Ευφυείς Τεχνικές για Εφαρμογές Αποθετηρίων

Α.-Γ. Σταφυλοπάτης

Ερευνητικό Πανεπιστημιακό Ινστιτούτο Συστημάτων Επικοινωνιών και Υπολογιστών

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εθνικό Μετσόβιο Πολυτεχνείο

ΕΥΦΥΕΙΣ ΥΠΗΡΕΣΙΕΣ ΑΝΑΖΗΤΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗΣ ΠΛΗΡΟΦΟΡΙΩΝ

Ανάκτηση - Επιλογή πληροφορίας

- Searching, mining, querying
- Εύκολη, ευέλικτη, διαισθητική πρόσβαση
- Ανοχή σε ανακρίβεια, αβεβαιότητα, ασάφεια
- Επέκταση ερωτήσεων, ευέλικτες ερωτήσεις
- Δυναμικό ταίριασμα αναζήτησης/αντικειμένου

Οργάνωση - Ανάλυση δεδομένων

- Ανάλυση κειμένου
- Συνόψιση (summarization)
- Προσθήκη δομής σε μη δομημένα δεδομένα
- Αυτόματη εξαγωγή μεταδεδομένων
- Σύνδεση με άλλες πηγές
- Μετα-αναζήτηση (metasearch)
- Αξιοποίηση σχεσιακών δεδομένων

ΕΥΦΥΕΙΣ ΥΠΗΡΕΣΙΕΣ ΑΝΑΖΗΤΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗΣ ΠΛΗΡΟΦΟΡΙΩΝ

Εξατομίκευση – Μοντελοποίηση χρήστη

- Προφίλ προτιμήσεων (user profiling)
- Ανάδραση σχετικότητας (relevance feedback)
- Συστάσεις - εισηγήσεις (recommender systems)

Οπτικοποίηση πολυδιάστατης πληροφορίας

- Κείμενο, εικόνες, πολυμεταβλητά δεδομένα
- 2Δ απεικονίσεις – Μείωση διαστάσεων
- Σημασιολογικοί χάρτες, θεματικοί χάρτες
- Πλοήγηση

ΒΑΣΙΚΕΣ ΤΕΧΝΙΚΕΣ

- Μοντελοποίηση
- Υπολογιστική νοημοσύνη (Computational Intelligence)
- Μηχανική μάθηση (Machine Learning)
    - Νευρωνικά δίκτυα (Neural Networks)
    - Ασαφή συστήματα (Fuzzy Systems)
    - Ταξινομητές πυρήνα / Support Vector Machines
    - Αυτο-οργανούμενοι χάρτες (Self Organizing Maps)
    - Πιθανοτικά μοντέλα και ταξινομητές

    - Ταξινόμηση (Classification)
    - Ομαδοποίηση (Clustering)

- Πραγματικές εφαρμογές με ερευνητικό ενδιαφέρον

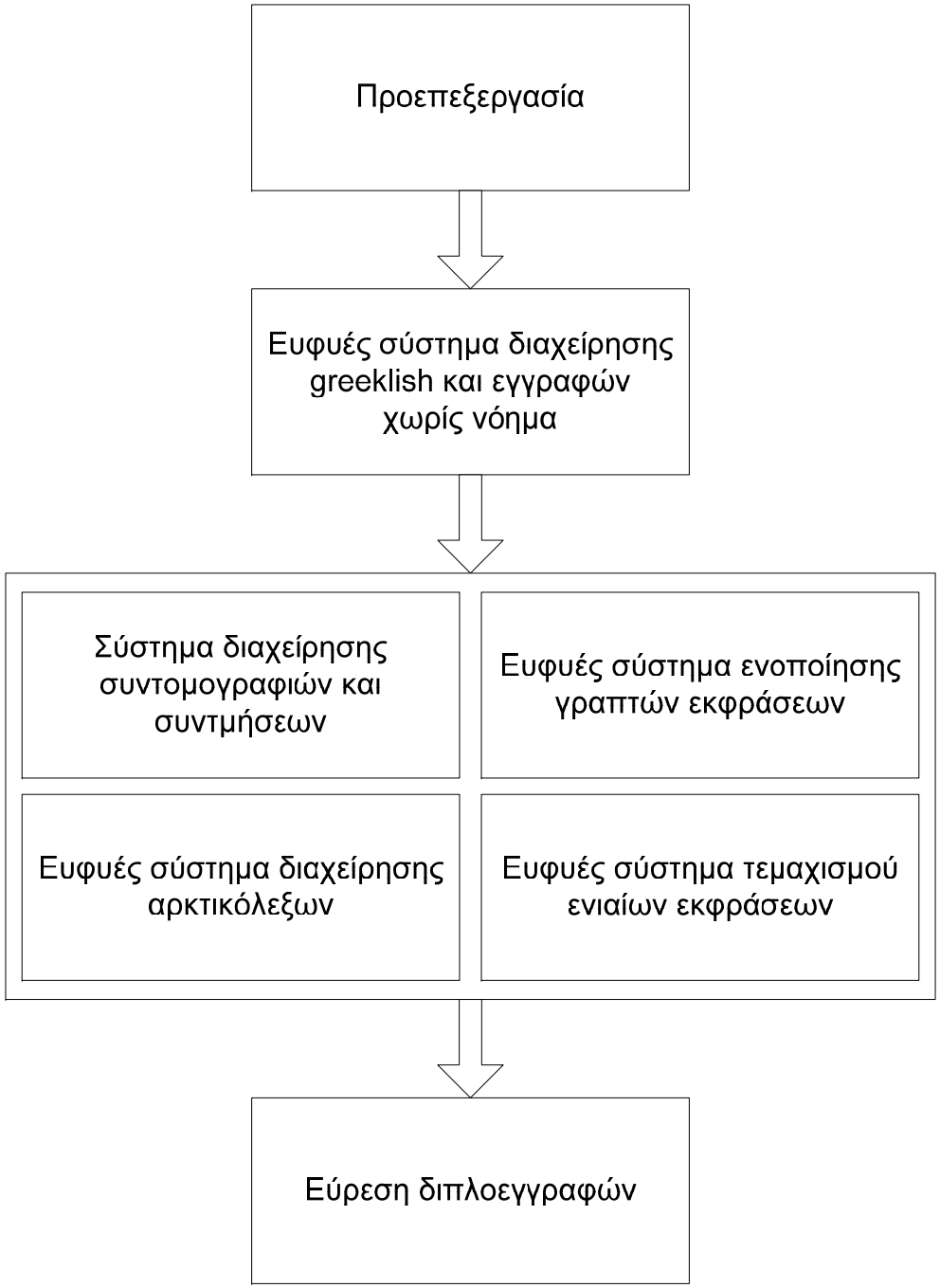
Καθαρισμός δεδομένων και εύρεση διπλοεγγραφών

- Ταυτόχρονη παρουσία εγγραφών σε διαφορετικές γλώσσες (ελληνικά/αγγλικά) ή και σε greeklish
- Παρουσία εγγραφών χωρίς νόημα (π.χ. εισαγωγή μόνο ειδικών χαρακτήρων σε πεδίο που θα έπρεπε να περιέχει όνομα)
- Ελλιπείς εγγραφές (με πολλά κενά πεδία)
- Ορθογραφικά ή τυπογραφικά λάθη
- Διαφορετικές γραπτές μορφές της ίδιας οντότητας (π.χ. συντμήσεις και αρκτικόλεξα σε αντιπαραβολή με τα αναπτύγματά τους)
- Ενιαίες εγγραφές που πρέπει να διαχωριστούν διότι περιλαμβάνουν πληροφορία που αντιστοιχεί σε διαφορετικά πεδία

- Διόρθωση του περιεχομένου της βάσης και διευκόλυνση της αναζήτησης διπλοεγγραφών
- Συνδυασμός ετερογενών βάσεων (λειτουργία κυρίως backoffice)

Καθαρισμός δεδομένων και εύρεση διπλοεγγραφών

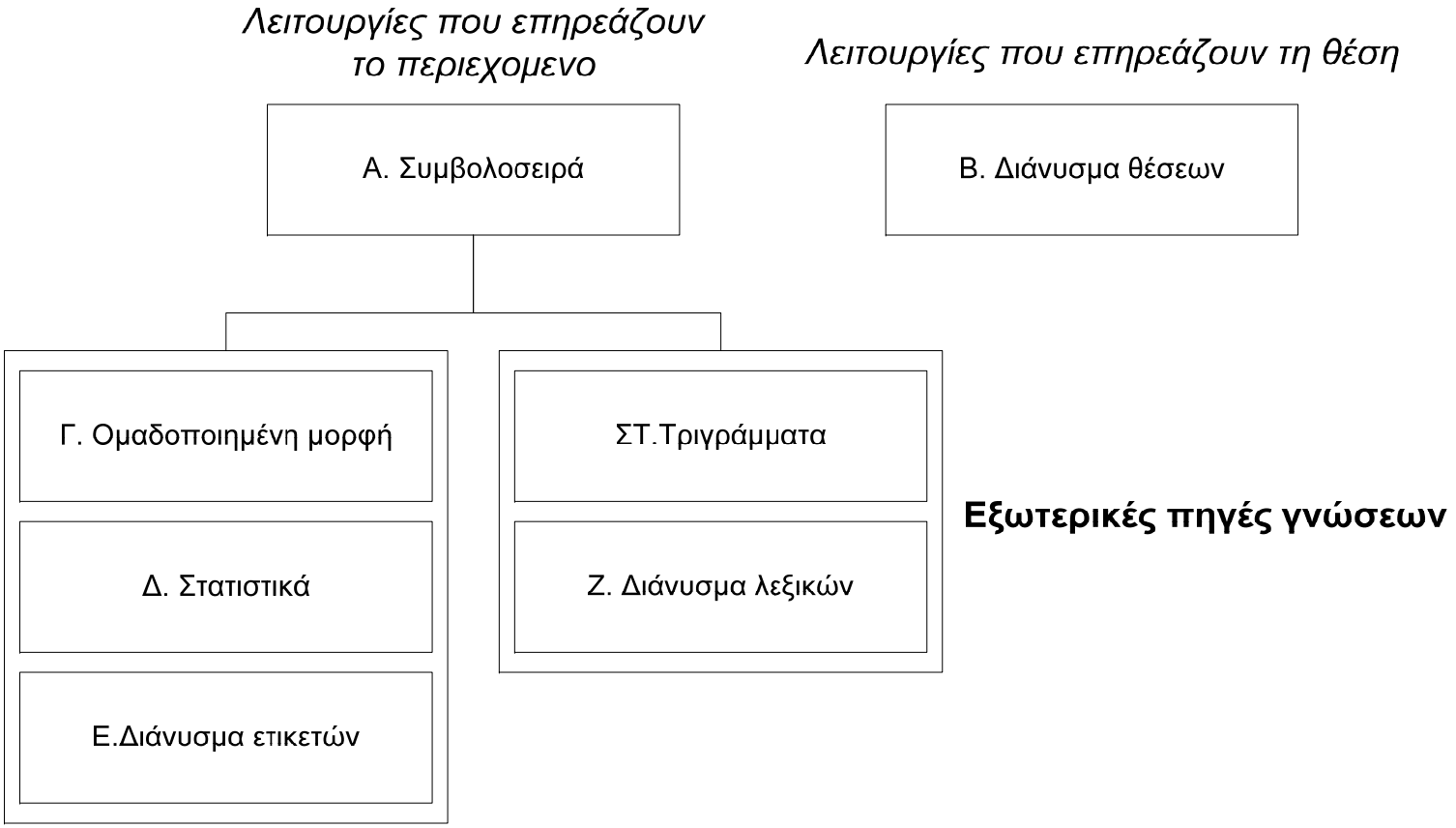
Αρχιτεκτονική του συστήματος



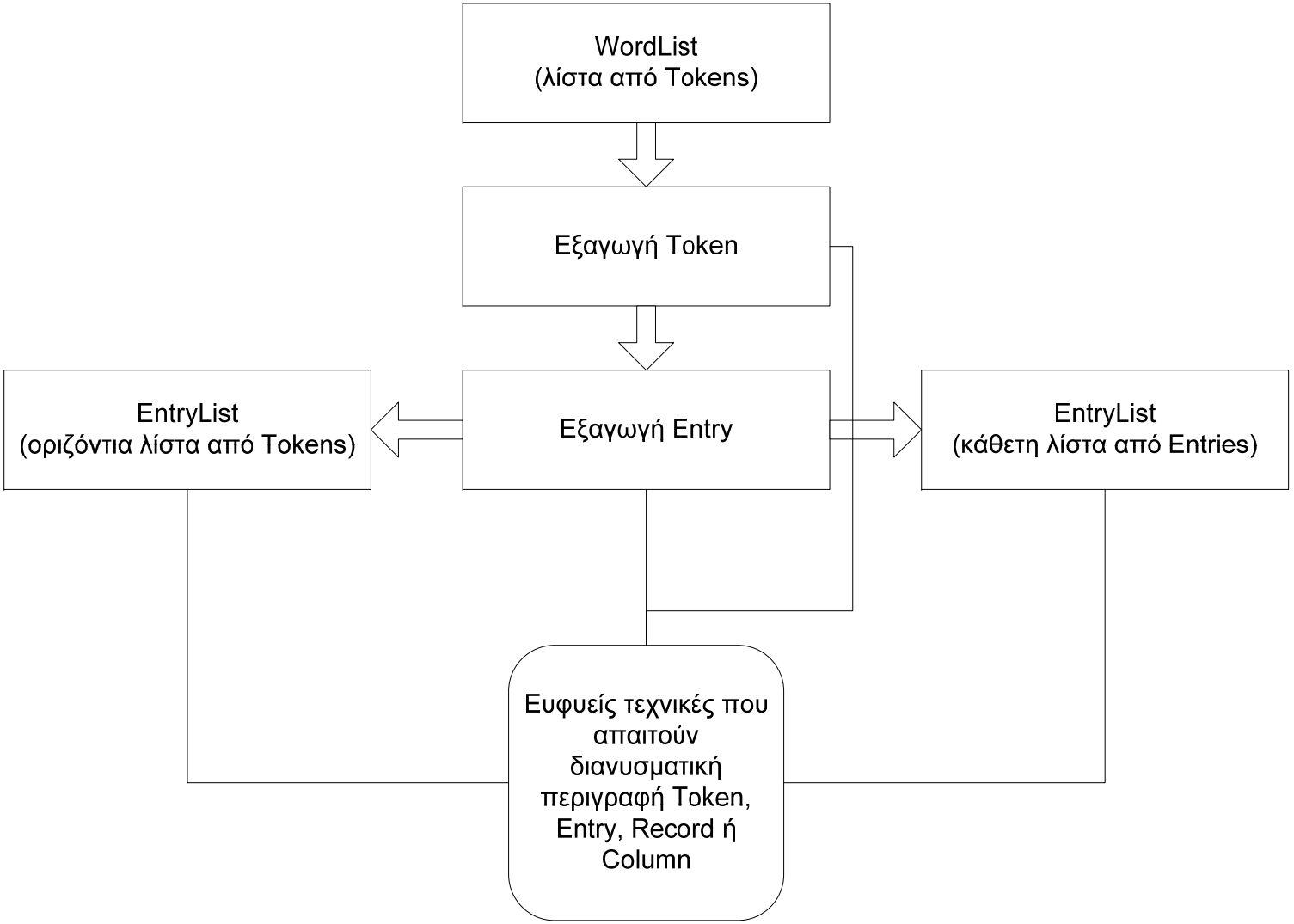
Σύστημα προεπεξεργασίας

- Αντικαταστάσεις και ομαδοποιήσεις (ανάθεση ετικετών)
- Δημιουργία αριθμητικών χαρακτηριστικών και αναπαράστασης του κειμένου
- Εμπλουτισμός των χαρακτηριστικών με εξωτερικές πηγές γνώσεων

Το βασικό αντικείμενο Token



Παραγωγή διανυσμάτων για ευφυείς τεχνικές από λίστα Tokens

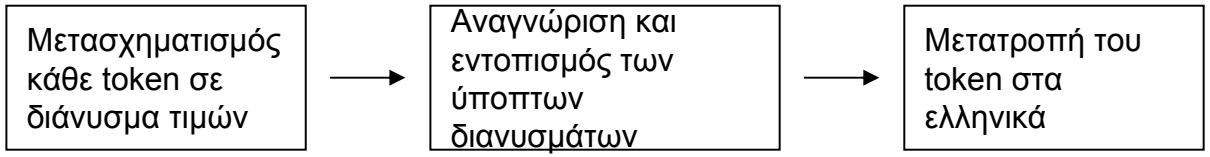


Καθαρισμός δεδομένων και εύρεση διπλοεγγραφών

- Αναγνώριση γλώσσας κειμένου/εγγραφών (ελληνικά/αγγλικά/greeklish) με χρήση αλφαβήτου
- Καθαρισμός από κείμενο/εγγραφές χωρίς νόημα – Εύρεση ποιότητας εγγραφών
- Διαχείριση κειμένου greeklish
- Διαδικασία μετατροπής λέξεων greeklish στα ελληνικά
  Βιβλιοθήκη Ινστιτούτου Επεξεργασίας του Λόγου (ΙΕΛ)

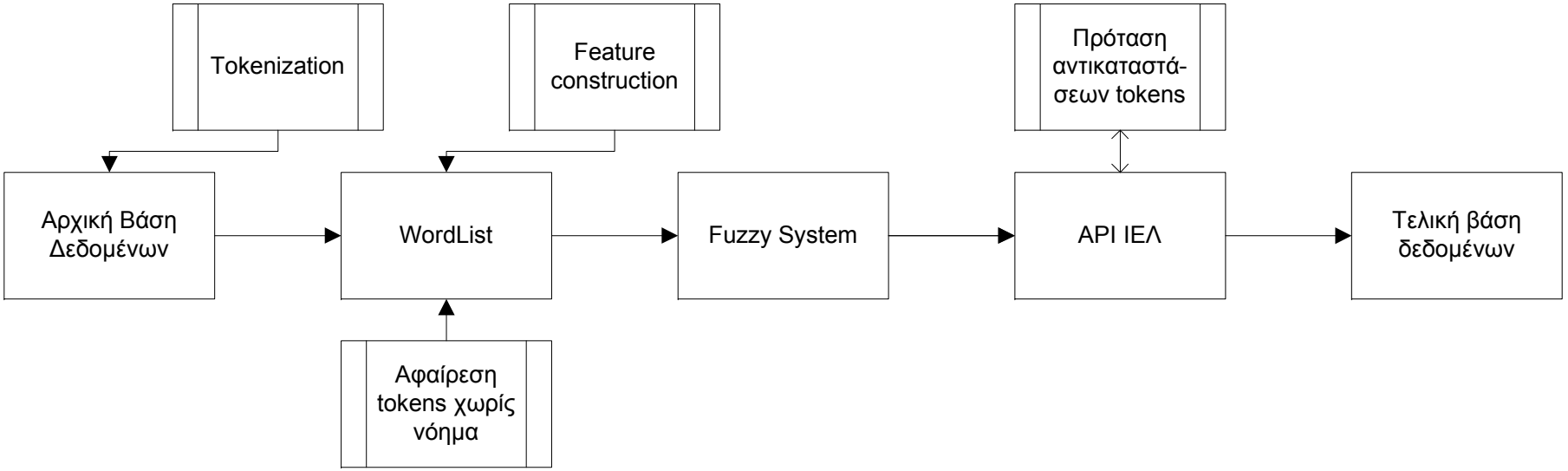
- Ανάπτυξη ελληνικού συστήματος Soundex

Διαχείριση κειμένου Greeklish

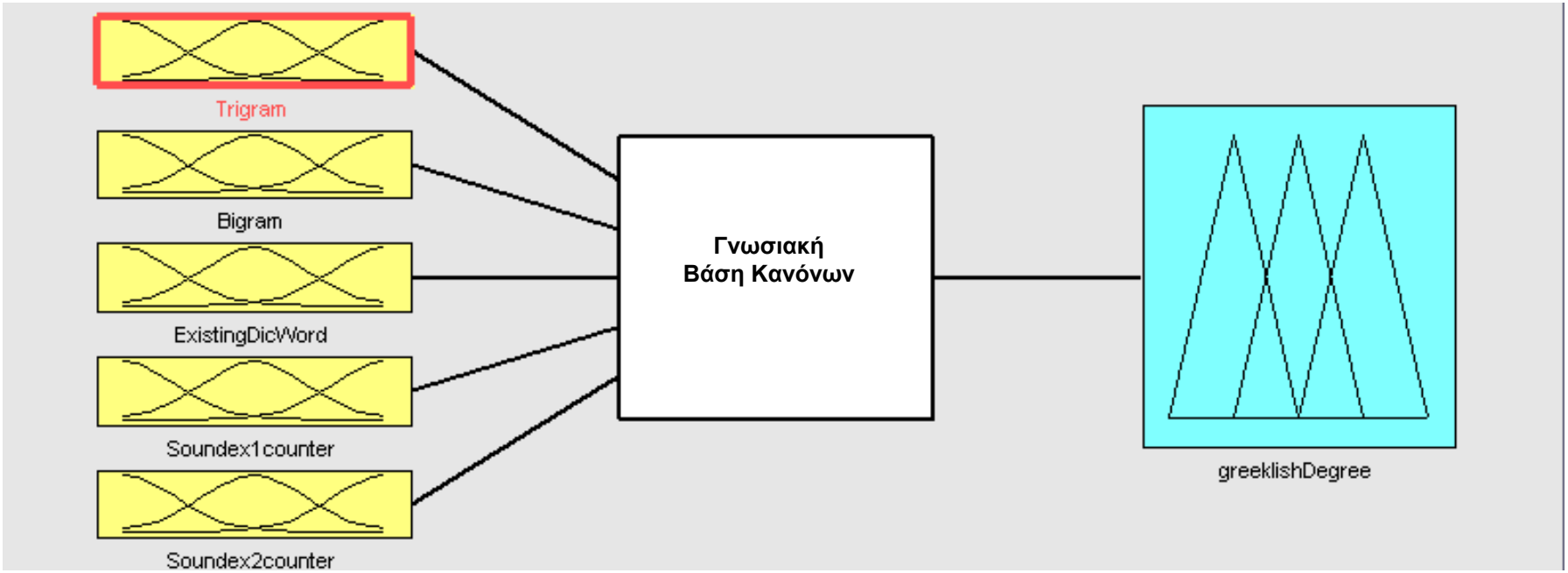


Μεταβλητές Διανύσματος			
Float	Float	Binary	Integer
Τριγραμματικό Σκορ	Διγραμματικό Σκορ	Ύπαρξη ή μη της λέξης στα αγγλοσαξονικά λεξικά	Πλήθος Ομόηχων λέξεων στα ελληνικά με βάση τον αλγόριθμο Soundex

Καθαρισμός δεδομένων και εύρεση διπλοεγγραφών



Ασαφής ταξινομητής λέξεων σε Greeklish ή Latin



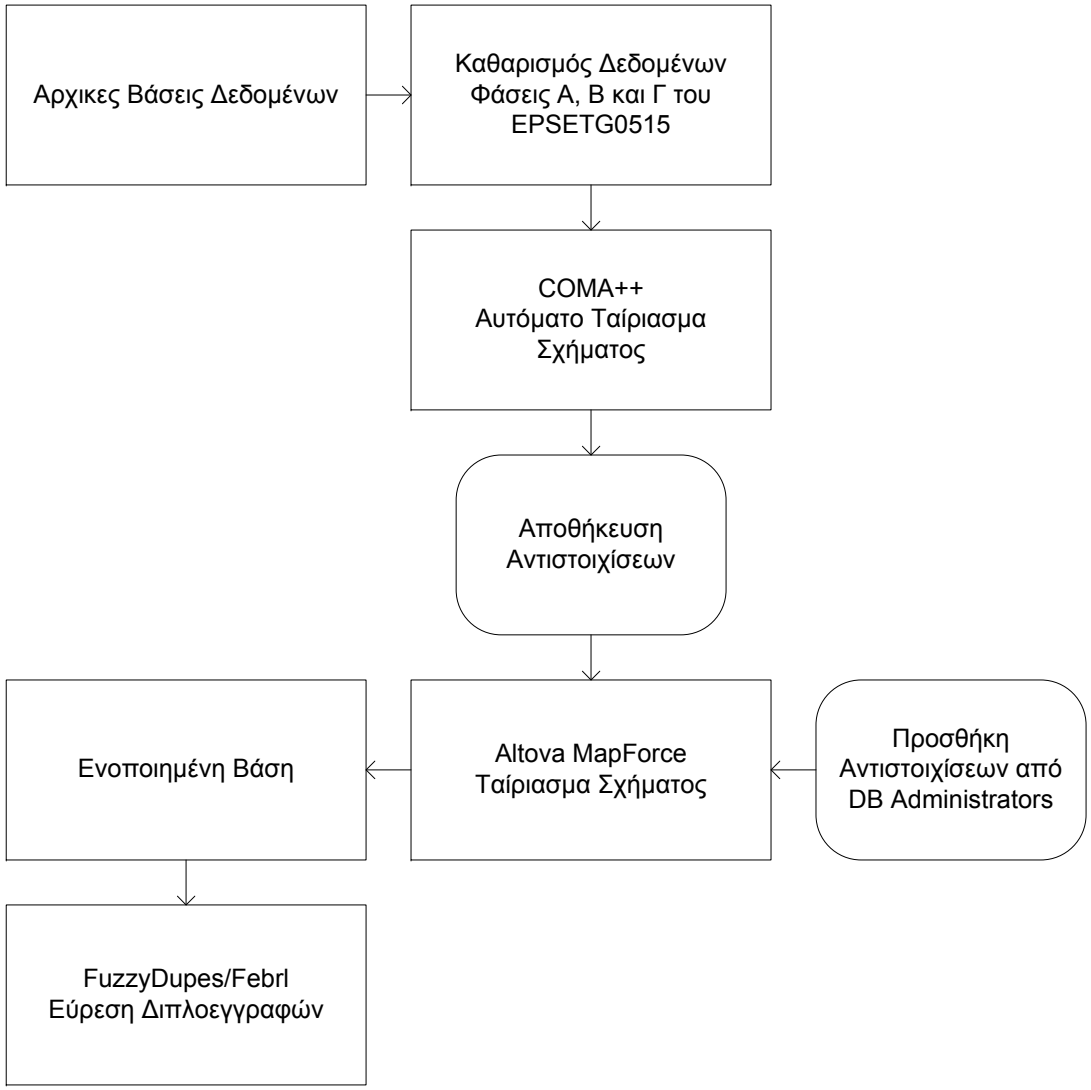
Γνωσιακή βάση κανόνων ασαφούς ταξινομητή

	Ασαφείς Κανόνες
1	ΕΑΝ *υπάρχει* στο αγγλικό λεξικό, ΤΟΤΕ η λέξη είναι μάλλον *αγγλική*
2	ΕΑΝ το Τριγραμματικό Σκορ είναι *χαμηλό* ΚΑΙ το Διγραμματικό Σκορ είναι *χαμηλό* ΚΑΙ *δεν υπάρχει* στο αγγλικό λεξικό ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-απλό είναι *χαμηλό* ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-επέκταση είναι *χαμηλό* , ΤΟΤΕ η λέξη είναι μάλλον *άγνωστη*.
3	ΕΑΝ το Τριγραμματικό Σκορ είναι *χαμηλό* ΚΑΙ *δεν υπάρχει* στο αγγλικό λεξικό ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-απλό ΔΕΝ είναι υψηλό, ΤΟΤΕ η λέξη είναι μάλλον *άγνωστη*
4	ΕΑΝ *δεν υπάρχει* στο αγγλικό λεξικό ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-απλό ΔΕΝ είναι *χαμηλό* ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-επέκταση ΔΕΝ είναι *χαμηλό* , ΤΟΤΕ η λέξη είναι μάλλον *greeklish*
5	ΕΑΝ το Τριγραμματικό Σκορ είναι *χαμηλό* ΚΑΙ *δεν υπάρχει* στο αγγλικό λεξικό ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-απλό ΔΕΝ είναι *χαμηλό* ΚΑΙ το πλήθος ομόηχων λέξεων με Soundex-επέκταση είναι *χαμηλό* , ΤΟΤΕ η λέξη είναι μάλλον *greeklish.*

Καθαρισμός δεδομένων και εύρεση διπλοεγγραφών

- Αντιστοίχιση αρκτικόλεξων με αναπτύγματά τους
- Διαχείριση συντομογραφιών και συντμήσεων
- Ευφυής ενοποίηση γραπτών εκφράσεων
- Εντοπισμός διπλοεγγραφών και διασύνδεση εγγραφών
- Εύρεση προτύπων σε δεδομένα
- Τεμαχισμός ενιαίων εκφράσεων
- Ταίριασμα σχήματος και ενοποίηση βάσεων δεδομένων

Καθαρισμός δεδομένων, ταίριασμα σχήματος μεταξύ βάσεων δεδομένων και εύρεση διπλοεγγραφών



Γενική αρχιτεκτονική του συστήματος

Καθαρισμός δεδομένων και εύρεση διπλοεγγραφών – Ταίριασμα σχήματος

- Εξόρυξη και επανάχρηση πληροφοριών
- Επίλυση προβλήματος πολλών διαφορετικων γραπτών εκφράσεων (αρκτικόλεξα, διαφορετικές γλώσσες, συντομογραφίες κλπ) που αναφέρονται σε σημασιολογικά ίδιο περιεχόμενο
- Προοπτική δημιουργίας οντολογιών από πολλές "επίπεδες" περιγραφές
- Επισήμανση λαθών/προβλημάτων στα δεδομένα - βελτίωση μελλοντικών διαδικασιών ανάκτησης και εισαγωγής δεδομένων, π.χ. χρήση δομημένων φορμών περιγραφής, πολλαπλών επιλογών, έλεγχος γλώσσας κλπ
- Συγκρότηση σημαντικών και επαναχρησιμοποιήσιμων γλωσσικών εργαλείων και βάσεων γνώσεων, όπου -ειδικά για τα ελληνικά- δεν υπάρχουν προηγούμενα διαθέσιμα σε μεγάλη κλίμακα

Εξαγωγή μεταδεδομένων, βιβλιογραφικών αναφορών και φράσεων-κλειδιών από διαφορετικούς τύπους κειμένων

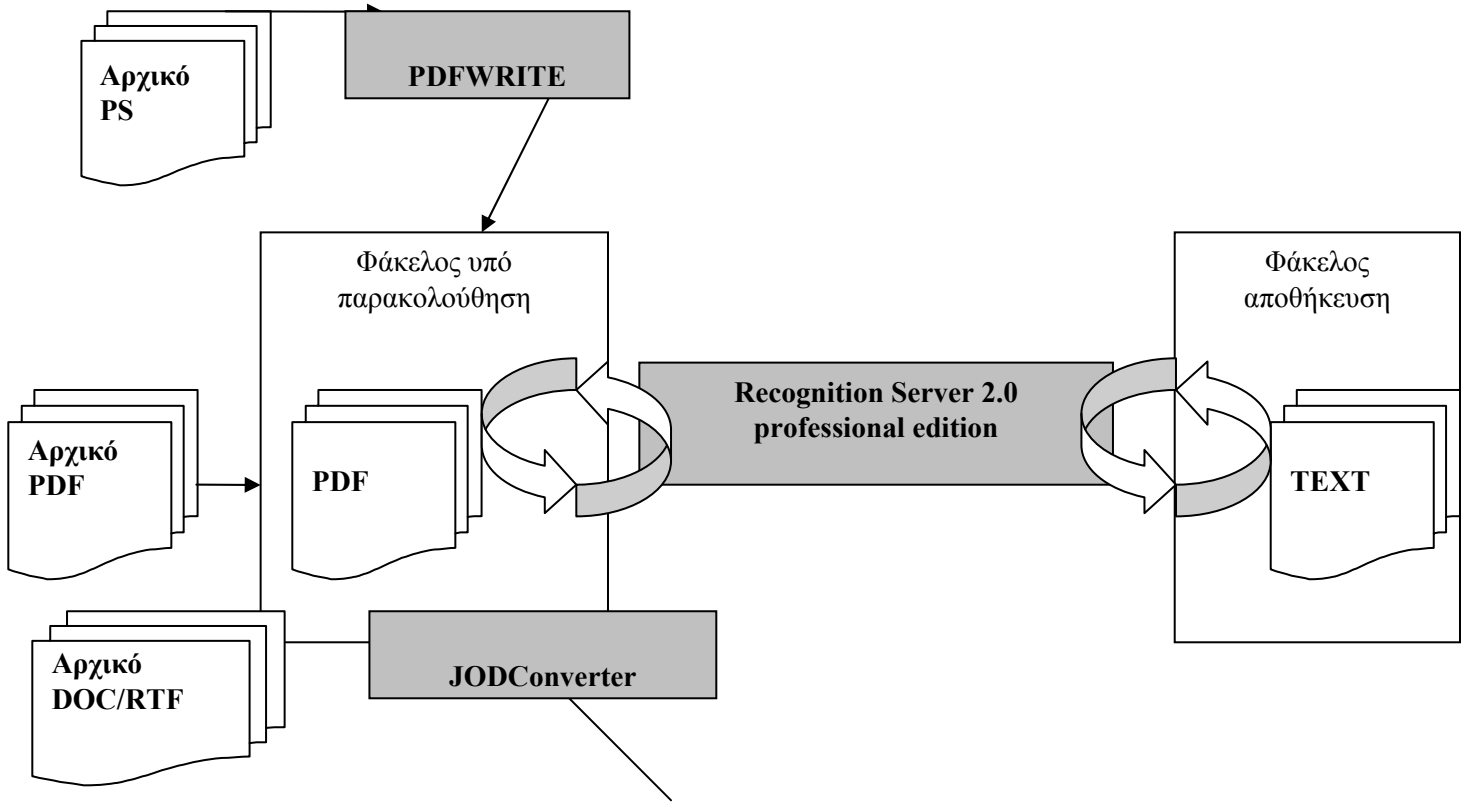
- Στην προοπτική ενός ηλεκτρονικού αποθετηρίου, μεγάλος όγκος πληροφορίας και χειροκίνητες διαδικασίες: ανάγκη για διευκόλυνση/επιτάχυνση της διαδικασίας αυτής με την αυτόματη εξαγωγή πληροφοριών από τα έγγραφα
- Χρήση ευφυών μεθόδων που επιτρέπουν την αυτόματη εξαγωγή προτεινόμενων τιμών για τα πεδία που θα πρέπει να συμπληρώσει ο χρήστης κατά την κατάθεση του εγγράφου του, όπως τίτλος, συγγραφέας κλπ., αλλά και την εξαγωγή φράσεων κλειδιών και βιβλιογραφικών αναφορών που θα συνοδεύουν το έγγραφο
- Διαχείριση διαφορετικών τύπων εγγράφων: επιστημονικά άρθρα, διατριβές

Σύστημα προεπεξεργασίας (ελληνικό/αγγλικό κείμενο)

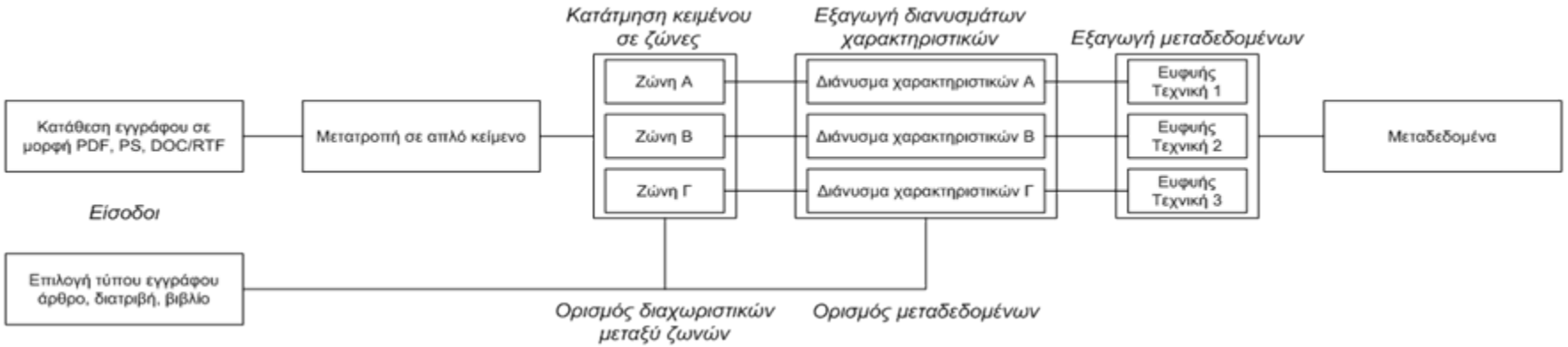
- Ανάγνωση δεδομένων κειμένου από διάφορους τύπους αρχείων (PDF, PS, DOC/RTF), αυτόματη ανάθεση ετικετών (tagging)
- Μετατροπή σε απλό κείμενο

- Αυτόματη κατάτμηση (segmentation) κειμένου σε ζώνες (zoning)
- Εξαγωγή διανυσμάτων χαρακτηριστικών (feature extraction) από κείμενο
- Προσδιορισμός σωμάτων δεδομένων για την εκπαίδευση και δοκιμή ευφυούς συστήματος και παραγωγή δεδομένων εκπαίδευσης (training set)

Μετατροπή Εγγράφων σε Απλό Κείμενο



Αρχιτεκτονική συστήματος εξαγωγής μεταδεδομένων



Έγγραφα τύπου επιστημονικής δημοσίευσης

HEADER PART

Ετικέτα μεταδεδομένου	HEADER _TAGS	Εξήγηση
Τίτλος	Title	Ο τίτλος του εγγράφου
Συγγραφέας	Author	Τα ονόματα των συγγραφέων του εγγράφου
Ίδρυμα	Affiliation	Το ίδρυμα που ανήκει ο κάθε συγγραφέας
Διεύθυνση	Address	Η διεύθυνση του κάθε συγγραφέα
Σημείωση	Note	Σημειώσεις σχετικά με copyright, αναφορές κλπ
Email	Email	Διευθύνσεις e-mail των συγγραφέων
Ημερομηνία	Date	Ημερομηνία δημοσίευσης
Περίληψη	Abstract	Περιγραφή του περιεχομένου και εισαγωγικό τμήμα του εγγράφου
Τηλέφωνο	Phone	Τηλέφωνα συγγραφέων
Λέξη κλειδί	Keyword	Λέξη κλειδί
Web	Web	URL συγγραφέα ή εγγράφου
Θέση	Degree	Λέξεις σχετικές με θέση
Αριθμός έκδοσης	Pubnum	Αριθμός έκδοσης του εγγράφου

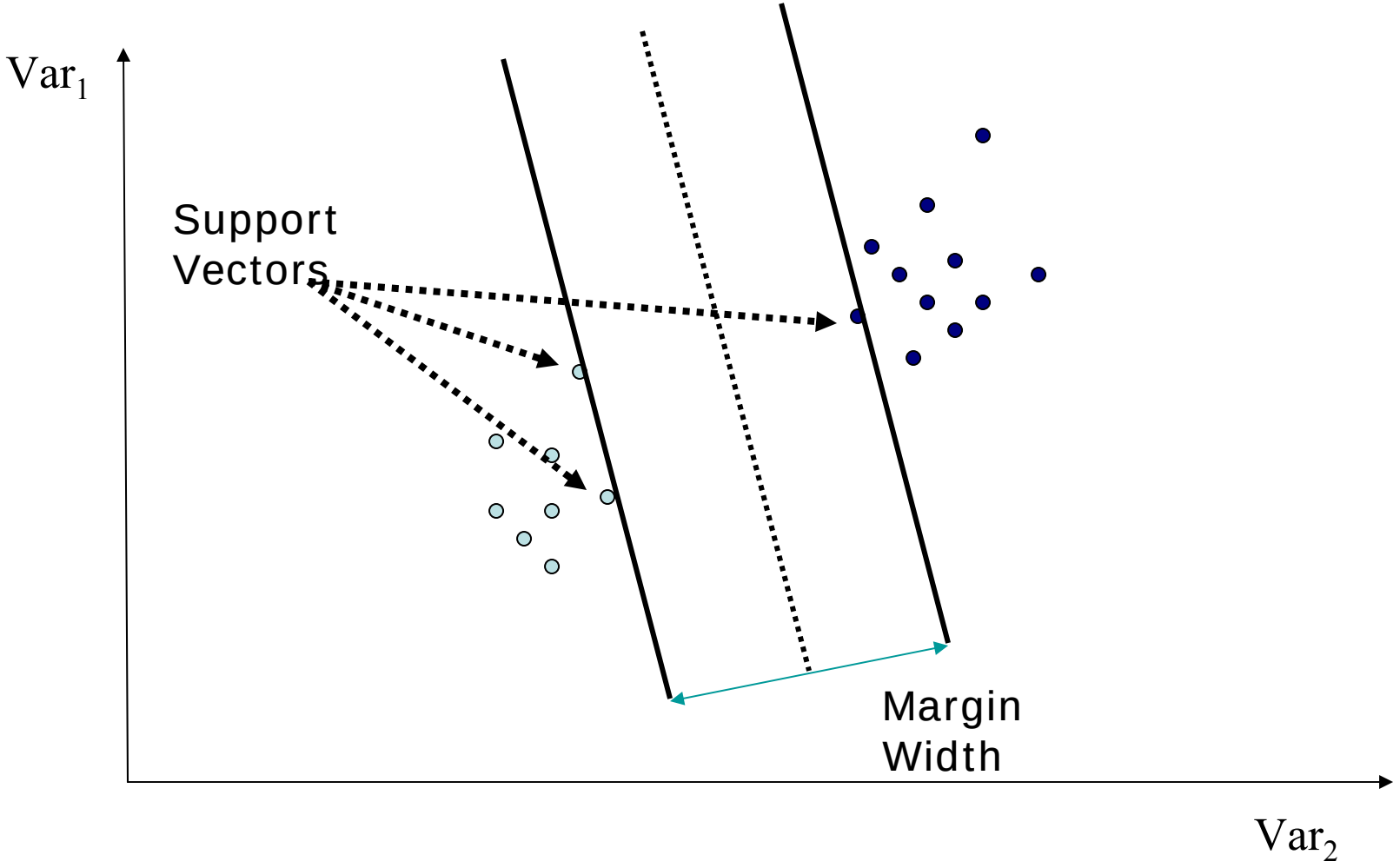
Έγγραφα τύπου επιστημονικής δημοσίευσης

REFERENCE PART

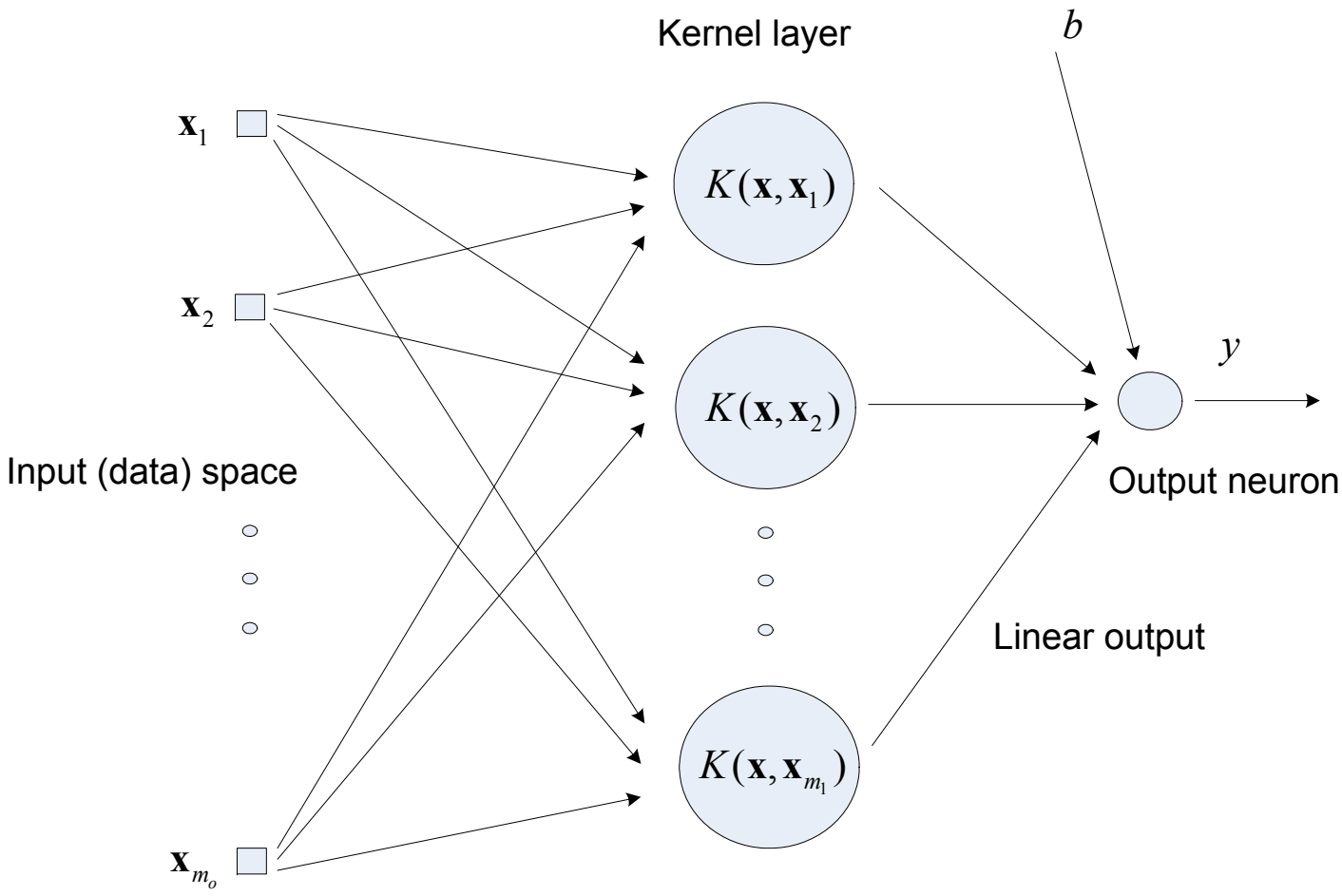
Ετικέτα μεταδεδομένου	**REF_TAGS**	**Εξήγηση**
Συγγραφέας	Author	Τα ονόματα των συγγραφέων
Τίτλος Proceedings	Booktitle	Τίτλος Proceedings
Ημερομηνία	Date	Ημερομηνία
Αρχισυντάκτης	Editor	Τα ονόματα των αρχισυντακτών
Ίδρυμα	Institution	Όνομα ιδρύματος
Περιοδικό	Journal	Τίτλος περιοδικού
Περιοχή	Location	Ονομασία περιοχής
Σημείωση	Note	Σημειώσεις σχετικά με copyright, submit, to appear κ.λ.π.
Σελίδες	Pages	Αναφορά σε σελίδες
Εκδότης	Publisher	Ονομασία εκδότη
Tech Report, Διδακτορικό, Master, Draft Paper	Tech	Χαρακτηρισμός όπως:Tech Report, Διδακτορικό, Master, Draft Paper
Τίτλος	Title	Τίτλος
Volume	Volume	Αριθμός Volume

Ευφυές Σύστημα Ταξινόμησης

Μηχανές Διανυσμάτων Στήριξης (Support Vector Machines-SVM)



Μηχανές Διανυσμάτων Στήριξης (Support Vector Machines-SVM)



HEADER PART - Σύστημα ταξινόμησης με συνδυασμό μοντέλων SVM

Κάθε γραμμή του HEADER PART (ή μέρος αυτής) κατατάσσεται σε μία ή περισσότερες από 14 κατηγορίες:

- title
- author
- address
- affiliation
- abstract
- phone
- pubnum
- web
- date
- degree
- email
- note
- keyword
- intro

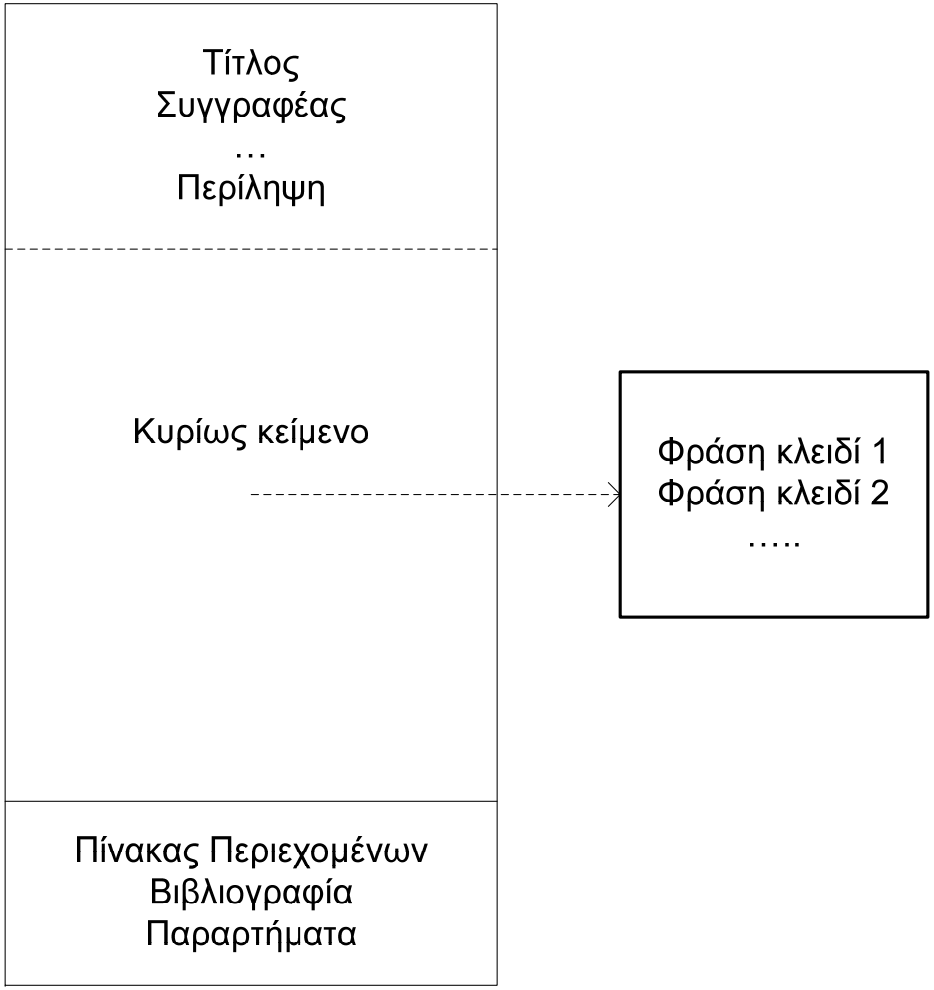
14 διαφορετικά υποπροβλήματα ταξινόμησης, έτσι ώστε κάθε υποπρόβλημα να αφορά την διάκριση μίας κατηγορίας από όλες τις άλλες (one-against-all)

REFERENCE PART - Σύστημα ταξινόμησης με συνδυασμό μοντέλων SVM

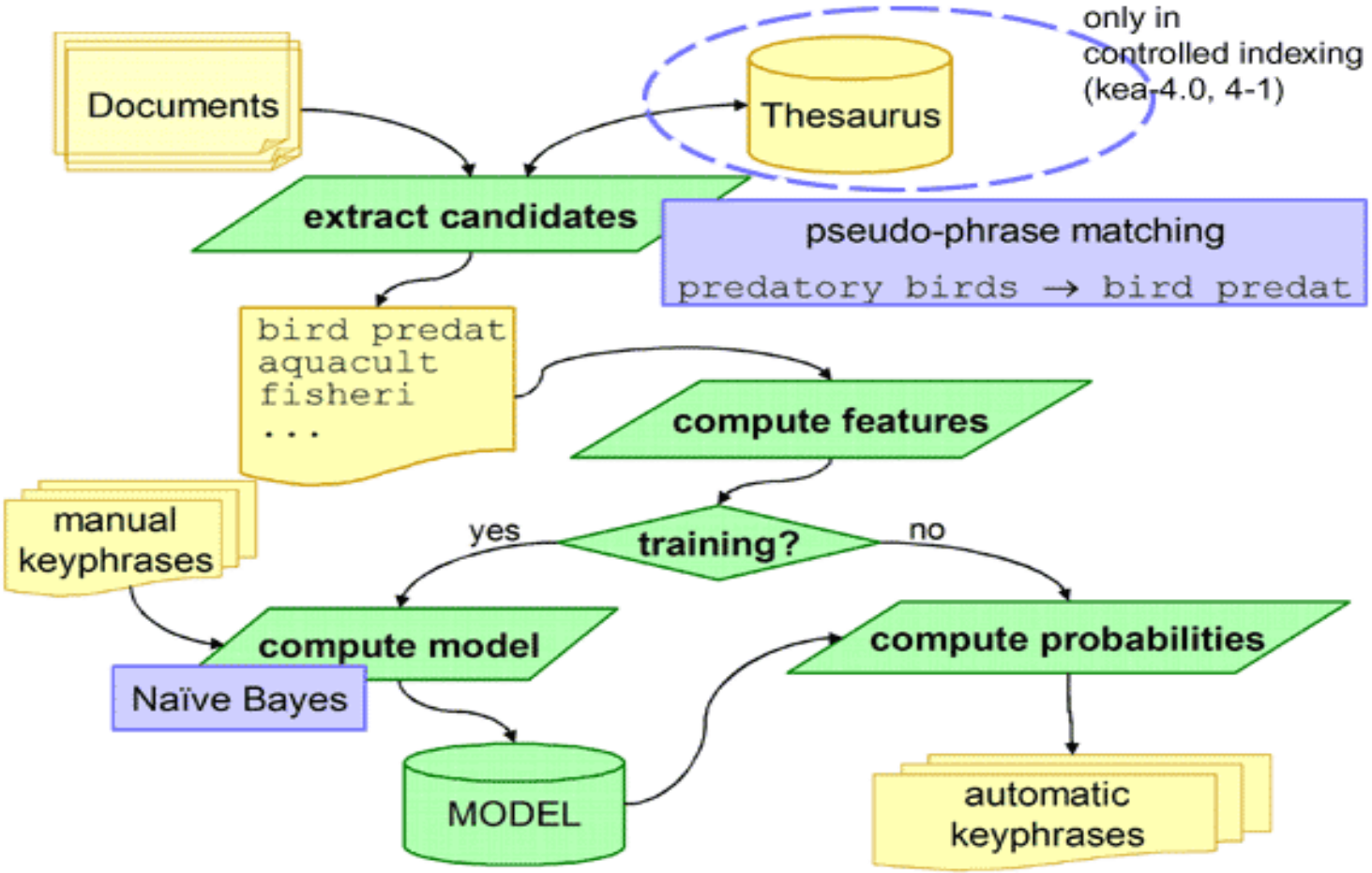
Κάθε τμήμα του REFERENCE PART κατατάσσεται σε μία ή περισσότερες από 12 κατηγορίες:

- Author
- Booktitle
- Date
- Editor
- Institution
- Journal
- Location
- Note
- Pages
- Tech
- Title
- Volume

Αυτόματη Εξαγωγή Λέξεων και Φράσεων Κλειδιών από Κείμενα Δημοσιεύσεων



Αυτόματη εξαγωγή λέξεων και φράσεων κλειδιών με το λογισμικό KEA



Εξαγωγή μεταδεδομένων, βιβλιογραφικών αναφορών και φράσεων-κλειδιών

- Ανάπτυξη ευφυών τεχνικών που κάνουν χρήση λεξικών, μορφολογίας και γενικότερα πολλών ιδιοτήτων του κειμένου και των εγγράφων (σημασιολογική ανάλυση/ανάγνωση)
- Δυνατότητα πολλαπλής αξιοποίησης:
  - διευκόλυνση της online κατάθεσης εγγράφων από τους χρήστες
  - offline βιβλιοθηκονομική αρχειοθέτηση του έγγραφου υλικού
- Μεταδεδομένα: συμπαγής περιγραφή, κατάλληλη για browsing και clustering
- Ενίσχυση της δυνατότητας πλοήγησης σε μεγάλους αριθμούς εγγράφων

Ερευνητική ομάδα ΕΠΙΣΕΥ-ΕΜΠ

- Α.-Γ. Σταφυλοπάτης, Καθ. ΕΜΠ, Επιστ. Υπεύθυνος

- Γ. Σιόλας, Δρ ΗΜΜΥ, Συντονιστής έργου

- Δ. Φροσυνιώτης, Δρ Πληροφορικής
- Μ. Περτσελάκης, Δρ ΗΜΜΥ
- Χ. Πατερίτσας, Δρ ΗΜΜΥ

- Α. Λαναρίδης, ΗΜΜΥ
- Γ. Σπανάκης, ΗΜΜΥ