγραμμα f-measure (4.5-5) παρουσιάζονται πάλι μαζί τα αποτελέσματα του FP για PL και AL. Όπως είναι φανερό, η AL απέφερε από την αρχή και μόλις των πειραμάτων βελτίωση των αποτελεσμάτων σε σχέση με την PL και για τις δύο κατηγορίες.

Παρατηρούμε ότι γενικά με την AL μπορούμε να φτάσουμε στο ίδιο επίπεδο επιδόσεων με λιγότερα παραδείγματα εκπαίδευσης (και άρα λιγότερο κόπο χειρωνακτικής επισημείωσης) σε σχέση με την PL. Πλέον των προαναφερθέντων πλεονεκτημάτων της AL έναντι της PL, **ο χρήστης δε χρειάζεται να αναλωθεί στη μονότονη και ιδιαίτερα κοπιαστική επισημείωση ολόκληρων κειμένων**. Το μόνο που χρειάζεται να κάνει είναι να κατατάξει εκείνες τις λεκτικές μονάδες για τις οποίες

ερωτάται από το σύστημα. (Οι Γραφικές Διεπαφές Χρήστη που παρέχονται διευκολύνουν περαιτέρω τη χειρωνακτική επισημείωση κατά την AL.) Άρα εκτός της ποιοτικής ανωτερότητας των αποτελεσμάτων **διαπιστώνεται και οικονομία καταβαλλόμενου κόπου και χρόνου** από τον άνθρωπο-επισημειωτή.



(4.5-5) f-measure αποτελεσμάτων FP

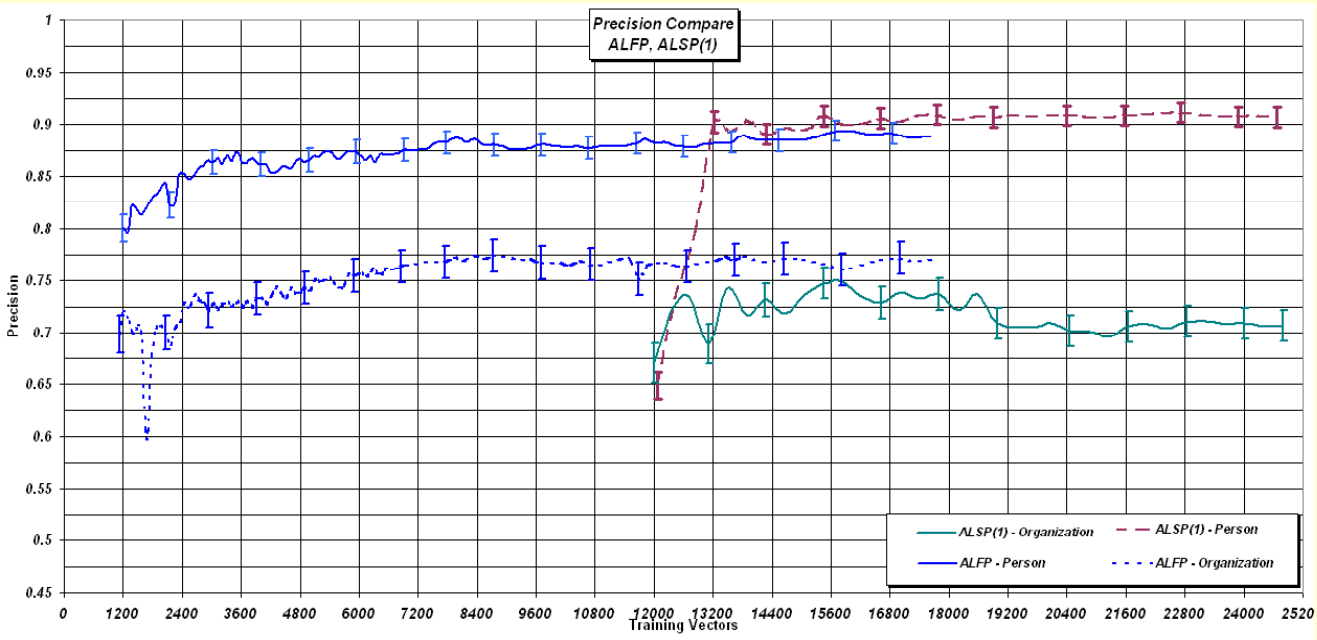


(4.5-6) Μέσο Πληροφοριακό Όφελος ανά δέσμη Πρώτου περάσματος

4.5.2 Δεύτερο Πέρασμα (SP)

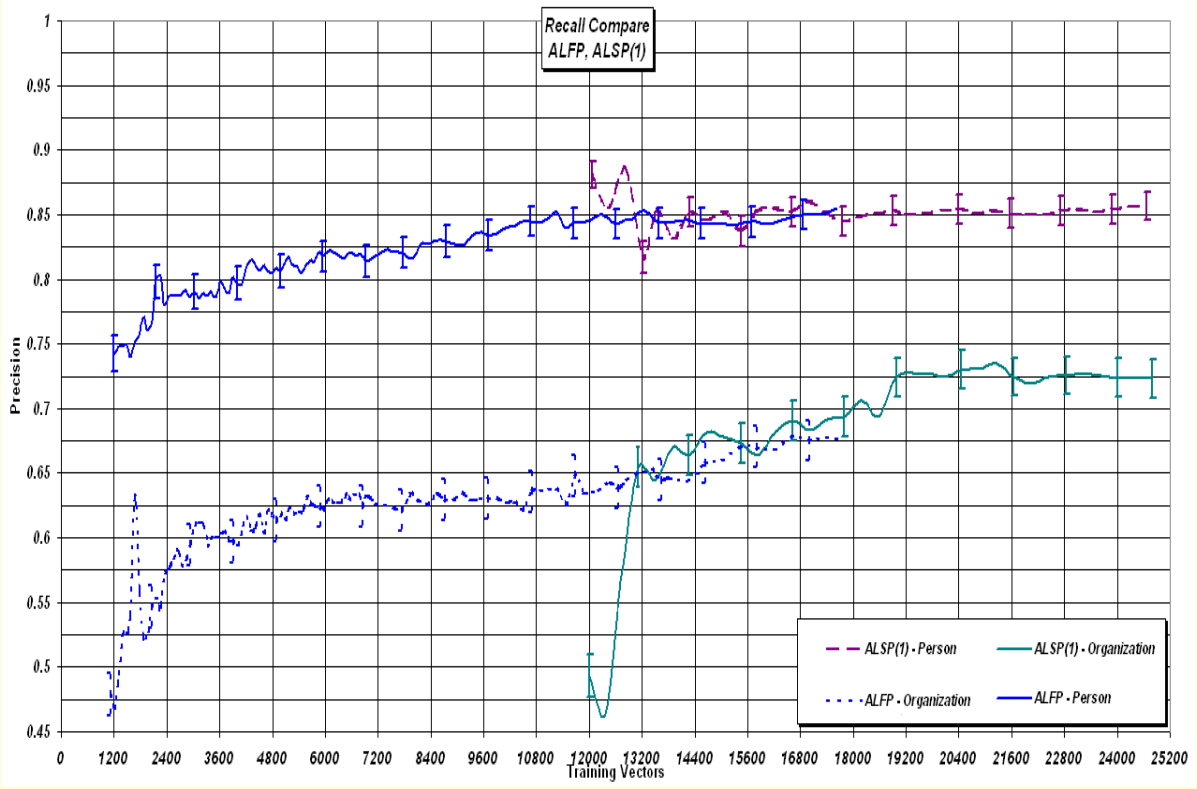
4.5.2.1 1η προσέγγιση

Ως γενική παρατήρηση προ του σχολιασμού των αποτελεσμάτων του Δεύτερου Περάσματος με τη 1η προσέγγιση εκτέλεσης (**AL SP(1)**), οφείλουμε να αναφέρουμε ότι στον οριζόντιο άξονα των διαγραμμάτων περιλαμβάνονται όλα τα διανύσματα εκπαίδευσης που έχουν επισημειωθεί χειρωνακτικά, δηλαδή ο οριζόντιος άξονας αντιστοιχεί στο **συνολικό κόπο που έχει καταβληθεί από το χρήστη**. Με άλλα λόγια στην περίπτωση του AL SP(1) στον οριζόντιο άξονα συνυπολογίζονται τα διανύσματα εκπαίδευσης των ΜΔΥ του FP (περίπου 10900), συν τα διανύσματα αρχικής εκπαίδευσης των ΜΔΥ του SP (περίπου 1100), συν τα διανύσματα για τα οποία ερωτάται από το σύστημα ο χρήστης κατά το SP.



(4.5-7.1) Σύγκριση ακρίβειας ALFP, AL SP (1)

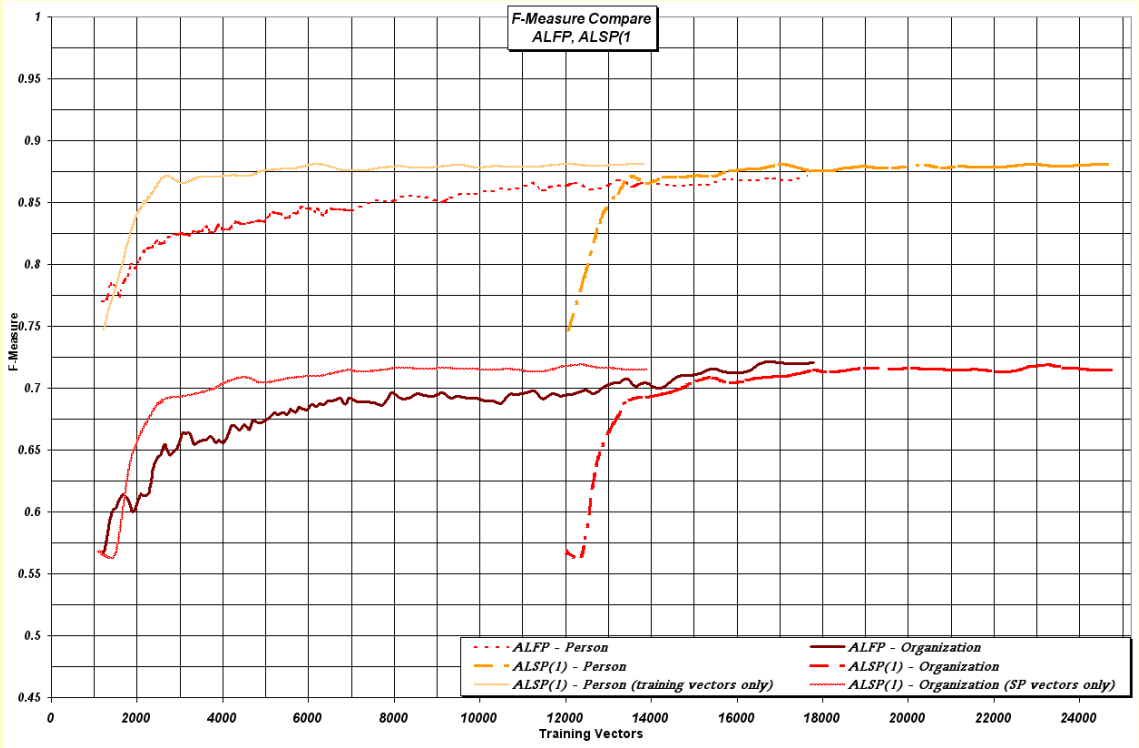
Το 4.5-7.1 δείχνει ότι στο SP η ακρίβεια των οργανισμών παρουσιάζει διακυμάνσεις και τελικά σταθεροποιείται στα ίδια περίπου επίπεδα όπου είχαμε καταλήξει με το FP. Αντίθετα, η ακρίβεια των προσώπων είναι σταθερότερη και -έστω και κατά ελάχιστο- βελτιωμένη σε σχέση με τα επίπεδα στα οποία είχαμε φτάσει με το FP.



(4.5-7.2) Διάγραμμα σύγκρισης ανάκλησης AL FP, AL SP (1)

Σε ό,τι αφορά τα πρόσωπα, η ανάκληση παραμένει σταθερή με κάποιες μικρές διακυμάνσεις σε σχέση με το πρώτο πέρασμα. Για τους οργανισμούς, ωστόσο, η ανάκληση έχει έντονα ανοδική τάση εις βάρος όμως της ακρίβειας.

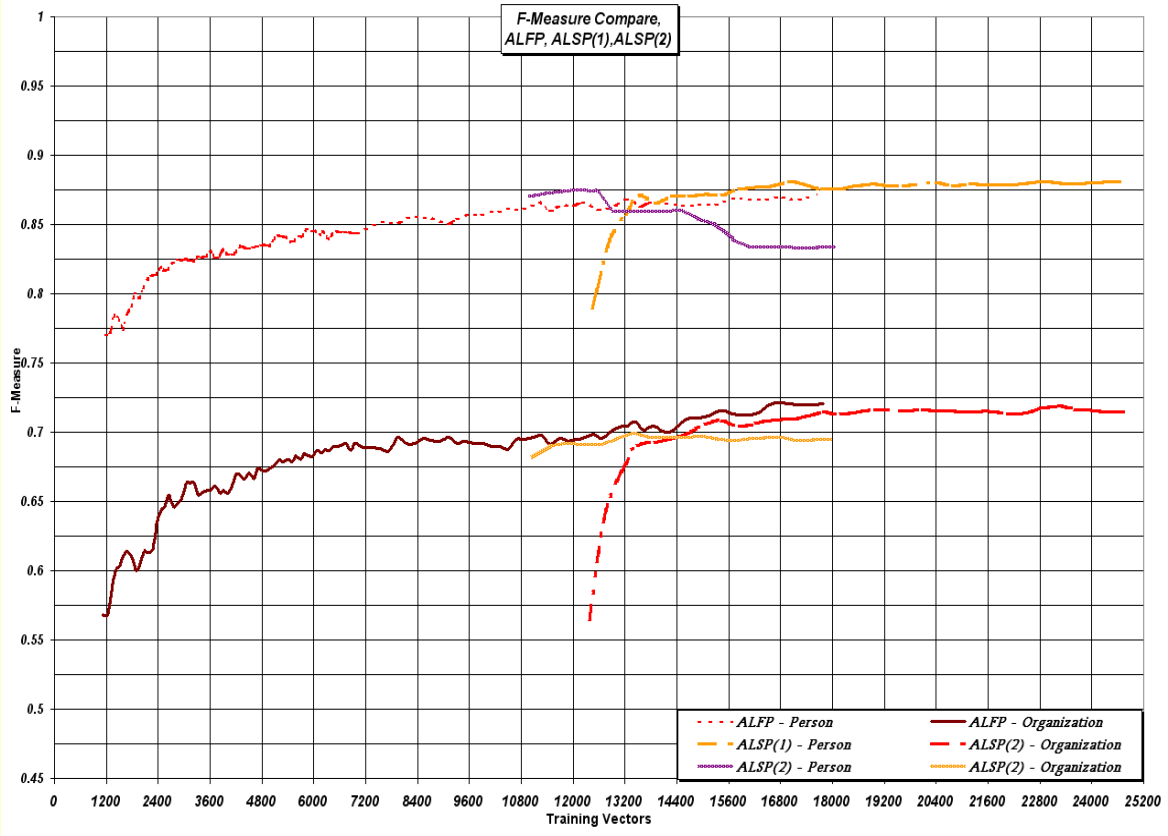
Πιο χαρακτηριστικό διάγραμμα από όλα αποτελεί το 4.5.-7.3, όπου και συνοψίζονται οι επιδόσεις του Δεύτερου Περάσματος (1η προσέγγιση). Στο διάγραμμα αυτό εμφανίζονται οι καμπύλες του Δεύτερου Περάσματος δύο φορές ανά κατηγορία. Τη πρώτη φορά μετρώνται στον οριζόντιο άξονα **μόνο τα ΔΕκ** της AL και στην άλλη μετρώνται τα **διανύσματα του συνολικού κόπου.**



(4.5-7.3) Σύγκριση αποτελεσμάτων f-measure AL FP, AL SP (1)

Αν περιορίσουμε τον σχολιασμό μας μόνο στις καμπύλες του f-measure συνολικού κόπου στο διάστημα [1200,18000], τότε θα δούμε ότι τα αποτελέσματα δεν επιφέρουν ουσιώδη βελτίωση σε σχέση με αυτά του FP. Για τους οργανισμούς μάλιστα τα αποτελέσματα είναι ελαφρώς χειρότερα. Στη πραγματικότητα η συμπεριφορά αυτή των καμπυλών έχει να κάνει με τον ***πλασματικό αριθμό των διανυσμάτων*** του οριζόντιου άξονα. Αντίθετα, αν παρατηρήσουμε τις καμπύλες με καταμετρημένα μόνο τα ΔΕκ της AL SP, τότε τα αποτελέσματα είναι εμφανώς βελτιωμένα. Ενδεικτικά αναφέρουμε ότι μόλις με 7200 διανύσματα τα αποτελέσματα είναι ίδια με εκείνα των 18000 διανυσμάτων του FP. Αυτό ισοδυναμεί με αναλογία ΔΕκ μόλις 1/2,5.

Όπως φαίνεται και στα διαγράμματα, τα πειράματα SP συνεχίστηκαν για περισσότερα παραδείγματα «συνολικού κόπου επισημείωσης» από ό,τι εκείνα του FP. Τονίζουμε ότι η συνέχιση των πειραμάτων με το FP μετά τα 18000 ΔΕκ θα ήταν πρακτικά ανέφικτη, επειδή η εκπαίδευση των ΜΔΥ του FP με τόσα πολλά διανύσματα εκπαίδευσης είναι εξαιρετικά χρονοβόρα. Υπενθυμίζουμε στον αναγνώστη ότι οι ΜΔΥ του SP εκπαιδεύονται στην πραγματικότητα σε 10900 + 1100 διανύσματα λιγότερα από τον αριθμό διανυσμάτων που αντιστοιχεί στο «συνολικό κόπο επισημείωσης».



(4.5-8) Σύγκριση αποτελεσμάτων f-measure
ALFP, AL SP (1) και AL SP (2)

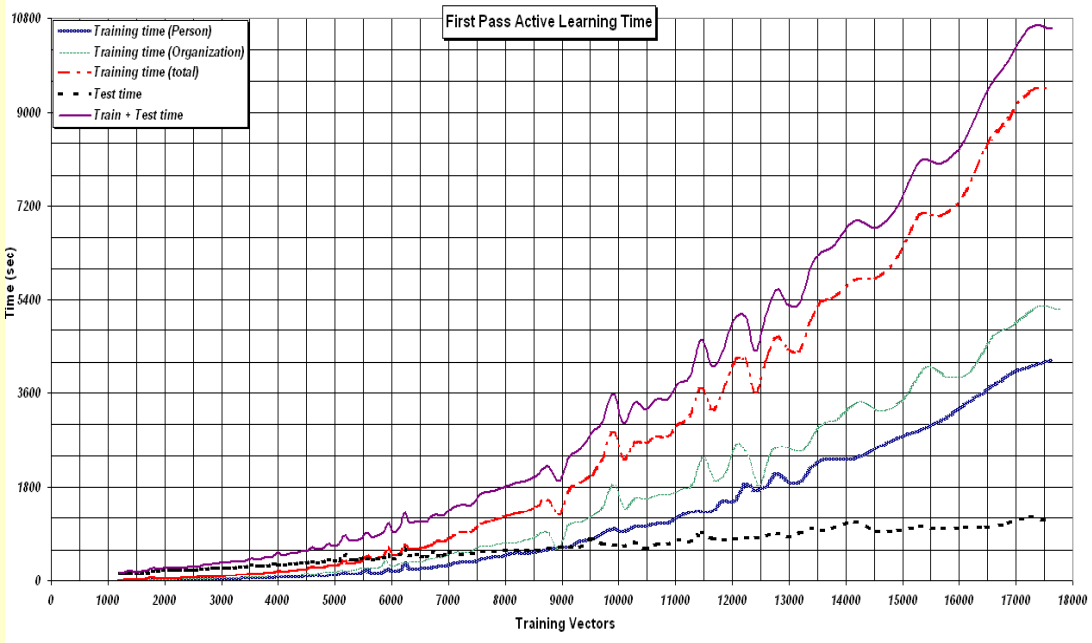
4.5.2.1 2η προσέγγιση

Όπως φαίνεται και από τα προηγούμενα διαγράμματα και κυρίως το 4.5-8, η προσέγγιση AL SP(2) είναι χειρότερη από την AL SP(1), κάτι που ενδεχομένως οφείλεται στο ότι οι ΜΔΥ του SP μαθαίνουν να θεωρούν πιο αξιόπιστες τις επιπλέον ιδιότητες του δεύτερου περάσματος (που αντανακλούν τις αποφάσεις του πρώτου περάσματος) από ό,τι είναι στην πραγματικότητα, όπως εξηγήθηκε στις προηγούμενες ενότητες.

4.5.3. Χρόνος διεξαγωγής πειραμάτων

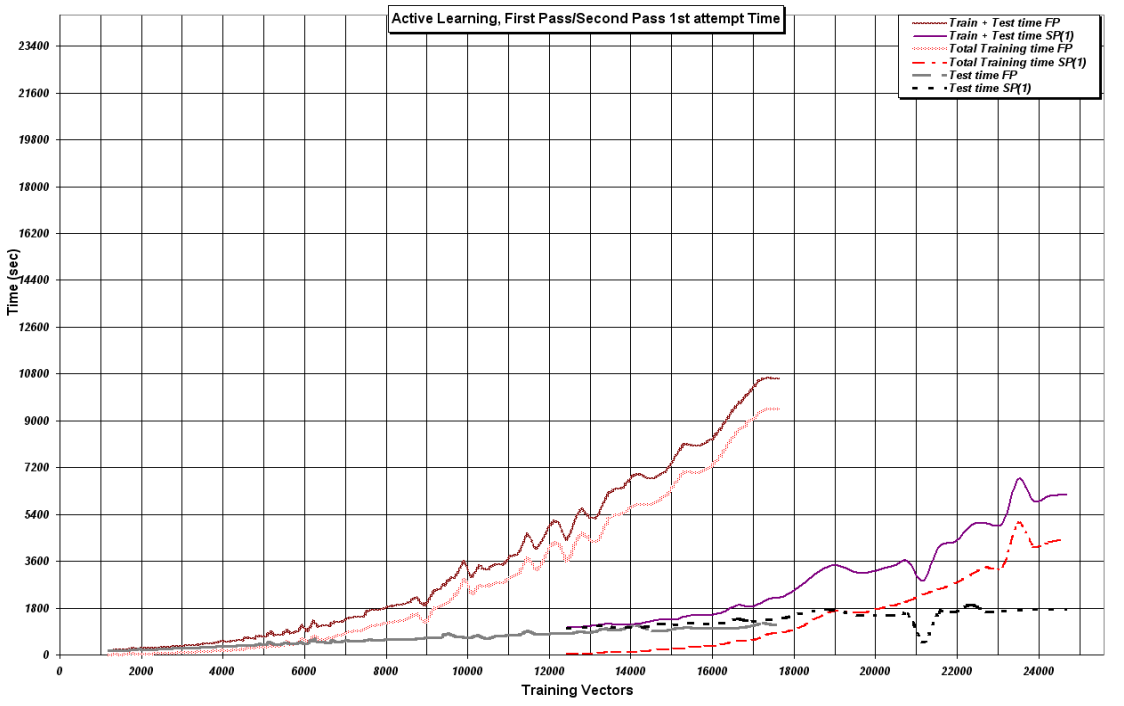
Στην περίπτωση της AL αναφέρουμε ότι ο χρόνος επισημείωσης κατά το πρώτο πέρασμα (ο χρόνος που χρειάζεται ο άνθρωπος-επισημειωτής για να κατατάξει χειρωνακτικά κάθε παράδειγμα) ήταν αρχικά πολύ μικρός, της τάξεως των **5 δευτερολέπτων ανά διάνυσμα**[6]. Στη συνέχεια ο χρόνος αυτός **αυξήθηκε σταδιακά** και κατέληξε να είναι περίπου ίσος με **6,5 δευτερόλεπτα ανά διάνυσμα**. Σε ότι αφορά το AL SP (1 και 2), ο αντίστοιχος χρόνος είναι εξαρχής περίπου 6,5 δευτερόλεπτα ανά διάνυσμα. *Αυτή η **σταδιακή αύξηση στο χρόνο** εξηγείται βάσει της δυσκολίας των παραδειγμάτων που εμφανίζει το σύστημα στο χρήστη: επιλέγονται περιπτώσεις για τις οποίες το σύστημα παραμένει αβέβαιο παρά την ως τότε εκπαίδευσή του, και οι περιπτώσεις αυτές είναι σταδιακά δυσκολότερες.*

Παραμένοντας λίγο ακόμα στην ανάλυση του SP, τονίζουμε ότι στο συνολικό χρόνο των πειραμάτων πρέπει να προσθέσουμε περίπου 35 λεπτά ανά πείραμα που είναι ο χρόνος υπολογισμού των αποφάσεων των ταξινομητών του FP. Τα επόμενα διαγράμματα δείχνουν τις καμπύλες χρόνου. Στον οριζόντιο άξονα μετρώνται τα ΔΕκ στα οποία εκπαιδεύονται οι αντίστοιχες ΜΔΥ και στον οριζόντιο ο χρόνος σε δευτερόλεπτα.

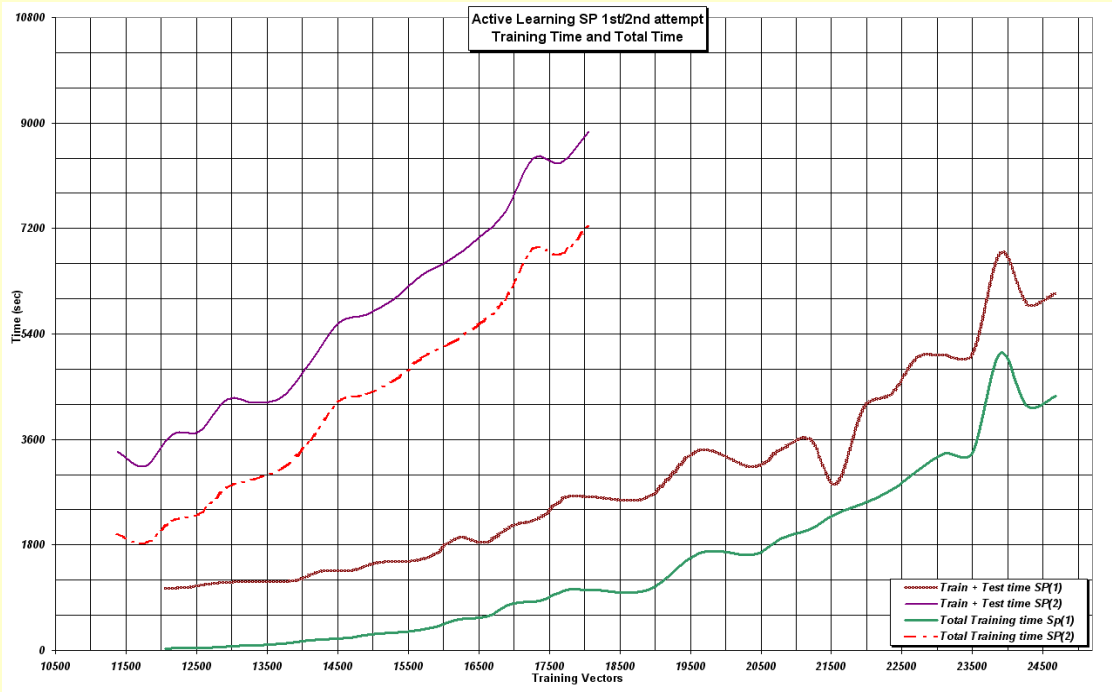


4.5-9 Διάγραμμα χρόνου πειραμάτων AL FP

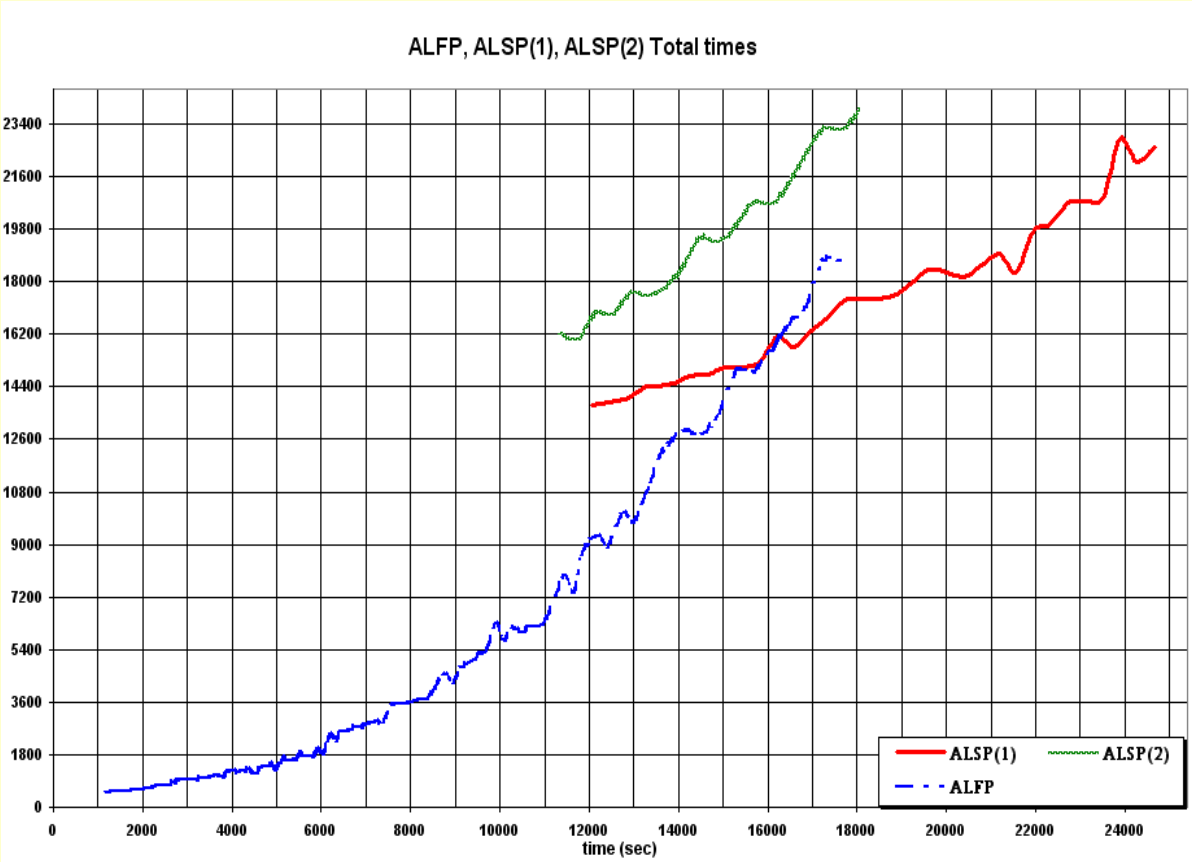
6 Σε αυτό το σημείο παίζει ρόλο και η ικανότητα του χρήστη ωστόσο η παράθεση χρονικών δεδομένων είναι ενδεικτική του χρόνου που απαιτείται.



4.5-10.1 Διάγραμμα σύγκρισης χρόνων πειραμάτων
AL FP – AL SP (1)



4.5-10.2 Διάγραμμα σύγκρισης χρόνων πειραμάτων
AL SP (1) – AL SP (2)



4.5-11 Διάγραμμα συνολικού χρόνου πειραμάτων AL

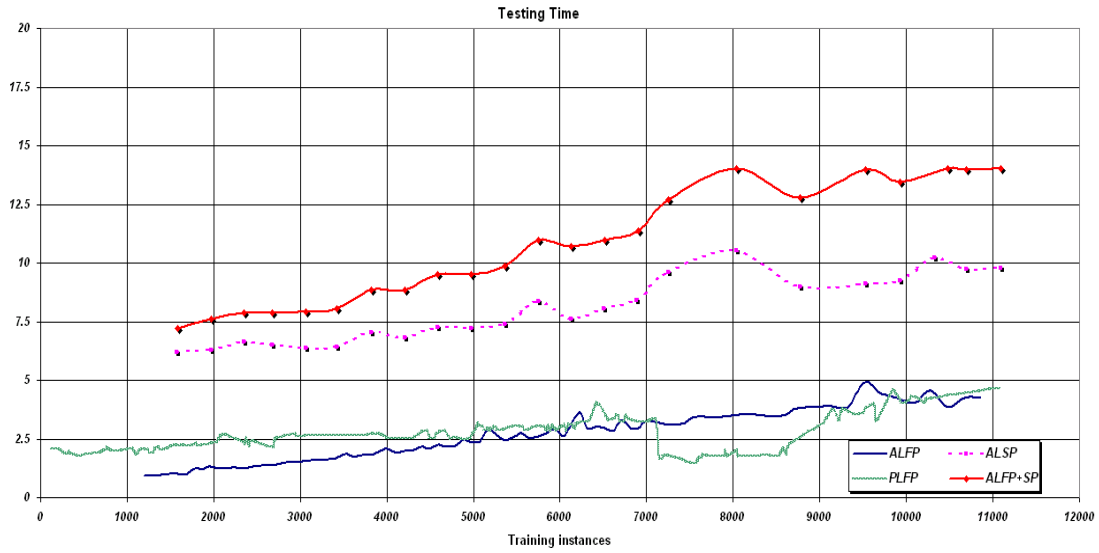
Στις καμπύλες περιλαμβάνονται οι επιμερισμένοι κατά πείραμα χρόνοι:

- εύρεσης των δεσμών (+15'' έως +55' για το Δεύτερο Πέρασμα 1ου τρόπου, +30' έως +65 για του 2ου, μεταβλητός χρόνος για το Πρώτο Πέρασμα από 5' έως 65'),
- επισημείωσης των διανυσμάτων (+45' για το Δεύτερο Πέρασμα, μεταβλητός χρόνος για το Πρώτο Πέρασμα από 5' έως 45'),
- υπολογισμού αποφάσεων Πρώτου Περάσματος για υπολογισμό ιδιοτήτων του Δεύτερου (+60 για το 1ο τρόπο, +100 για τον 2ο),
- μετατροπής των διανυσμάτων του Πρώτου σε Διανύσματα Δεύτερου Περάσματος 1ου τρόπου (+100).

Τα προηγούμενα μεταφράζονται κατά μέσο όρο σε + 300' για το Δεύτερο Πέρασμα 2ου τρόπου και +160 για το Δεύτερο Πέρασμα 1ου τρόπου, και +75 για το Πρώτο Πέρασμα

Από τις καμπύλες χρόνου παρατηρούμε ότι η εκπαίδευση των οργανισμών απαιτεί γενικά περισσότερο χρόνο σε σύγκριση με εκείνο για τα πρόσωπα, που υπονοεί ότι ο διαχωρισμός των θετικών και αρνητικών περιπτώσεων είναι δυσκολότερος στην περίπτωση των οργανισμών, κάτι που συμφωνεί με την γενική μας παρατήρηση ότι η κατηγορία των **οργανισμών είναι πιο «δύσκολη» απ' ότι εκείνη των προσώπων**.

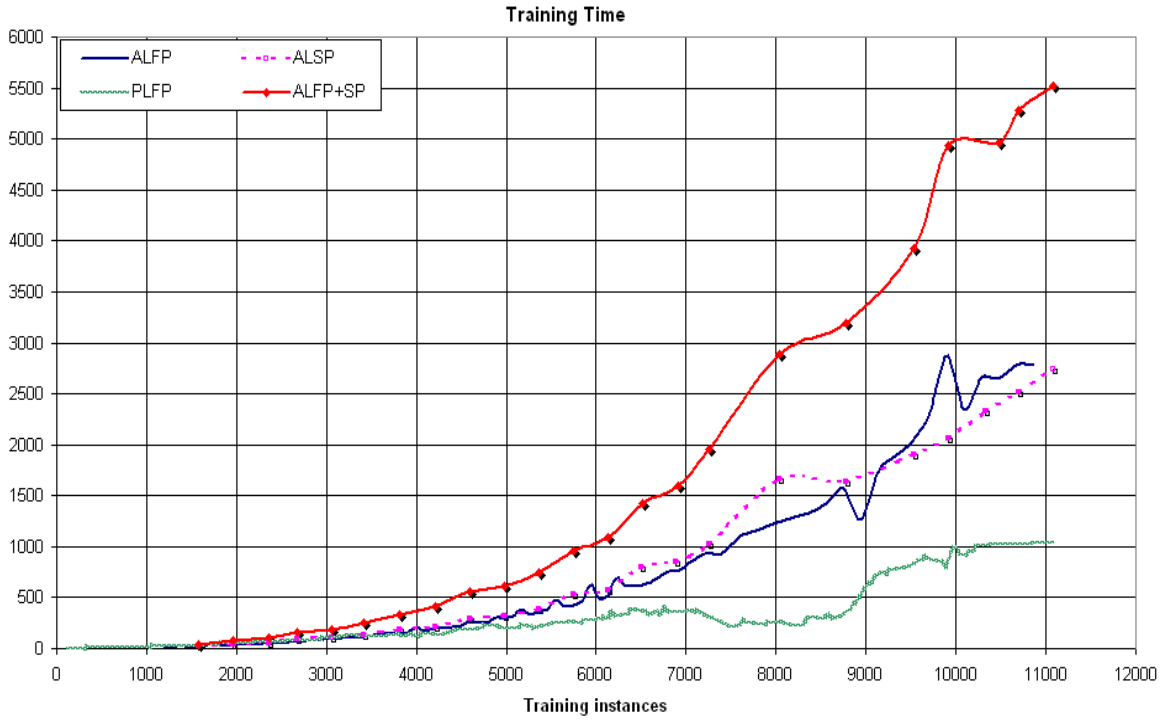
Στο προηγούμενο διάγραμμα παρατηρούμε ότι στην AL SP(2), λόγω των λιγότερων ΔΕκ στα οποία εκπαιδεύονται οι ΜΔΥ του δεύτερου περάσματος, οι χρόνοι που συνολικά απαιτούνται είναι μικρότεροι σε σχέση με την 2η προσέγγιση και το πρώτο πέρασμα. Διαπιστώνεται συνεπώς, ότι ο **1η προσέγγιση SP είναι χρονικά επικερδέστερη έναντι της 2ης και του FP**. Ίσως να συμφέρει γενικά να αλλάζουμε πέρασμα μετά από ένα εύλογο αριθμό ΔΕκ[7]. Τα επόμενα περάσματα βελτιώνουν λίγο τα αποτελέσματα f-measure, ωστόσο η ταχύτητα εκπαίδευσης είναι πολύ μεγαλύτερη.



4.5.12 Διάγραμμα χρόνου ελέγχου PL και AL για FP και SP

Ο κατακόρυφος άξονας αναφέρεται στα msec ανά διάνυσμα ελέγχου που χρειάστηκαν για να ολοκληρωθεί ο έλεγχος. Με αυτό τον τρόπο μπορούμε να εξάγουμε συμπεράσματα ανεξάρτητα από το πλήθος των διανυσμάτων ελέγχου. Το τελευταίο, θεωρείται αρκετά μεγάλο (167182 ΔΕλ), και περιλαμβάνει όλα τα διανύσματα των κειμένων ελέγχου, τόσο εκείνα για τα οποία αποφαίνονται οι ΜΔΥ, όσο και αυτά τα οποία απορρίπτονται από τις λίστες τετριμένων λέξεων ή τους ασφαλείς κανόνες.

[7] όταν για παράδειγμα η ακρίβεια των αποτελεσμάτων βελτιώνεται πολύ



4.5.13 Διάγραμμα χρόνου εκπαίδευσης PL και AL
για FP και SP

Ενδιαφέρον στοιχείο αυτού του διαγράμματος, πέραν της σύγκρισης μεταξύ των διαφορετικών περασμάτων που προσφέρει, είναι η καμπύλη ALFP+SP. Αυτή η καμπύλη παρουσιάζει το συνολικό χρόνο που χρειάστηκε το σύστημα για το πρώτο πέρασμα, συν το χρόνο για το δεύτερο πέρασμα με το πρώτο τρόπο (στον οριζόντιο άξονα προσμετρούμε μόνο τα διανύσματα εκπαίδευσης, όχι τα διανύσματα συνολικού κόπου).

5. Προτάσεις για βελτίωση του συστήματος

Στο τέλος αυτής της προσπάθειας για αναγνώριση ονομάτων οντοτήτων σε ελληνικά κείμενα εφημερίδων με χρήση ΜΔΥ, έχουμε αποκτήσει εμπειρία ως προς τα λάθη και τις παραλήψεις που σημειώθηκαν, και ιδέες που δε προλάβαμε να υλοποιήσουμε. Προτείνουμε στους μελλοντικούς χρήστες του συστήματος να προβούν σε αξιολόγηση των ακόλουθων ιδεών και προτάσεων, ώστε το ελεύθερα διαθέσιμο σύστημα να βελτιωθεί και να εξαχθούν καινούργια συμπεράσματα.

1. Κατά τη διάρκεια της PL, κάθε φορά που προσθέταμε νέα ΔΕκ επανυπολογίζονταν οι ιδιότητες και όλων των προηγούμενων ΔΕκ. (Μερικές από τις ιδιότητες χρησιμοποιούν λίστες λεκτικών μονάδων που εμφανίζονται συχνά κοντά σε ονόματα προσώπων και οργανισμών. Οι λίστες αυτές προκύπτουν από τα ΔΕκ που έχουμε συναντήσει και επομένως μεταβάλλονται όταν προστίθενται νέα ΔΕκ.) Αντίθετα **κατά την AL** του πρώτου περάσματος, **απλώς προσθέταμε τα επόμενα ΔΕκ** στα προηγούμενα, **χωρίς να επανυπολογίζονται οι ιδιότητες των προηγούμενων ΔΕκ**, κάτι που **αδικεί την εκπαίδευση** με AL για το πρώτο πέρασμα. Αυτό το πρόβλημα δεν αφορά την AL δεύτερου περάσματος, επειδή στο δεύτερο πέρασμα οι λίστες είχαν ήδη παγώσει από τα 9500 διανύσματα εκπαίδευσης του πρώτου περάσματος. *Προτείνεται η βελτίωση σε αυτό το σημείο για την AL FP.*

2. Η αξιολόγηση των υποσυνόλων των ιδιοτήτων για μια κατηγορία ονομάτων έγινε με κριτήριο τη μεγιστοποίηση του F-measure των αποτελεσμάτων για αυτή τη κατηγορία μεμονωμένα. Το σύστημα όμως ερωτά και τις δύο ΜΔΥ προκειμένου να ταξινομήσει κάθε διάνυσμα ελέγχου. Συνεπώς θα ήταν πιο ορθό να στοχεύαμε όχι μόνο στη βελτίωση του F-measure των αποτελεσμάτων μίας κατηγορίας, αλλά στη ταυτόχρονη βελτίωση και για τις δύο κατηγορίες.

3. Κατά την ενεργητική μάθηση, το **σύστημα φαίνεται να επιλέγει συχνά λεκτικές μονάδες που ξεκινούν από κεφαλαίο γράμμα**. Η πλειοψηφία αυτών των λεκτικών μονάδων προέρχεται από ένα σταθερό σύνολο με πιο συχνές τις λεκτικές μονάδες «Πρωθυπουργός», «Αρχιεπίσκοπος», «Τσάνπιονς Λίγκ», και λέξεις που προδίδουν εθνικότητα ή κράτος όπως «Δανός», «Γερμανός», «ΗΠΑ» κ.α. Θα μπορούσαν λοιπόν να δοκιμαστούν κάποιες ιδιότητες, οι οποίες να στοχεύουν στο να αυξηθεί η βεβαιότητα του συστήματος για την κατηγορία τέτοιων λεκτικών μονάδων. *Θα μπορούσαν, για παράδειγμα, να προστεθούν λίστες με εθνικότητες ή / και ονόματα κρατών, καθώς και*

ιδιότητες που να εξετάζουν αν οι λεκτικές μονάδες ανήκουν σε αυτές τις λίστες. Θα μπορούσαν, επίσης, να κατασκευαστούν αυτόματα τέτοιες λίστες από τα δεδομένα εκπαίδευσης, εντοπίζοντας λέξεις που ξεκινούν από κεφαλαία (ή έχουν μόνο κεφαλαία) και δεν είναι πρόσωπο / οργανισμοί .

4. Κατά την εκτέλεση των πειραμάτων AL έγινε αντιληπτό ότι το **παράθυρο λεκτικών μονάδων μεγέθους 7** (η υπό κατάταξη λεκτική μονάδα και οι τρεις προηγούμενες και επόμενες) **δεν είναι αρκετό**, αφού ήταν συχνά αδύνατο για το χρήστη να αποφανθεί με σιγουριά για την κατηγορία ενός παραδείγματος εξετάζοντας μόνο αυτό το παράθυρο. *Συμβουλεύουμε, λοιπόν, να αυξηθεί το μήκος του παραθύρου σε 11. Καλό θα ήταν, όμως, να κρατηθεί (με επιλογή ιδιοτήτων) ο τωρινός αριθμός ιδιοτήτων, ώστε το κόστος υπολογισμού των ιδιοτήτων να μην αυξηθεί υπερβολικά.*

5. Κατά την κατάταξη μιας λεκτικής μονάδας, ακολουθούσαμε την **απόφαση της ΜΔΥ (προσώπων ή οργανισμών) με τη μεγαλύτερη θετική βεβαιότητα .** Επίσης αν μόνο η μία ΜΔΥ επέστρεφε αρνητική απάντηση, τότε αυτομάτως κατατάσσαμε τη λεκτική μονάδα στην άλλη κατηγορία. Αν επέστρεφαν και οι δύο αρνητική απάντηση, τότε κατατάσσαμε τη λεκτική μονάδα ως *τίποτα*. Θα μπορούσε το σύστημα να δίνει τις βεβαιότητες των δυο ΜΔΥ σε *μια τρίτη ΜΔΥ, που θα εκπαιδευόταν (πάνω σε νέα δεδομένα) στο να παίρνει σωστές αποφάσεις βάσει των «εισηγήσεων» των άλλων δύο ΜΔΥ*, μια μορφή μάθησης γνωστή ως **stacking**.

6. Στη παρούσα εργασία, η **μεγάλη δεξαμενή μη επισημειωμένων κειμένων** για τα πειράματα AL **είχε χωριστεί σε 10 τμήματα** για λόγους οικονομίας χρόνου κατά την εύρεση των διανυσμάτων της δέσμης. Δεδομένου ότι το σύστημα μπορεί να υποστηρίξει ήδη μεθόδους συντόμευσης των υπολογισμών για την επιλογή διανυσμάτων εκπαίδευσης (βλ. [3] και ενότητα 3.3), *προτείνεται ο χωρισμός τη δεξαμενής κειμένων σε λιγότερα (ή κανένα) τμήματα. Αναμένουμε έτσι γρηγορότερη βελτίωση των αποτελεσμάτων του συστήματος. Τονίζουμε ότι το προηγούμενο προϋποθέτει και αρκετή διαθέσιμη μνήμη (περίπου 1GB RAM είναι υπεραρκετό για 5 τμήματα).*

7. ***Τέλος, προτείνουμε να μελετηθεί και η κατηγορία ονομάτων τοποθεσιών.*** *Η εισαγωγή μίας επιπλέον κατηγορία ονομάτων στο σύστημα, ως μια ακόμα ΜΔΥ, θα το επιβαρύνει ως προς την εκπαίδευση. Ενδέχεται ωστόσο να βελτιώσει περαιτέρω τα αποτελέσματα των άλλων δύο κατηγοριών. Υπενθυμίζουμε πως παρατηρήθηκε συχνά το φαινόμενο να αμφιβάλλει το σύστημα σχετικά με ονόματα δρόμων, τοποθεσίες, κράτη κ.ο.κ.*

6. Αναφορές σε άλλες εργασίες

[1] *«Αναγνώριση και Κατάταξη Ονομάτων σε κείμενα»,*
Λουκαρέλλι Γεώργιος (Διπλωματική Εργασία Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Επιστήμη Υπολογιστών, Οικονομικό Πανεπιστήμιο Αθηνών, 2005).
http://www.aueb.gr/users/ion/docs/lucarelli_msc_final_report.pdf

[2] *«Active Learning with Support Vector Machines»,*
Ανδρέας Βλάχος (School of informatics, University of Edinburgh, 2004)
http://www.inf.ed.ac.uk/publications/thesis/online/IM040138.pdf

[3] *«Fast Uncertainty Sampling for Labeling Large e-mail Corpora»,*
Richard Segal, Ted Markowitz, William Arnold (2006)
http://www.ceas.cc/2006/20.pdf

[4] *«LibSVM -A Library for Support Vector Machines»,*
Chih-Chung Chang and Chih-Jen Lin, (2005-2006)
http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[5] *«WEKA 3: A Java 2 implementation of machine learning algorithms for data mining tasks»,*
(1993 - 2006)
http://www.cs.waikato.ac.nz/~ml/weka/index.html

[6] *«MUC-6, Proceedings of the Sixth Message Understanding Conference»,*
(1995)

[7] *Άρθρα γενικού περιεχομένου (οικονομία, αθλητικά κ.α.) για την εκπαίδευση και τον έλεγχο του συστήματος:*

- ***"TA NEA"*, http://ta-nea.dolnet.gr**
- ***"TO BHMA"*, http://tovima.dolnet.gr**

[8] *«MUC-7, Proceedings of the Seventh Message Understanding Conference»,*
(1998)

[9] *«Incorporating Diversity in Active Learning with Support Vector Machines»,*
Klaus Brinker (2003)
http://www.hpl.hp.com/conferences/icml2003/papers/20.pdf

[10] «*Investigating the Effects of Selective Sampling on the Annotation Task»,*
Ben Hachey, Beatrice Alex and Markus Becker
http://www.ltg.ed.ac.uk/seer/papers/conll05.pdf

[11] *«Multi-Criteria-based Active Learning for Named Entity Recognition»,*
Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, Chew-Lim Tan
http://www.coli.uni-saarland.de/~dshen/publications/MScThesis.pdf
http://www.coli.uni-saarland.de/~dshen/publications/ACL04.pdf

[12] *«Statistical Learning Theory»,*
V. P. Vapnik (1998)

[13] *«An Introduction to Support Vector Machines»,*
N. Cristianini and J. Shawe-Taylor, (Cambridge University Press, 2000).

[14] *Java Native Interface JNI*
(jdk 6, 2006)
www.java.sun.com/tutorial

[15] *«A Greek Entity-Recognizer that Uses Support Vector Machines and Active Learning»,*
Λουκαρέλλι Γεώργιος και Ανδρουτσόπουλος Ιωάννης (Πρακτικά 4ου Πανελλήνιου Συνεδρίου Τεχνητής Νοημοσύνης (ΣΕΤΝ 2006), Ηράκλειο Κρήτης, 2006).
http://www.aueb.gr/users/ion/docs/setn2006_paper.pdf

ΠΑΡΑΡΤΗΜΑ

I. Επεξήγηση συντομογραφιών

Ελληνικοί χαρακτήρες

ΜΔΥ	Μηχανή Διανυσμάτων Υποστήριξης
ΔΕκ	Διανύσματα Εκπαίδευσης
Δελ	Διανύσματα Ελέγχου

Λατινικοί χαρακτήρες

AL	Active Learning (Ενεργητική Μάθηση)
AL FP	Active Learning First Pass (Ενεργητική Μάθηση Πρώτου Περάσματος)
AL SP	Active Learning Second Pass (Ενεργητική Μάθηση Δεύτερου Περάσματος)
AL SP (1)	Active Learning Second Pass, 1st approach (Ενεργητική Μάθηση Δεύτερου Περάσματος, 1η προσέγγιση)
AL SP (2)	Active Learning Second Pass, 2st approach (Ενεργητική Μάθηση Δεύτερου Περάσματος, 2η προσέγγιση)
FP	First Pass (Πρώτο Πέρασμα) NERCNamed Entity Recognition and Categorization (Αναγνώριση και Κατηγοριοποίηση Ονομάτων Οντοτήτων)
PL	Passive Learning (Παθητική Μάθηση)
PL FP	Passive Learning First Pass (Παθητική Μάθηση)
PL SP	Passive Learning Second Pass (Παθητική Μάθηση)
SP	Second Pass (Δεύτερο Πέρασμα)

II. Επεξήγηση υπολογισμού διαστημάτων εμπιστοσύνης

*Τα **διαστήματα εμπιστοσύνης,** αποτελούν διαστήματα τιμών ενός στατιστικού μεγέθους που μελετήθηκε πάνω σε ένα στατιστικό δείγμα. Αυτά τα διαστήματα φανερώνουν, με μια **πιθανότητα α%,** τις τιμές του μεγέθους σε ολόκληρο το σύνολο του πληθυσμού.*

Στη δική μας περίπτωση, τα μεγέθη που μας ενδιαφέρουν είναι η **ακρίβεια** και η **ανάκληση**. Επίσης, ως πληθυσμό θεωρούμε όλες τις λεκτικές μονάδες από όλα τα κείμενα Ελληνικών άρθρων γενικού θεματολογίου, και ως στατιστικό δείγμα τις λεκτικές μονάδες από τα κείμενα που χρησιμοποιήθηκαν κατά τον έλεγχο του συστήματος.

*Έστω ότι θέλουμε να υπολογίσουμε το διάστημα εμπιστοσύνης του μεγέθους της ανάκλησης **για τη κατηγορία των Προσώπων**, με πιθανότητα 95%, **και με δεδομένα ότι:***

*1ον **η ανάκληση έχει εκτιμηθεί από τον έλεγχο του συστήματος ότι είναι ίση με** p*

και

*2ον **έχουμε Π τω πλήθος διανύσματα ιδιοτήτων των λεκτικών μονάδων προερχόμενων από το σύνολο των κειμένων ελέγχου. (στο π περιλαμβάνονται και όσα ΔΕλ απορρίπτονται απευθείας από τους ασφαλείς κανόνες)***

Ισχύει οτι $\sigma = \sqrt{(p * (1-p) / n)}$ *και το διάστημα εμπιστοσύνης είναι:*

$p +/- 1.96 * \sigma$.

*Ο υπολογισμός του διαστήματος εμπιστοσύνης για το μέγεθος της ακρίβειας **είναι** αντίστοιχος, με τη διαφορά ότι p είναι η μέτρηση της ακρίβειας για τα πρόσωπα πάνω στο σύνολο ελέγχου. Επίσης, Π είναι ο αριθμός των ΔΕλ που η μέθοδός μας κατατάσσει ως Πρόσωπα (ανεξαρτήτως της ορθότητας της απόφασης).*

Συνεπώς $\sigma = \sqrt{(p * (1-p) / n)}$ *και το διάστημα εμπιστοσύνης είναι:*

$p +/- 1.96 * \sigma$.

Όλοι οι παραπάνω υπολογισμοί ισχύουν υπό την προϋπόθεση ότι

$n * p > 5 \quad n * (1-p) > 5)$