Δίκαρος Νίκος
Δ/νση Μηχανογράνωσης κ΄Η.Ε.Σ.
Υπουργείο Εσωτερικών.
Τελική εργασία Κ΄ Εκπαιδευτικής Σειράς Ε.Σ.Δ.Δ.
Επιβλέπων: Ηρακλής Βαρλάμης

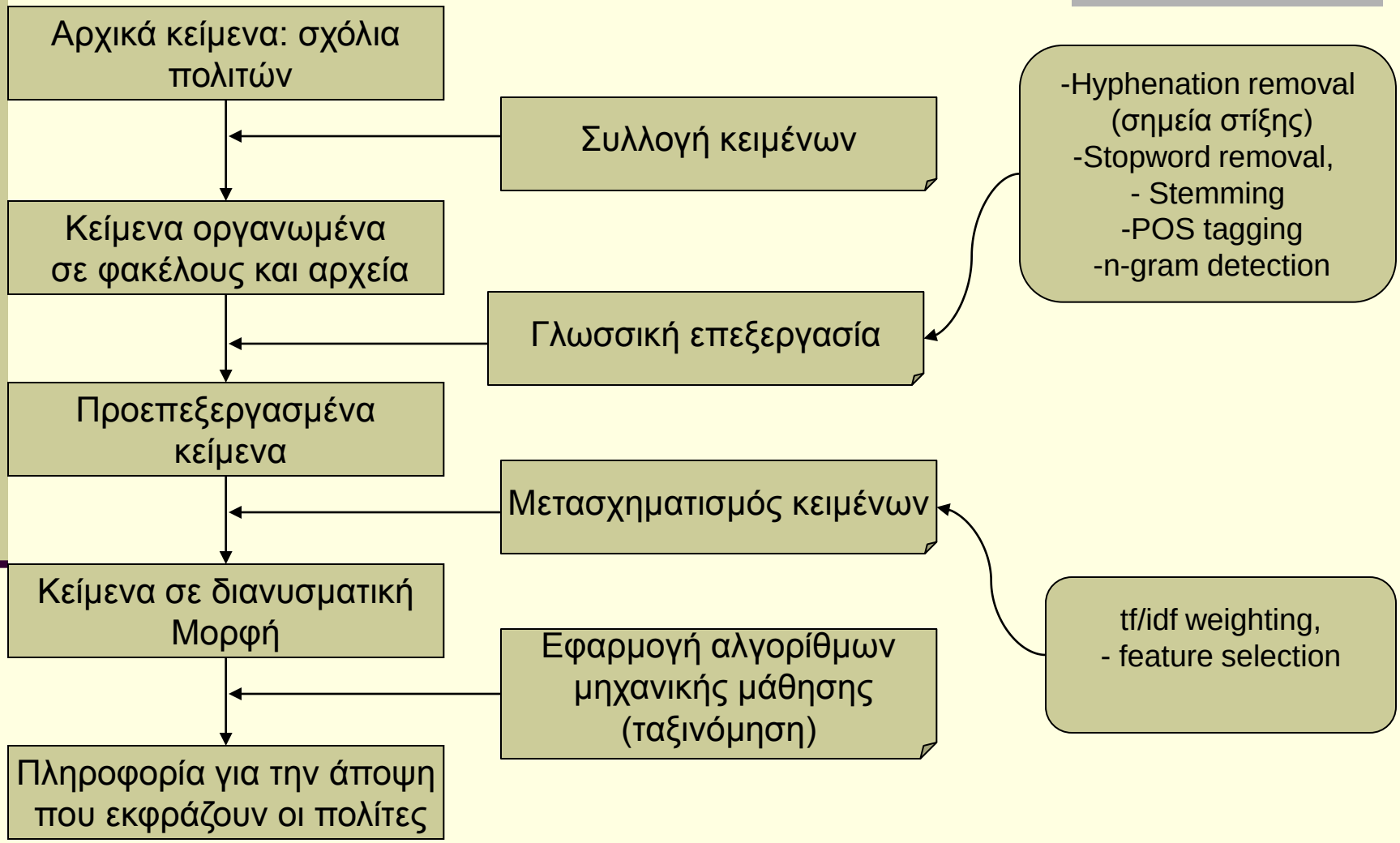
ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ

Εξόρυξη γνώμης πολιτών από ελεύθερο κείμενο

Κεντρική ιδέα

- Προβληματισμοί
  - Πως θα μπορούσαμε να ωφεληθούμε απ’ τον μεγάλο όγκο πληροφορίας υπό μορφή ελευθέρου κειμένου στο διαδίκτυο;
  - Μπορεί η Δημόσια Διοίκηση να εκμεταλλευτεί την πληροφορία αυτή;
- Προτάσεις
  - Εξαγωγή της γνώμης που εκφράζεται σε αυτά με αυτοματοποιημένο τρόπο, χωρίς την ανάγκη ανθρώπινης παρέμβασης.
- Οφέλη
  - Σφυγμομέτρηση λαϊκής βούλησης.
  - Διευκόλυνση στη λήψη αποφάσεων.
  - Προώθηση ηλεκτρονικής δημοκρατίας.

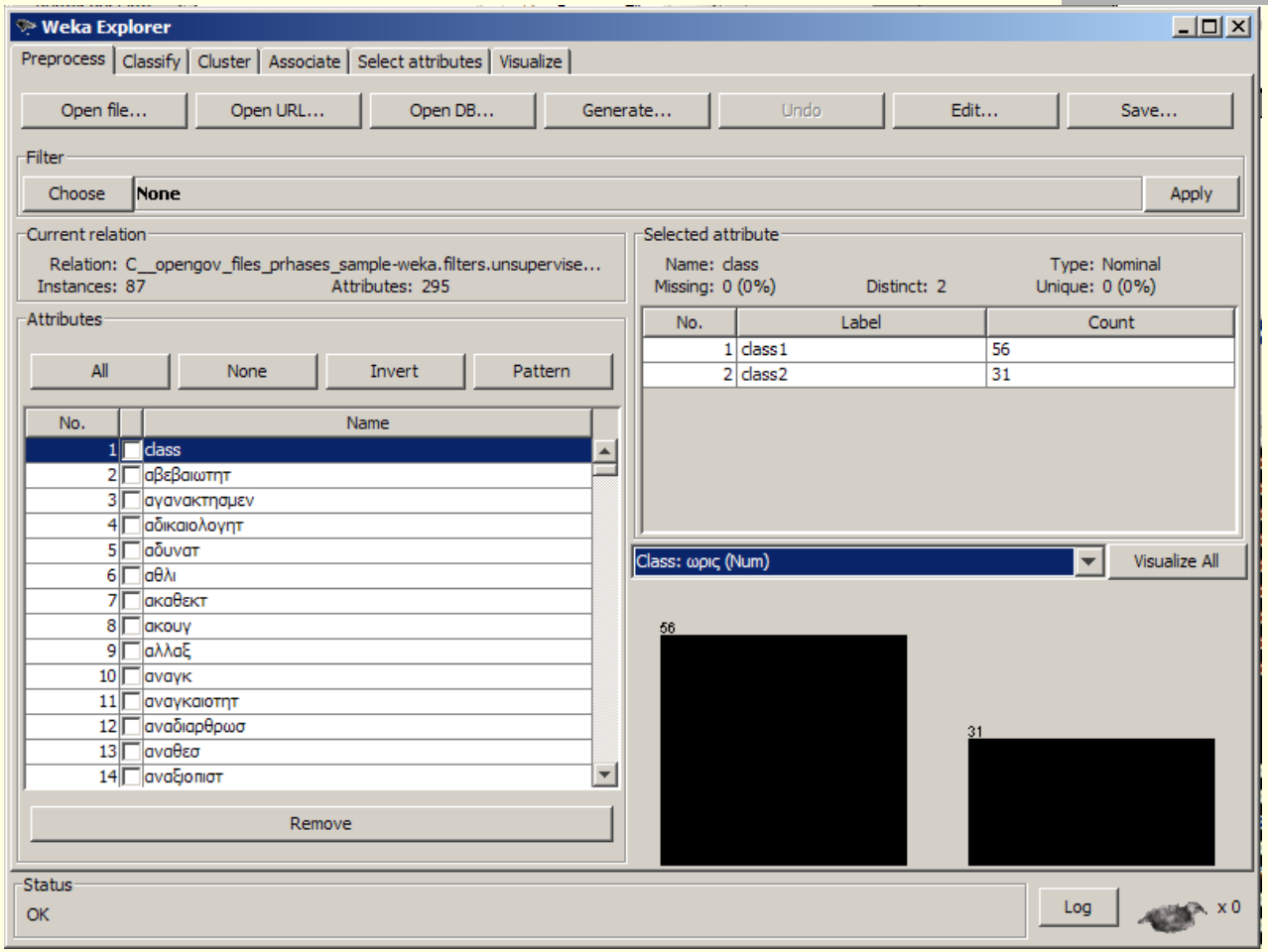
Διαδικασία εξόρυξης γνώμης



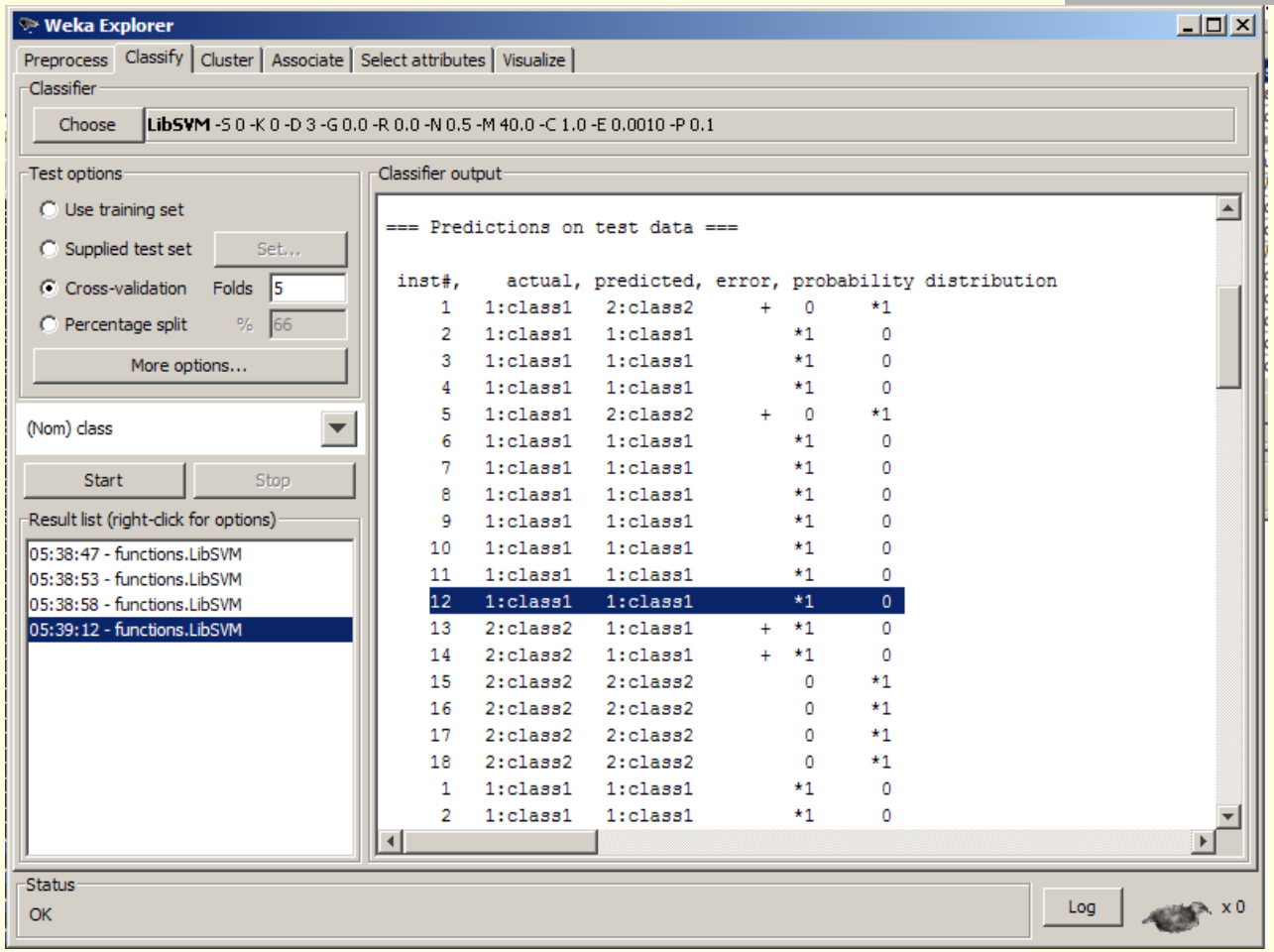
Λογισμικό Weka

- Ελεύθερη χρήση υπό την GNU General Public License
- Διαθέτει συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για ανάλυση δεδομένων και δημιουργία μοντέλων πρόβλεψης.
- Εύκολη προγραμματιστική ενσωμάτωση σε τρίτες εφαρμογές.
- Γραφική διεπαφή χρήστη που καθιστά εύκολη τη χρήση των δυνατοτήτων του.
- Υποστήριξη των πιο πολλών εργασιών που υπάγονται στον επιστημονικό κλάδο της εξόρυξης γνώσης.
- Μεγάλη φορητότητα λόγω υλοποίησης σε γλώσσα Java.

Λογισμικό Weka (2)



Λογισμικό Weka (3)



Ενδεικτική προεπεξεργασία (1)- αναρτημένα σχόλια στο διαδίκτυο

Σχόλια « 1 2 3

25 Ιανουαρίου 2010, 23:58 | [redacted] Μόνιμος Σύνδεσμος

ΤΙ να αξολογήσεις σήμερα από τους νομάρχες οι περισσότεροι είναι ελεγχόμενοι από κόμματα και κομματάρχες δυστυχώς δεν χαράζουν πολιτικές για τον τόπο τους ως όφειλαν, αλλά έγιναν διεκπεραιωτές ρουσφετιών από βουλευτές, μητροπολίτες, ιερείς, κόμματα, κ.λ.π. Είναι όλοι εξαρτώμενοι (υπάρχουν βέβαια και εξαιρέσεις - που πράγματι εργάστηκαν σκληρά κάτω από αντίξοες συνθήκες, άνθρωποι που αγάπησαν και υπηρέτησαν το θεσμό.) Γι΄αυτό θα πρέπει να αλλάξει εκ θεμελίων το σύστημα και να μπούμε σε καινούργιες και καινοτόμες λύσεις. Όπως πολύ σωστά κ. Υπουργέ βάλατε σε διαβούλευση αυτές της προτάσεις. Σύστημα περιφερειακής διακυβέρνησης (ΠΕΡΙΦΕΡΙΑΡΧΕΣ ΑΝΤΙΠΕΡΙΦΕΡΙΑΡΧΕΣ), εκτελεστική επιτροπή , ηλεκτρονική περιφερειακή διακυβέρνηση, πράσινη ανάπτυξη, περιφερειακό συμβούλιο - θα πρέπει να δοθούν πολλές αρμοδιότητες και πόροι χρηματοδότησής τους σαν μια τοπική κυβέρνηση, αλλά με αυστηρό ελεγκτικό μηχανισμό για να μην υπάρχει η αυθαιρεσία και η ανομία. Επειδή θα διαχειρίζονται τεράστια κονδύλια και θα πρέπει όλες οι αποφάσεις τους να δημοσιεύονται στο διαδύκτιο.Σωστό είναι και εδώ να εφαρμοσθεί ο περιφερειακός συνήγορος για τον πολίτη και την επιχείρηση.

Αναφέρετε το Σχόλιο

25 Ιανουαρίου 2010, 23:40 | [redacted] Μόνιμος Σύνδεσμος

Τοπική αυτοδιοίκηση

Η χώρα έχει 52 Νομούς.
Να φτιαχτούν 52 Δήμοι.

Και 8 περιφέρεις.
Κρήτης, Πελοποννήσου, Δωδεκανήσων, Επτανήσιων, Αττικής, Στερεάς, Ηπείρου, Μακεδονίας Θράκης.

Αυτό γιατί.

πχ. στην Κρήτη, ο βόρειος άξονας είναι όλες οι πόλεις όλα τα ξενοδοχεία όλος ο πληθυσμός θα είναι μαζεμένα όλα τα έσοδα, όλα τα λεφτά. Και στην Νότια Κρήτη με τα άπειρα χιλιόμετρα δρόμων, σωλήνων νερού, αποχετεύσεων, δεν θα έχουν καθόλου χρήματα για έργα.

Ο δήμαρχος στην πόλη δεν θα ξέρει τι να τα κάνει τα λεφτά. Γιατί όλες οι πόλης έχουν και σωλήνες νερού. Και αποχετεύσεις. Και δρόμους και από όλα.
Και ο δήμαρχος της υπαίθρου, ο κακομοίρης, θα θέλει να κάνει κάτι και δεν θα μπορεί.

Όποιος έχει λεφτά να έχει και έξοδα. Θέλεις την πόλη. Να έχεις και επαρχεία.

Γιατί αυτό που γίνετε τώρα είναι ότι στις πόλης φτιάχνονται τα ίδια και τα ίδια έργα συνέχεια. Γίνετε δηλαδή κακοδιαχείριση. Πετιούνται εκατομμύρια. Και ο άλλος, που έχει κάτι να κάνει δεν μπορεί.

Έχεις έσοδα να έχεις και έξοδα. Αυτό μπορεί να ρυθμιστεί αν φτιαχτούν μεγάλοι δήμοι που να έχουν και πόλη και ύπαιθρο. Και ο ποιο απλός τρόπος είναι ένας δήμος ολόκληρος ο νομός.

Ενδεικτική προεπεξεργασία(2)-αρχικό σχόλιο

- θα ήθελα κατ αρχήν να εκφράσω την χαρά και την ικανοποίηση μου για τον Καλλικράτη. Ασχολούμαι χρόνια με την τοπική αυτοδιοίκηση από τη θέση του ειδικού συνεργάτη, αλλά και μέσα από εθελοντικές δραστηριότητες. Επί της ουσίας του νομοσχεδίου δεν έχω να κάνω ιδιαίτερες παρατηρήσεις έαν ο σχεδιασμός, που το ελπίζω, εφαρμοστεί και λειτουργήσει, πιστεύω βαθιά, ότι θα αλλάξει η μορφή της Ελλάδας εξ αλλού, αυτή η αλλαγή, αποτελούσε επί χρόνια ελπίδα ανθρώπων που είχαν, άδω και δεκαετίες, οραματιστεί αυτή την αειφόρο ανάπτυξη των τόπων τους και μέχρι πρόσφατα αντιμετωπίζονταν από τους μικροεργολαβους της περιφερειακής πολίτικης σκηνής, λίγο ως πολύ, ως ψώνια

Ενδεικτική προεπεξεργασία(3)-σχόλιο μετά το stemming και την αφαίρεση stopwords

ηθελ κατ αρχην εκφρασ χαρ ικανοποιησ καλλικρατ ασχολ χρον τοπικ αυτοδιοικησ θεσ ειδ συνεργατ εθελοντικ δραστηριοτητ ουσι νομοσχεδ ιδιαιτερ παρατηρησ σχεδιασμ ελπιζ εφαρμοστ λειτουργ πιστευ βαθ αλλαξ μορφ ελλαδ εξ αλλ αλλαγ αποτελ χρον ελπιδ ανθρωπ δεκαετι οραματιστ αειφορ αναπτυξ τοπ προσφατ αντιμετωπιζ μικροεργολαβ περιφερειακ πολιτικ σκην ως ως ψων

Ενδεικτική προεπεξεργασία(4)- διανυσματική απεικόνιση (TF-IDF)

{8 4.787492,16 4.094345,24 4.787492,44 4.787492,57 4.787492,108 4.787492,110 4.787492,114 4.787492,143 4.787492,150 4.787492,152 4.787492,156 2.389596,158 4.094345,176 4.094345,214 4.094345,272 4.787492,287 4.787492}

- Κάθε πεδίο περιλαμβάνει 2 αριθμούς εκ των οποίων ο πρώτος συμβολίζει την αύξουσα σειρά του χαρακτηριστικού και ο δεύτερος την τιμή του στατιστικού μέτρου (εδώ TF-IDF)

Επιλογή αλγορίθμων, παράμετροι

- Χρησιμοποιούμενοι αλγόριθμοι
  - Αλγόριθμος KNN.
  - Αλγόριθμος Naïve Bayes.
  - Support Vector Machines.
- Τεχνικές-Παραμετροποίησης
  - Χρήση N-fold cross validation
  - Χρήση ξεχωριστού training set με προτάσεις που φέρουν έντονο εννοιολογικό προσανατολισμό (semantic orientation).

Παρουσίαση Αλγοριθμων (1)

- Αλγόριθμος (1)-KNN
  - Υπολογίζει την απόσταση ενός στιγμιότυπου από τους K κοντινότερους γείτονές του με βάση την ευκλείδεια απόσταση

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + ... + (x_{ip} - x_{lp})^2}$$

Παρουσίαση Αλγοριθμων(2)

- Αλγόριθμος Naïve Bayes που είναι απλός πιθανοτικός ταξινομητής βασισμένος στην εφαρμογή του θεωρήματος του Bayes.

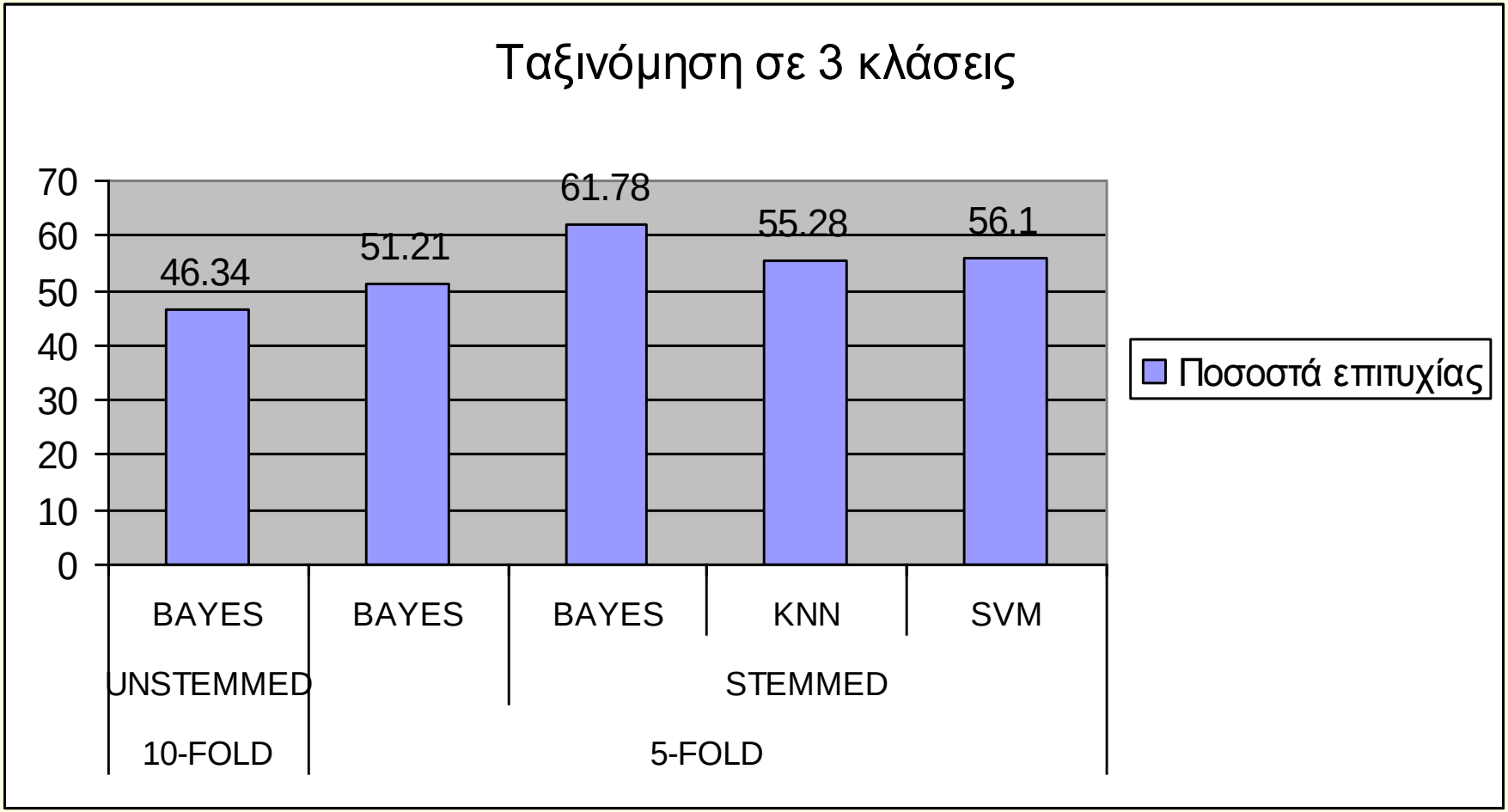
$$P(c \mid x) = P(c) \cdot \prod_i P(x_i \mid c)$$

- Μηχανές διανυσμάτων υποστήριξης(support vector machines)
  - Αλγόριθμος που δίνει παρόμοια αποτελέσματα με τα νευρωνικά δίκτυα σε προβλήματα κατηγοριοποίησης σε πολύ μικρότερο υπολογιστικό χρόνο.
  - Εκτενής χρήση του σε προβλήματα κατηγοριοποίησης κειμένου στην υπάρχουσα βιβλιογραφία

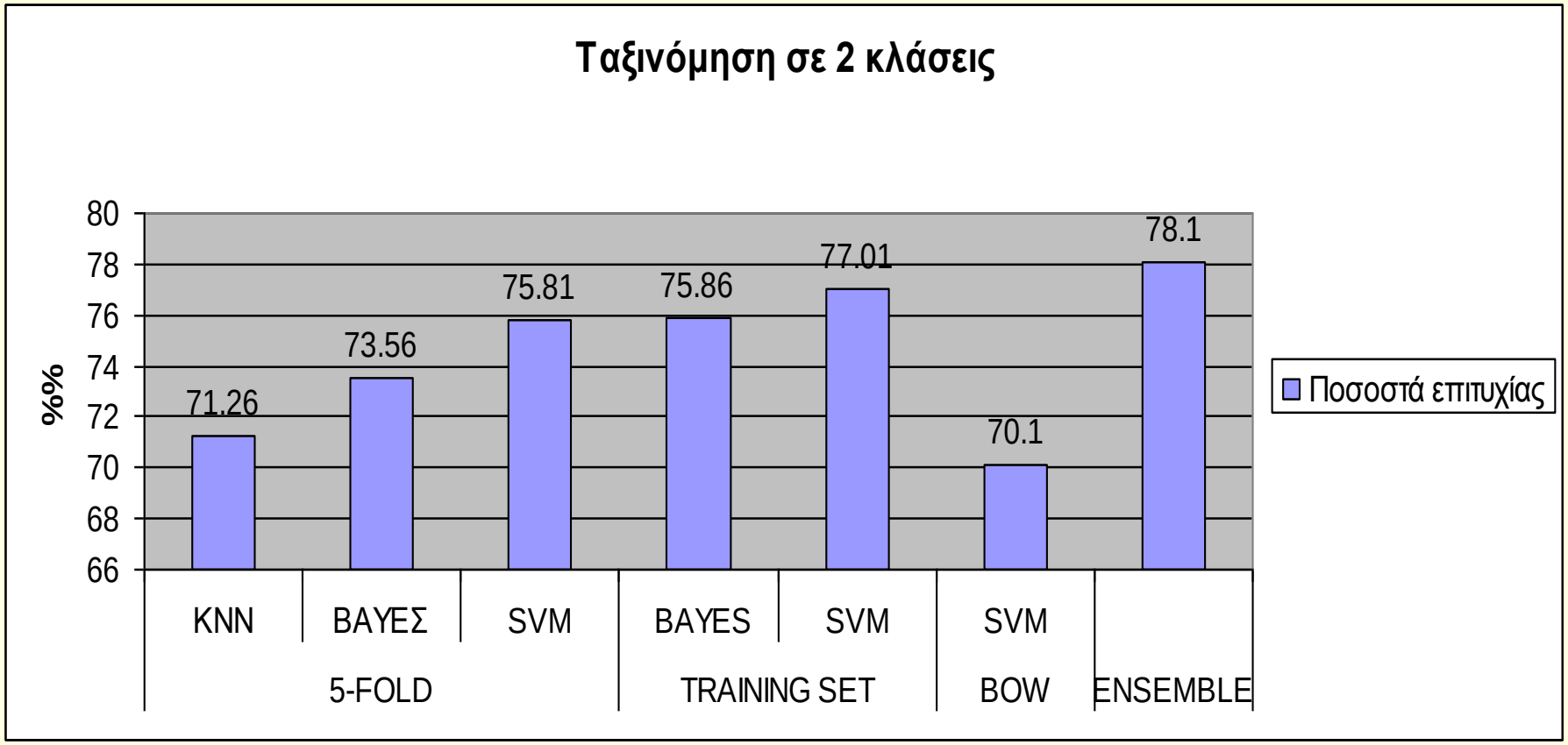
Αποτελέσματα πειραμάτων

- Δοκιμή διαχωρισμού του δείγματος σε 3 κλάσεις (θετικά, αρνητικά και ουδέτερα).
  - Ποσοστά επιτυχίας που κυμάνθηκαν από 44%-62%
- Δοκιμή διαχωρισμού του δείγματος σε 2 κλάσεις (θετικά, αρνητικά).
  - Ποσοστά επιτυχίας που κυμάνθηκαν από 68%-77%

Αποτελέσματα-3 κλάσεις



Αποτελέσματα-2 κλάσεις



Συμπεράσματα

- Οι αλγόριθμοι που χρησιμοποιήθηκαν προσέφεραν ικανοποιητικά αποτελέσματα τα οποία μπορούν να βελτιωθούν με τη χρήση τεχνικών γλωσσικής τεχνολογίας.
- Η συνδυασμένη χρήση αλγορίθμων (μεταταξινομητής) ξεπέρασε τις επιδόσεις των μεμονωμένων.
- Μεγάλο περιθώριο για περεταίρω έρευνα με δημιουργία νέων πειραματικών διατάξεων και δοκιμής νέων αλγορίθμων.

Μελλοντικές κατευθύνσεις

- Χρήση εξειδικευμένων εργαλείων γλωσσικής τεχνολογίας για βελτίωση των αποτελεσμάτων της κατηγοριοποίησης (π.χ Ellogon και BoosTexter).
- Δημιουργία ολοκληρωμένης εφαρμογής που θα χρησιμοποιεί λειτουργικές μονάδες του weka και άλλων λογισμικών και υλοποιεί πειραματικές διατάξεις κατηγοριοποίησης κειμένων.
- Δημιουργία web crowler για αυτοματοποίηση διαδικασίας συλλογής κειμένων από ιστοτόπους.

Ευχαριστώ για την προσοχή σας!!