ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΑΝΟΙΚΤΑ ακαδημαϊκά μαθήματα ΠΠ

# Εισαγωγή στη Βιοπληροφορική

## Ενότητα 9: Text Mining

Μακρής Χρήστος, Τσακαλίδης Αθανάσιος, Ιωάννου Μαρίνα

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ και Πληροφορικής

# Σκοποί ενότητας

- Σκοπός της ενότητας είναι η παρουσίαση του Text Mining στις εφαρμογές της βιοπληροφορικής

# Περιεχόμενα ενότητας

- Εξόρυξη γνώσης από δεδομένα, κείμενα και κείμενα βιολογικού περιεχομένου
- Εφαρμογές
- Συσταδοποίηση
- Αλγόριθμοι

# Βασικές Βιβλιογραφικές Πηγές στις οποίες βασίζονται οι διαφάνειες

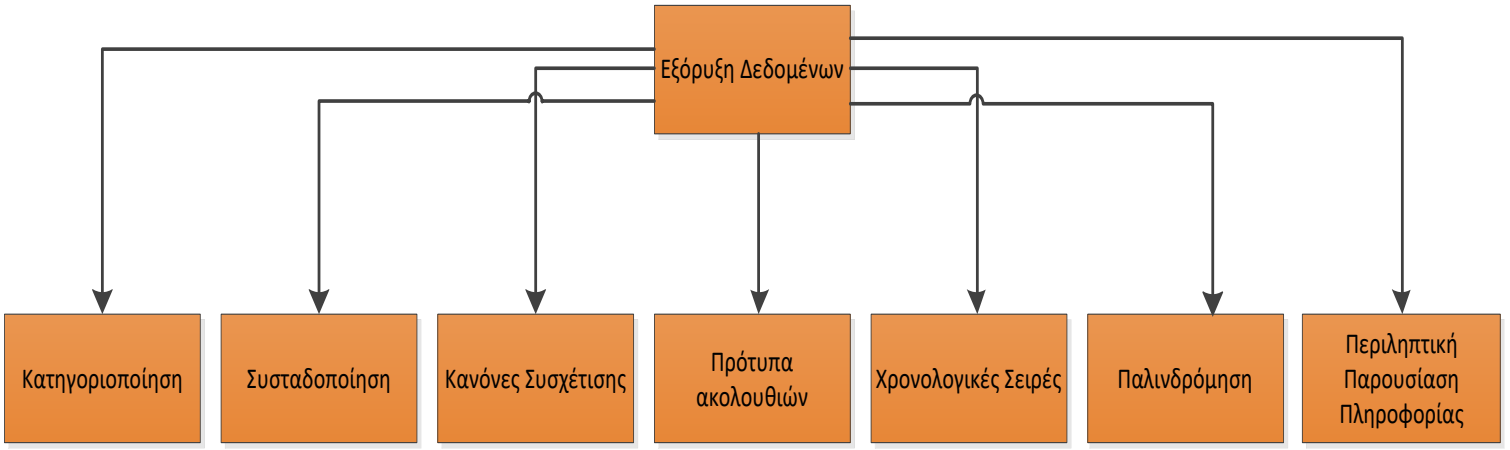
- Zafeiria-Marina Ioannou, Christos Makris, George P. Patrinos, Giannis Tzimas: A set of novel mining tools for efficient biological knowledge discovery. Artif. Intell. Rev. 42(3): 461-478 (2014)

# Text Mining

# Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

- Εξόρυξη Γνώσης
  - Ανακάλυψη γνώσης από βάσεις δεδομένων
  - Τεχνικές για την ανάλυση και εξόρυξη δεδομένων

- Μέθοδοι Εξόρυξης Γνώσης




# Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

- Κατηγοριοποίηση (Classification)
  - Βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου και στην αντιστοίχηση του βάση αυτών των χαρακτηριστικών σε ένα προκαθορισμένο σύνολο κλάσεων.

- Συσταδοποίηση (Clustering)
  - Διαχωρισμός ενός συνόλου δεδομένων σε ένα σύνολο συστάδων (clusters).
  - Διαφοροποιείται από την κατηγοριοποίηση διότι η συσταδοποίηση δεν διαθέτει προκαθορισμένες κατηγορίες.
  - Τα δεδομένα οργανώνονται σε συστάδες με βάση της ομοιότητας που έχουν μεταξύ τους.

# Εξόρυξη Γνώσης από Κείμενα (Text Mining)

- Text Mining
  - Αποσκοπεί στην εξαγωγή χρήσιμης πληροφορίας από πηγές δεδομένων μέσω της αναγνώρισης και της διερεύνησης ενδιαφερόντων προτύπων.
  - Οι πηγές δεδομένων είναι **συλλογές κειμένων**
  - Τα ενδιαφέροντα πρότυπα αναζητούνται σε **μη δομημένα δεδομένα κειμένων**, δηλαδή στα έγγραφα της συλλογής και όχι σε δομημένα δεδομένα Βάσεων δεδομένων.

# Εξόρυξη Γνώσης από Κείμενα Βιολογικού Περιεχομένου (Biomedical Text Mining)

## ❑Πρόβλημα

- Ο όγκος των δημοσιεύσεων βιοϊατρικής έρευνας και οι αντίστοιχες βάσεις βιοϊατρικών δεδομένων, επεκτείνονται και αυξάνονται ραγδαία.

## ❑Στόχος βιοϊατρικής έρευνας

- Η ανακάλυψη γνώσης και η χρησιμοποίηση της στη διάγνωση και θεραπεία.
- Ο ραγδαίος ρυθμός αύξησης των δημοσιεύσεων βιοϊατρικής έρευνας, καθιστά πιο δύσκολη την αναγνώριση σημαντικών συνδέσεων μεταξύ των επιμέρους στοιχείων της βιοϊατρικής γνώσης.

# Εξόρυξη Γνώσης από Κείμενα Βιολογικού Περιεχομένου (Biomedical Text Mining)

- Εξόρυξη κειμένου (Text Mining)
  - Τομέας της επιστήμης των υπολογιστών που μπορεί να βοηθήσει τους ερευνητές στην αντιμετώπιση της πληθώρας πληροφοριών.
- Στόχος
  - Αναγνώριση της πληροφορίας με αποδοτικό τρόπο
  - Αναγνώριση των σχέσεων που υποσκιάζονται από τον μεγάλο όγκο πληροφορίας
  - Εφαρμόζοντας αλγοριθμικές, στατιστικές μεθόδους και μεθόδους διαχείρισης δεδομένων

# Biomedical Text Mining
# Ερευνητικά Πεδία

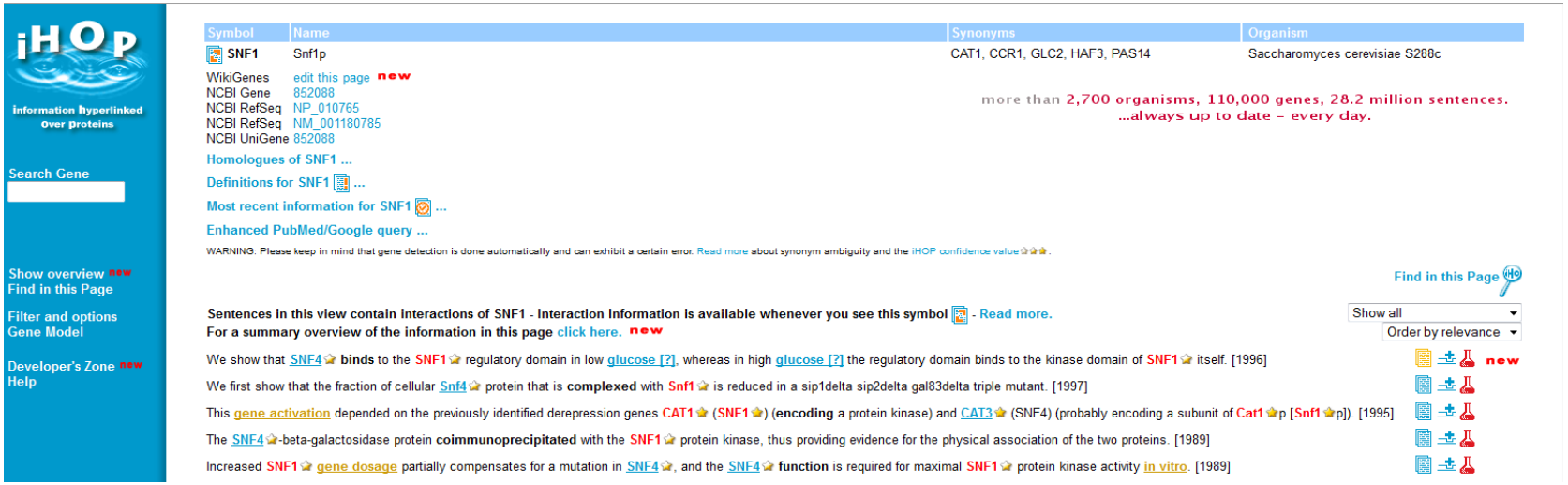
- **Αναγνώριση Ονοματικών Οντοτήτων (Named Entity Recognition)**
  - Όλα τα ονόματα των φαρμάκων μέσα σε μια συλλογή άρθρων, ή όλα τα ονόματα γονιδίων.
- **Κατηγοριοποίηση Κειμένων (Text Classification)**
  - Καθορίζει αυτοματοποιημένο τρόπο εάν ένα κείμενο ή μέρος ενός κειμένου έχει συγκεκριμένα χαρακτηριστικά
- **Εξόρυξη Συσχετίσεων**
  - συσχετίσεις ανάμεσα σε γονίδια και πρωτεΐνες

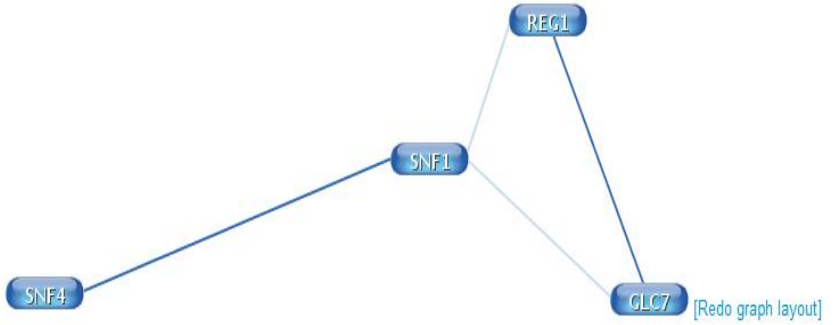
# Biomedical Text Mining
# Web Εφαρμογές

- **iHop**
  - Ανακτά τις προτάσεις που περιέχουν συγκεκριμένα γονίδια, επισημαίνει τις βιοϊατρικές οντότητες στα γονίδια και παρέχει γραφήματα των συσχετίσεων μεταξύ όλων των οντοτήτων.

- Το iHop παρέχει στους ερευνητές:
  1. Φιλτράρισμα και ταξινόμηση των ανακτηθέντων προτάσεων που ταιριάζουν στο δοθέν γονίδιο ή πρωτεΐνη με βάση την σπουδαιότητά τους, το Impact factor, την ημερομηνία δημοσίευσης και σύνταξης
  2. Εξερεύνηση ενός δικτύου αλληλεπιδράσεων γονιδίων και πρωτεϊνών

# Biomedical Text Mining Web Εφαρμογές



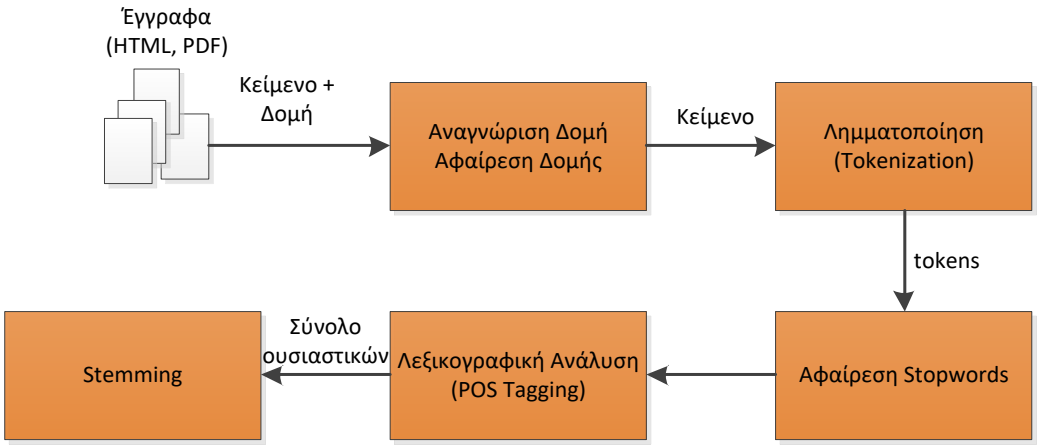



# Text Mining

- Τα βασικά βήματα για την ανάλυση κειμένων είναι:
  - Προεπεξεργασία Κειμένων
  - Αναπαράσταση Κειμένων
  - Εξαγωγή Χαρακτηριστικών Γνωρισμάτων των κειμένων

# Προεπεξεργασία Κειμένων

- Αναγνώριση και αφαίρεση της δομής των κειμένων
- Λημματοποίηση (Tokenization)
- Αφαίρεση των stopwords
- Λεξικογραφική Ανάλυση (POS Tagging)
- Αποκατάληξη (Stemming)



# Προεπεξεργασία Κειμένων (2)

- **Αφαίρεση Δομής**
  - πχ. Μετατροπή των PDF και HTML αρχείων σε απλό κείμενο .txt
- **Λημματοποίηση (Tokenization)**
  - Διαχωρισμός των προτάσεων σε ξεχωριστούς όρους (tokens) που μπορεί να είναι λέξεις ή σημεία στίξης ή αριθμοί.
- **Αφαίρεση Stopwords:**
  - Σύγκριση κάθε όρου με μια γνωστή συλλογή από stopwords.
- **Λεξικογραφική Ανάλυση (POS Tagging)**
  - αναγνώριση του μέρους του λόγου που ανήκει η κάθε λέξη, δηλαδή ουσιαστικό, ρήμα, επίθετο κλπ.
- **Επιλογή των ουσιαστικών**
  - Τα ουσιαστικά επιφέρουν τη σημαντικότερη πληροφορία των κειμένων.

# Διανυσματικό Μοντέλο (Vector Space Model)

- Αναπαριστούμε τα κείμενα σε μια μορφή που να είναι επεξεργάσιμη.
- Η πιο γνωστή μέθοδος αναπαράστασης κειμένων είναι η διανυσματική αναπαράσταση.
- Κάθε κείμενο και κάθε ερώτημα αναπαρίσταται ως ένα **διάνυσμα** m όρων, όπου m είναι ο αριθμός των μοναδικών όρων (unique terms) της συλλογής.
- Για κάθε όρο υπολογίζουμε το βάρος.

# TF-IDF

- Η πιο γνωστή μέθοδος απόδοσης βάρους

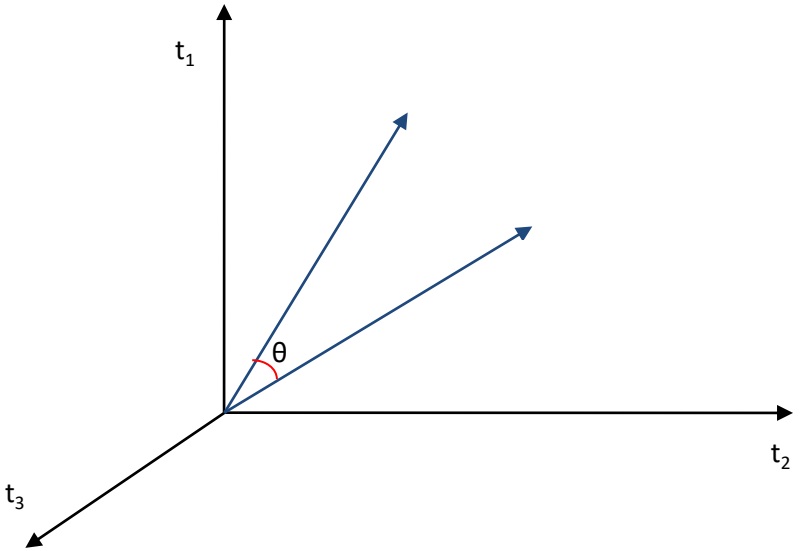
- Το TF-IDF αποτελείται από τις εξής ποσότητες:
  - TF είναι η συχνότητα εμφάνισης ενός όρου σε ένα κείμενο.
  - IDF αποτελεί ένα βάρος που δηλώνει τη σημαντικότητα ενός όρου στο κείμενο, σε σχέση με ολόκληρη τη συλλογή κειμένων.
  - Το τελικό βάρος TF-IDF προκύπτει από τον πολλαπλασιασμό των TF και IDF.

- TF-IDF έχει μεγάλη τιμή για έναν όρο και επομένως είναι σημαντικός για ένα κείμενο, όταν ο όρος εμφανίζεται συχνά σε ένα κείμενο και σπάνια στα υπόλοιπα κείμενα της συλλογής.

# Ομοιότητα Συνημίτονου

- Η πιο γνωστή μέθοδος υπολογισμού της ομοιότητας, βασίζεται στο συνημίτονο της εμπεριεχόμενης γωνίας των δυο διανυσμάτων.

$$\mathrm{sim}(d_j, q) = \frac{(\vec{d_j} \bullet \vec{q})}{|\vec{d_j}|\,|\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij} \times w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}\sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$



# Latent Semantic Indexing

- Το Latent Semantic Indexing (LSI) είναι μια σημαντική τεχνική δεικτοδότησης και ανάκτησης.

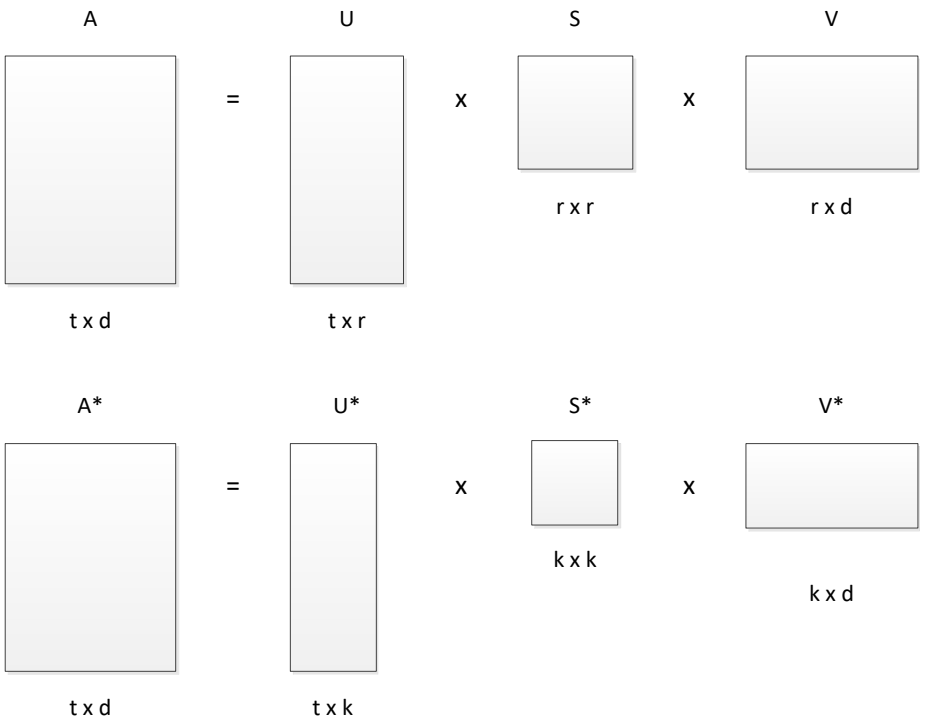
- Χρησιμοποιεί τη μέθοδο Singular Value Decomposition (SVD) για να ανακαλύψει πρότυπα και συσχετίσεις μεταξύ των όρων και των εννοιών που περιέχονται σε μη δομημένες συλλογές κειμένων.

- Ο αρχικός πίνακας A, αναλύεται σε ένα γινόμενο τριών απλών πινάκων:

$$A=USV^T$$

- Κρατάμε μόνο τις k μεγαλύτερες ιδιοτιμές σύμφωνα με ένα κατώφλι και παράγουμε τους πίνακες $U_k$, $S_k$, και $V_k$.

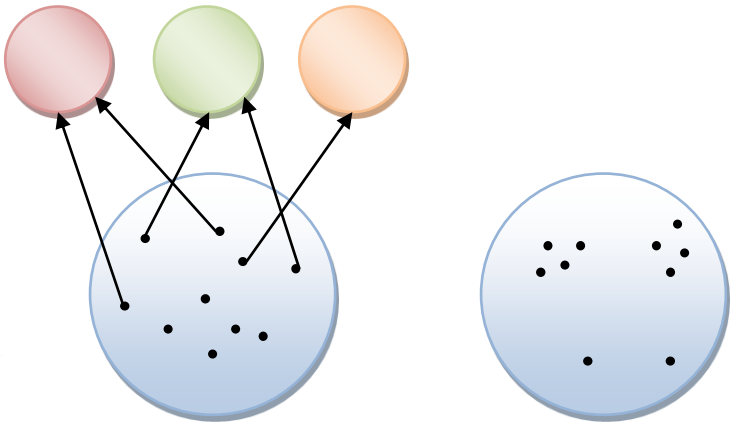
# Singular Value Decomposition (SVD)




- Μειώνοντας το διανυσματικό χώρο σε k διαστάσεις, εξαλείφεται ο θόρυβος που προκαλεί κακή απόδοση στα συστήματα ανάκτησης πληροφορίας

# Συσταδοποίηση (Clustering)

- Το πρόβλημα της συσταδοποίησης σχετίζεται με την τμηματοποίηση (partitioning, clustering) ενός συνόλου δεδομένων σε συστάδες, έτσι ώστε τα στοιχεία που ανήκουν σε μία συστάδα να είναι περισσότερο όμοια μεταξύ τους από ότι είναι με τα στοιχεία των άλλων συστάδων.
- Δεν υπάρχουν προκαθορισμένες κατηγορίες ούτε κάποια άλλη προηγούμενη γνώση σχετικά με την σχέση μεταξύ των στοιχείων.
- Αντίθετα, η κατηγοριοποίηση είναι η διαδικασία με την οποία ένα σύνολο αντικειμένων αντιστοιχίζεται σε ένα σύνολο προκαθορισμένων κατηγοριών εξετάζοντας τα χαρακτηριστικά κάθε αντικειμένου.



# Μέθοδοι Συσταδοποίησης

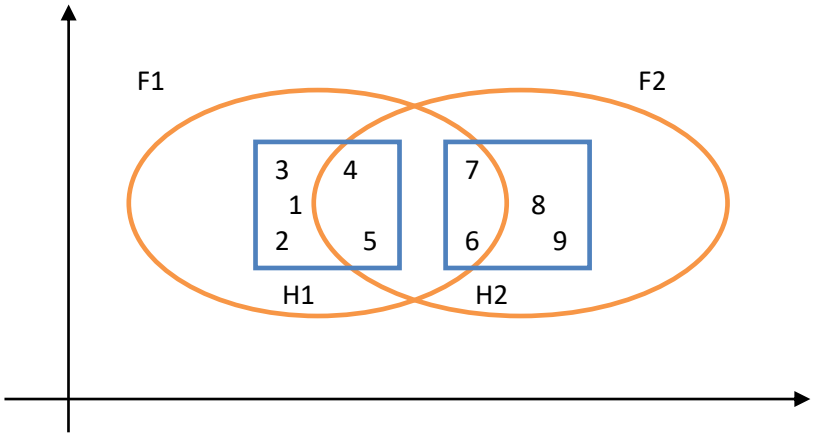
- Διαιρετική Συσταδοποίηση (Partitional Clustering)
- Ασαφής Συσταδοποίηση (Fuzzy Clustering)
- Μη ασαφής Συσταδοποίησης (Crisp Clustering)
- Συσταδοποίηση με δίκτυα Kohonen
- Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

# Ασαφής Συσταδοποίηση (Fuzzy Clustering)

- Hard Clustering
  - Τα στοιχεία διαχωρίζονται σε μη ασαφείς συστάδες (crisp clusters), όπου κάθε στοιχείο ανήκει σε ακριβώς μία συστάδα. Με τον τρόπο αυτό παράγονται συστάδες που είναι μη επικαλυπτόμενες (crisp clustering).
- Αλγόριθμοι Ασαφής Συσταδοποίησης (Fuzzy Clustering)
  - Θεωρούν ότι ένα στοιχείο μπορεί να ανήκει σε περισσότερες από μια συστάδες ορίζοντας ένα βαθμό συμμετοχής κάθε στοιχείου σε κάθε συστάδα.
  - Η τιμή του βαθμού συμμετοχής ενός στοιχείου i στην συστάδα j, δείχνει την πιθανότητα να ανήκει το στοιχείο αυτό στην συγκεκριμένη συστάδα.

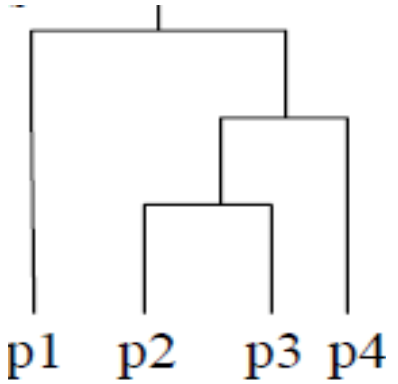
# Ασαφής Συσταδοποίηση (Fuzzy Clustering)

- Το αποτέλεσμα της Fuzzy Clustering τεχνικής μπορεί να μετατραπεί σε Hard Clustering.
- Κάθε στοιχείο ανήκει σε μία μόνο συστάδα, στη συστάδα στην οποία έχει τον **μεγαλύτερο βαθμό συμμετοχής**.



# Ιεραρχικοί Αλγόριθμοι (Hierarchical Algorithms)

- Παράγουν μια ακολουθία διχοτομήσεων ή συγχωνεύσεων, η οποία μπορεί να αναπαρασταθεί ως ένα δέντρο, το οποίο ονομάζεται δενδρόγραμμα.
- Κάθε επίπεδο του δενδρογράμματος απεικονίζει τη συγχώνευση δύο συστάδων του χαμηλότερου επιπέδου.



# Ιεραρχικοί Αλγόριθμοι (Hierarchical Algorithms)

- Συσσωρευτικοί Ιεραρχικοί Αλγόριθμοι
  - αρχικά κάθε στοιχείο ως μια ξεχωριστή συστάδα.
  - Σε κάθε βήμα, συγχωνεύουν το ζεύγος συστάδων με την μεγαλύτερη ομοιότητα ή το πλησιέστερο ζεύγος συστάδων.
  - Για να βρεθεί η ομοιότητα ή η απόσταση δύο συστάδων απαιτείται ο προσδιορισμός ενός κριτηρίου.
- Διαιρετικοί Ιεραρχικοί Αλγόριθμοι
  - αρχικά όλα τα στοιχεία ως μια μοναδική συστάδα
  - σε κάθε βήμα διαχωρίζουν μια συστάδα έως ότου καταλήξουμε σε ένα σύνολο συστάδων, όπου κάθε μια αποτελείται από ένα μόνο στοιχείο.
  - Σε αυτή την περίπτωση, θα πρέπει να ορίσουμε ποια συστάδα θα διαχωριστεί σε κάθε βήμα, καθώς και τον τρόπο διαχωρισμού της.

# Απλός Συσσωρευτικός Ιεραρχικός Αλγόριθμος

- Υπολογίζουμε την ομοιότητα όλων των ζευγών συστάδων (δηλαδή υπολογίζουμε έναν πίνακα ομοιότητας, όπου το στοιχείο (i,j) ορίζει την ομοιότητα των συστάδων i και j.

- Συγχωνεύουμε τις δύο πιο όμοιες (πιο κοντινές) συστάδες.

- Ανανεώνουμε τον πίνακα ομοιότητας για να απεικονίζει την ομοιότητα μεταξύ της νέας συστάδας και των αρχικών συστάδων.

- Επαναλαμβάνουμε τα βήματα 2 και 3  έως ότου μείνει μια μόνο συστάδα.

# Κριτήρια Ομοιότητας

- **Intra-Cluster Similarity Technique (IST)**
  - Ομοιότητα όλων των κειμένων της συστάδας με το κέντρο (centroid) της συστάδας.
  - Η επιλογή του ζεύγους συστάδων που θα συγχωνευθεί πραγματοποιείται καθορίζοντας ποιο ζεύγος συστάδων θα οδηγήσει στην μικρότερη μείωση ομοιότητας.

$$Sim(X) = \sum_{d \in X} cosine(d, c)$$

- **Centroid Similarity Technique (CST)**
  - Ορίζει την ομοιότητα των δυο συστάδων, ως την ομοιότητα του συνημιτόνου μεταξύ των κέντρων δύο συστάδων.

- **UPGMA**
  - Η ομοιότητα των συστάδων ορίζεται ως εξής:

$$similarity(Cluster1, Cluster2) = \frac{\sum_{\substack{d_1 \in cluster1 \\ d_2 \in cluster2}} cosine(d_1, d_2)}{size(cluster1) * size(cluster2)}$$

# Αλγόριθμος K-means

- Ο Αλγόριθμος ξεκινά αρχικοποιώντας με τυχαίο τρόπο τα κέντρα των συστάδων.
- Στη συνέχεια, αναθέτει κάθε στοιχείο του συνόλου δεδομένων στη συστάδα της οποίας το κέντρο βρίσκεται πιο κοντά και ξαναυπολογίζει τα νέα κέντρα που προκύπτουν.
- Τα νέα κέντρα των συστάδων υπολογίζονται χρησιμοποιώντας τον μέσο όρο των σημείων της κάθε συστάδας.
- Η διαδικασία αυτή επαναλαμβάνεται έως ότου τα να κέντρα των συστάδων σταματήσουν να αλλάζουν.

# K-means με Διχοτόμηση (Bisecting K-means)

- Παραλλαγή του αλγόριθμου K-means
- Αρχικά, αντιστοιχεί όλα τα αντικείμενα-δεδομένα σε μια συστάδα.
- Επανάληψη 3 βημάτων έως ότου επιτύχουμε τον επιθυμητό αριθμό συστάδων
  - Επιλογή της συστάδας που θα διασπαστεί.
  - Διαχωρισμός αυτής της συστάδας σε δυο υπο-συστάδες χρησιμοποιώντας τον βασικό Αλγόριθμο K-means.
  - Bisecting βήμα το οποίο επαναλαμβάνεται για έναν αριθμό επαναλήψεων, προκειμένου να επιλέξουμε το διαχωρισμό με την υψηλότερη συνολική ομοιότητα.

# Spherical K-means

- **K-means**
  - Χρησιμοποιεί την ευκλείδεια απόσταση, ωστόσο αυτή η μετρική απόστασης είναι συχνά ακατάλλλη για την συσταδοποίηση κειμένων.
- **Spherical K-means**
  - Χρησιμοποιεί την ομοιότητα συνημιτόνου, η οποία υπολογίζει το συνημίτονο της εσωτερικής γωνίας των διανυσμάτων των κειμένων.
  - Τα διανύσματα βρίσκονται πάνω στην μοναδιαία σφαίρα.

# Αλγόριθμος

Ο αλγόριθμος της εφαρμογής βασίζεται σε:

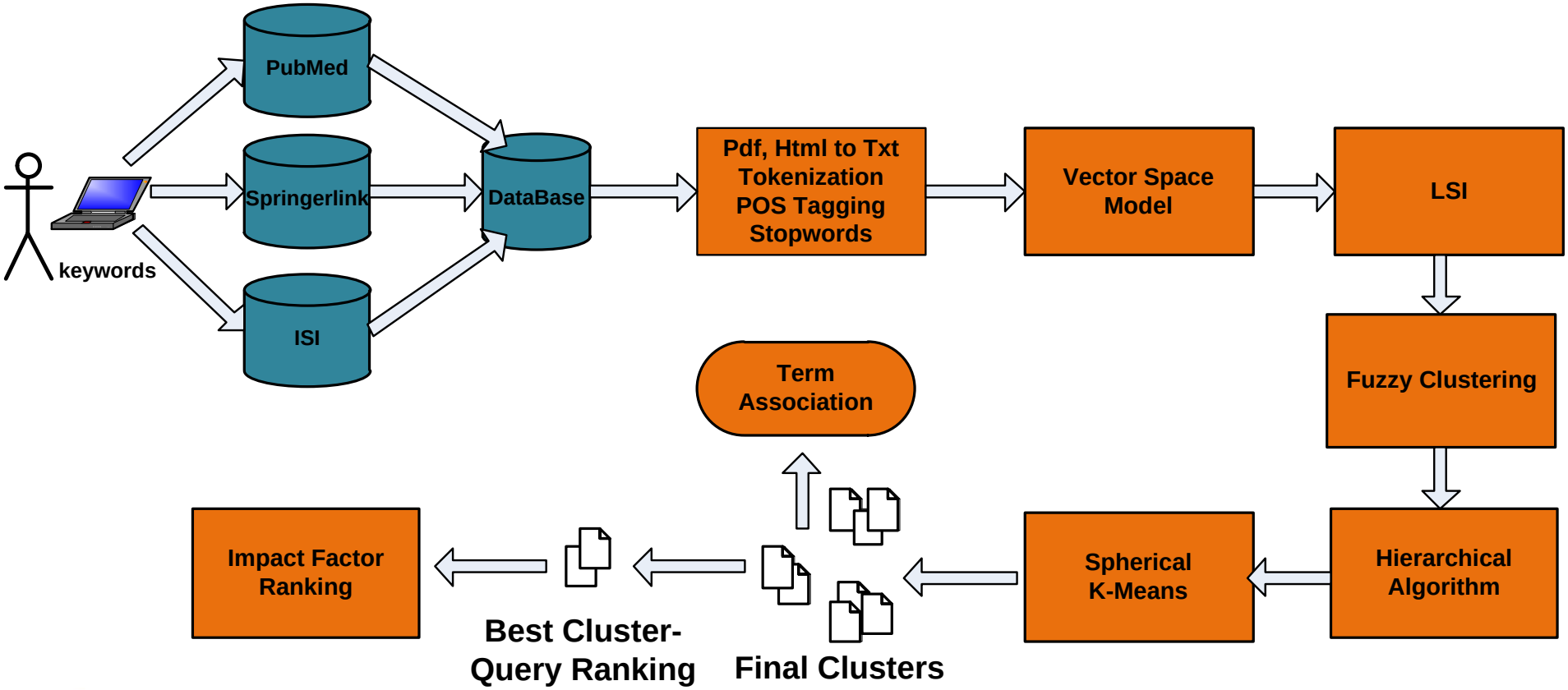
- **Τεχνικές Συσταδοποίησης (Clustering)**

  Όπως Ιεραρχικός Αλγόριθμος (Hierarchical Algorithm),

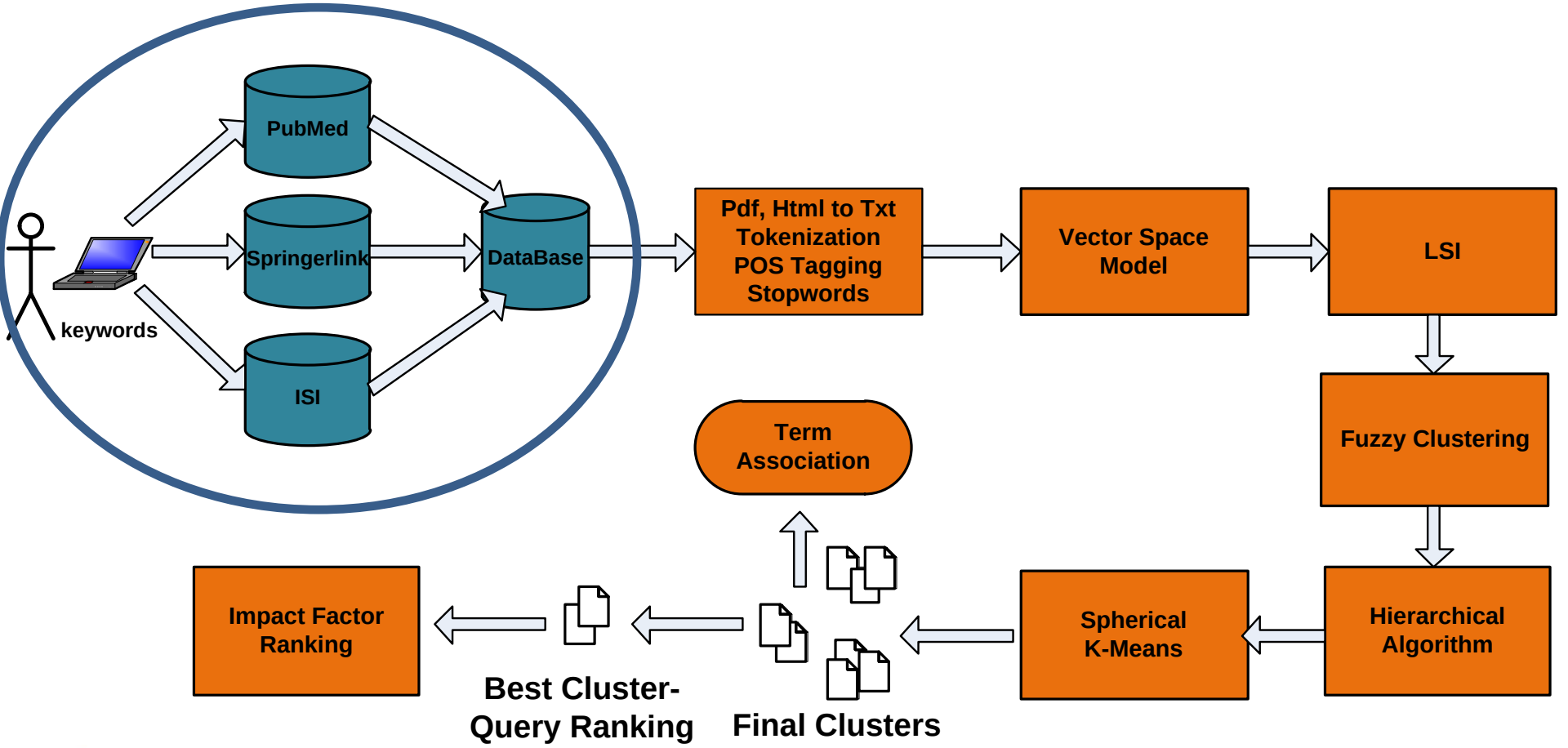
  Spherical K-means Αλγόριθμος.

- **Τελική ταξινόμηση** με βάση το Impact Factor των κειμένων που ανακτήθηκαν.

# Βασικά Βήματα



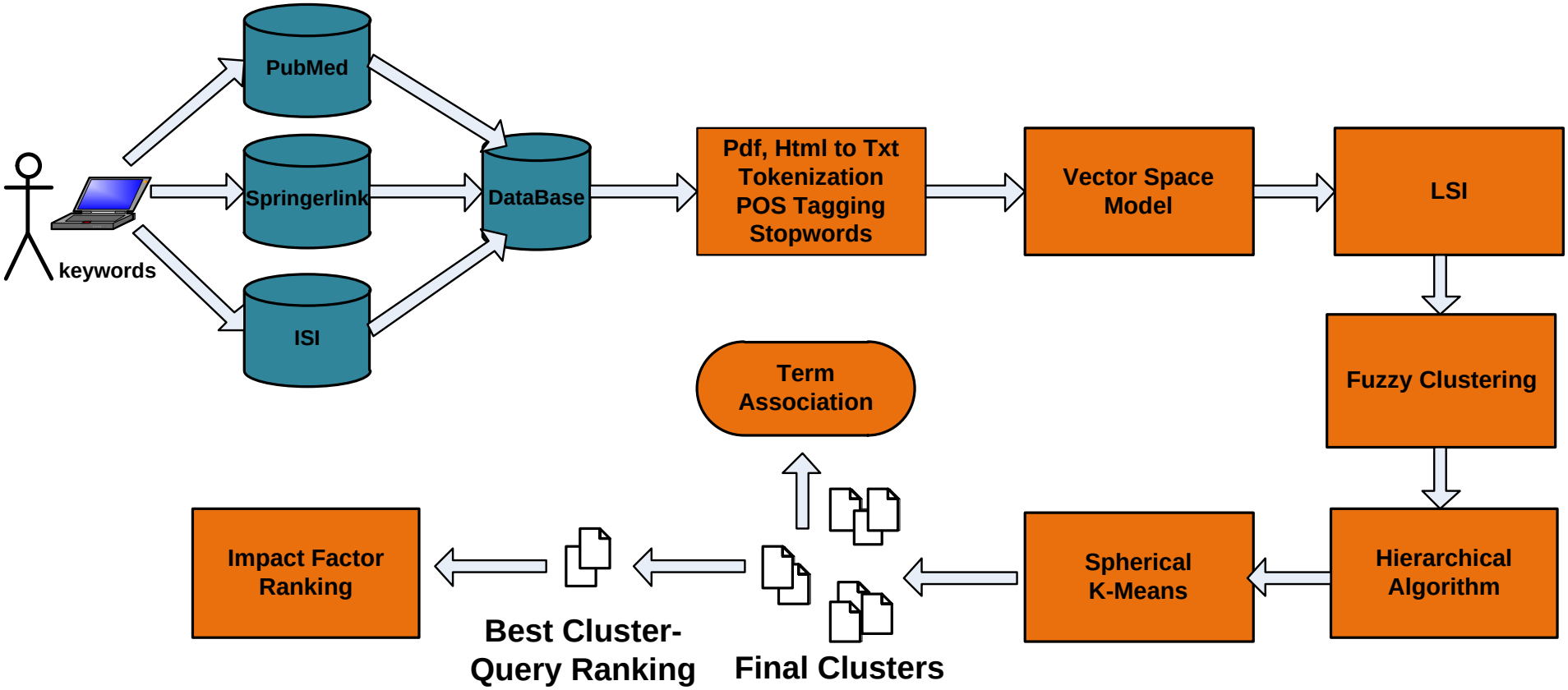
# Βήμα 1


PubMed
Springerlink
ISI
DataBase
keywords
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

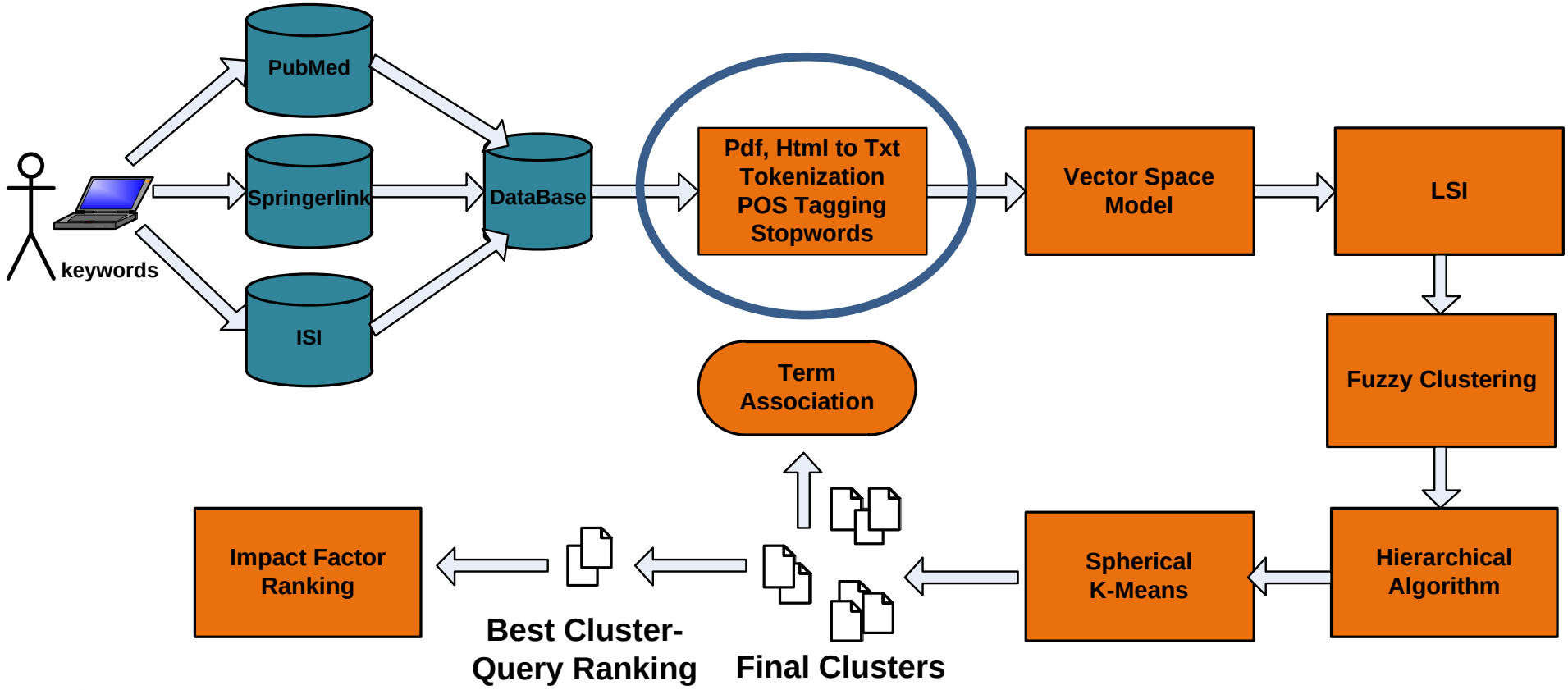
# Βάση Δεδομένων του Συστήματος

- Ο χρήστης δίνει τις λέξεις-κλειδιά (keywords) στη φόρμα αναζήτησης.
- Οι λέξεις-κλειδιά δίνονται στη συνέχεια ως ερώτημα στις βάσεις δεδομένων του **PubMed** και του **Springerlink**.
- Αποθήκευση των κορυφαίων αποτελεσμάτων που επιστρέφονται στη βάση δεδομένων του συστήματος.
- Εύρεση του **Impact Factor** των περιοδικών από τη βάση δεδομένων του **ISI Web of Knowledge**.
- Στη Βάση Δεδομένων αποθηκεύονται επίσης πληροφορίες για κάθε άρθρο όπως (Url του περιοδικού και της περίληψης, τίτλος περιοδικού, ονόματα συγγραφέων κλπ.)

# Βήμα 2


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

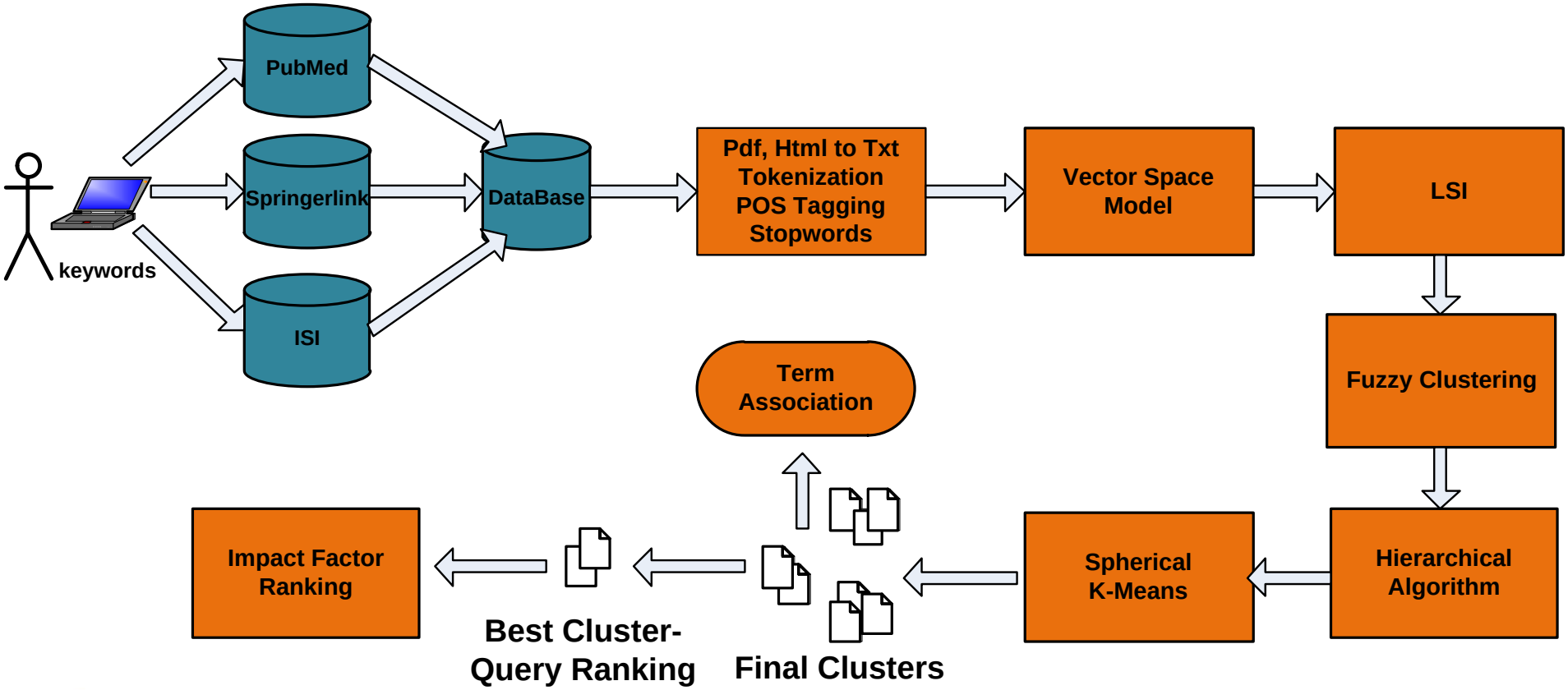
# Βήμα 2


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space Model
LSI
Fuzzy Clustering
Hierarchical Algorithm
Spherical K-Means
Term Association
Final Clusters
Best Cluster-Query Ranking
Impact Factor Ranking

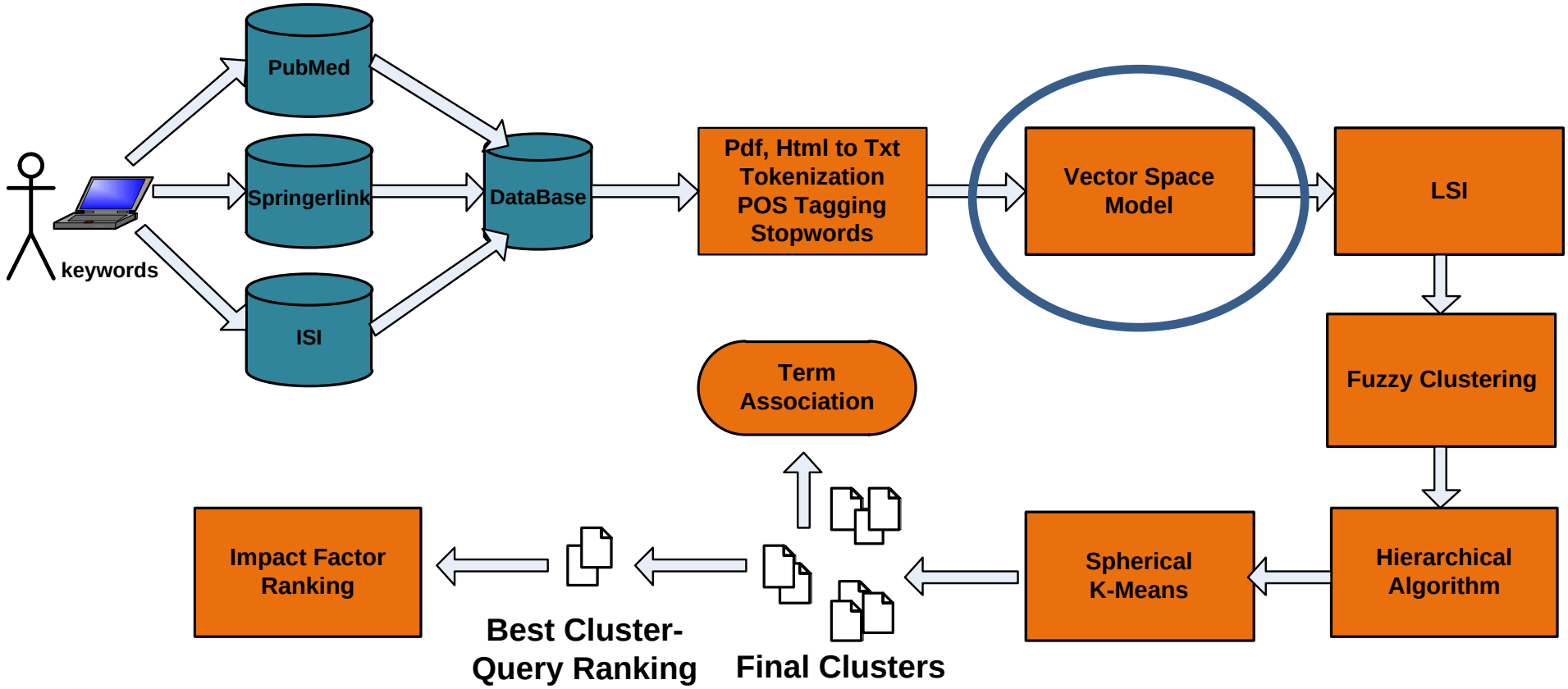
# Προεπεξεργασία Κειμένων

- **Αφαίρεση Δομής:** Μετατροπή των PDF και HTML αρχείων σε απλό κείμενο .txt
- **Λημματοποίηση (Tokenization)**: Διαχωρισμός των προτάσεων σε ξεχωριστούς όρους (tokens) που μπορεί να είναι λέξεις ή σημεία στίξης ή αριθμοί.
- **Αφαίρεση Stopwords:** Σύγκριση κάθε όρου με μια γνωστή συλλογή από stopwords.
- **Λεξικογραφική Ανάλυση (POS Tagging):** αναγνώριση του μέρους του λόγου που ανήκει η κάθε λέξη, δηλαδή ουσιαστικό, ρήμα, επίθετο κλπ. Χρησιμοποιήσαμε τον GENIA Tagger, ο οποίος είναι εξειδικευμένος στην ανάλυση κειμένων βιολογικού περιεχομένου.
- **Επιλογή των ουσιαστικών:** Τα ουσιαστικά επιφέρουν τη σημαντικότερη πληροφορία των κειμένων.

# Βήμα 3


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Βήμα 3


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Διανυσματικό Μοντέλο (Vector Space Model)

- Κάθε κείμενο και κάθε ερώτημα αναπαρίσταται ως ένα **διάνυσμα** m όρων, όπου m είναι ο αριθμός των μοναδικών όρων (unique terms) της συλλογής.

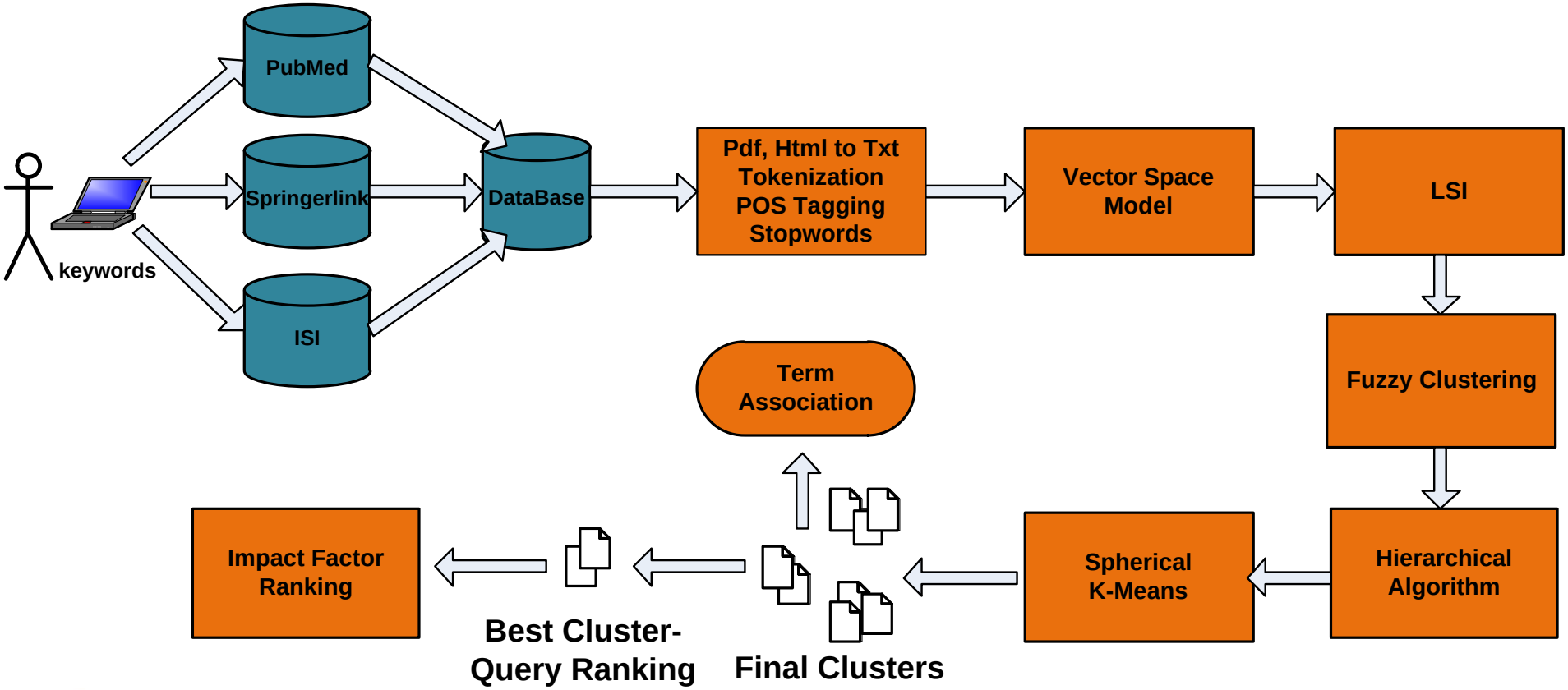
$$q_i = (w_1, w_2, \ldots, w_m)$$

- Για κάθε όρο υπολογίζουμε το βάρος **TF-IDF** που αντιστοιχεί για κάθε κείμενο.

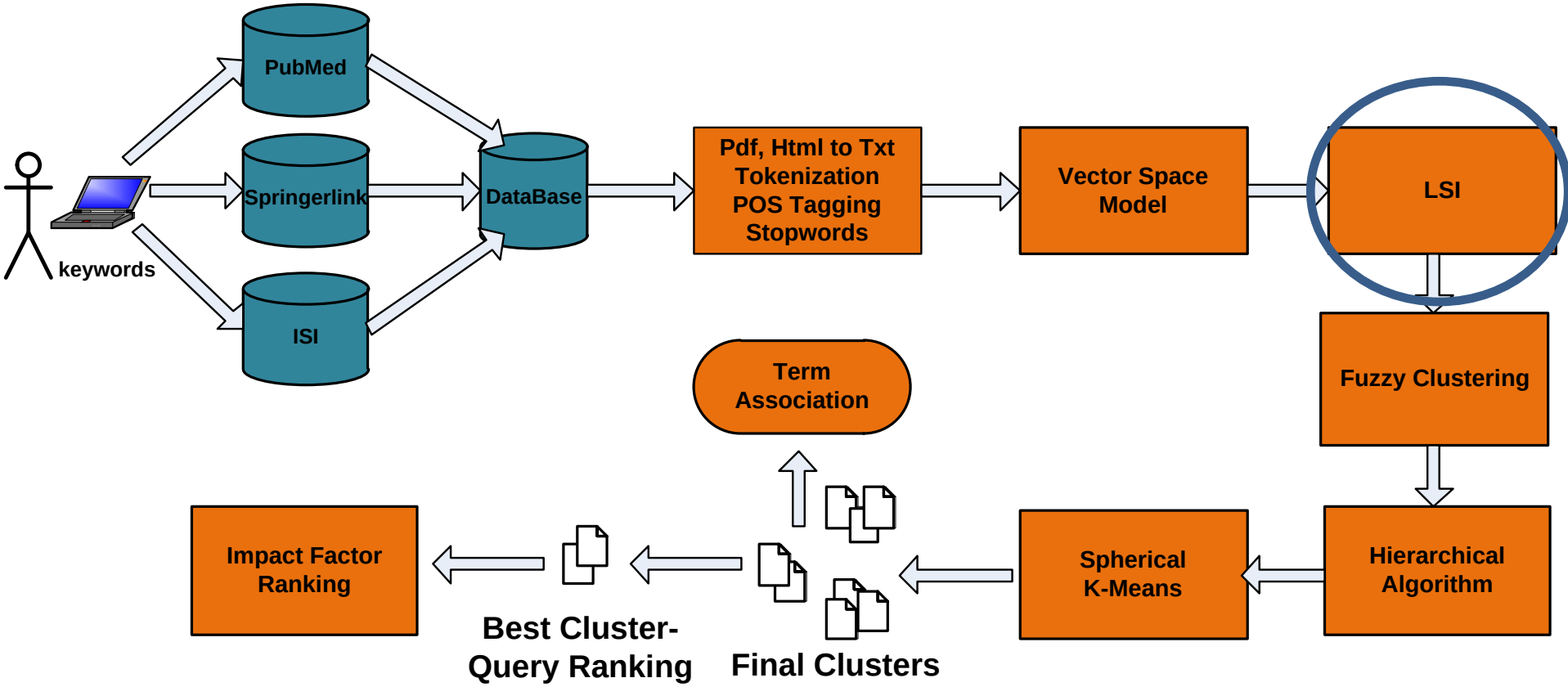
- Στο σχήμα TF-IDF, η **συχνότητα εμφάνισης** TF του όρου στο κείμενο πολλαπλασιάζεται με την **αντίστροφη συχνότητα** (IDF - inverse document frequency) του όρου αυτού στα κείμενα της συλλογής.

- **Αποτέλεσμα**: Η δημιουργία ενός πίνακα A mxn, όπου m είναι ο αριθμός των μοναδικών όρων και n ο αριθμός των κειμένων.

# Βήμα 4


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Βήμα 4


PubMed
Springerlink
ISI
DataBase
keywords
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

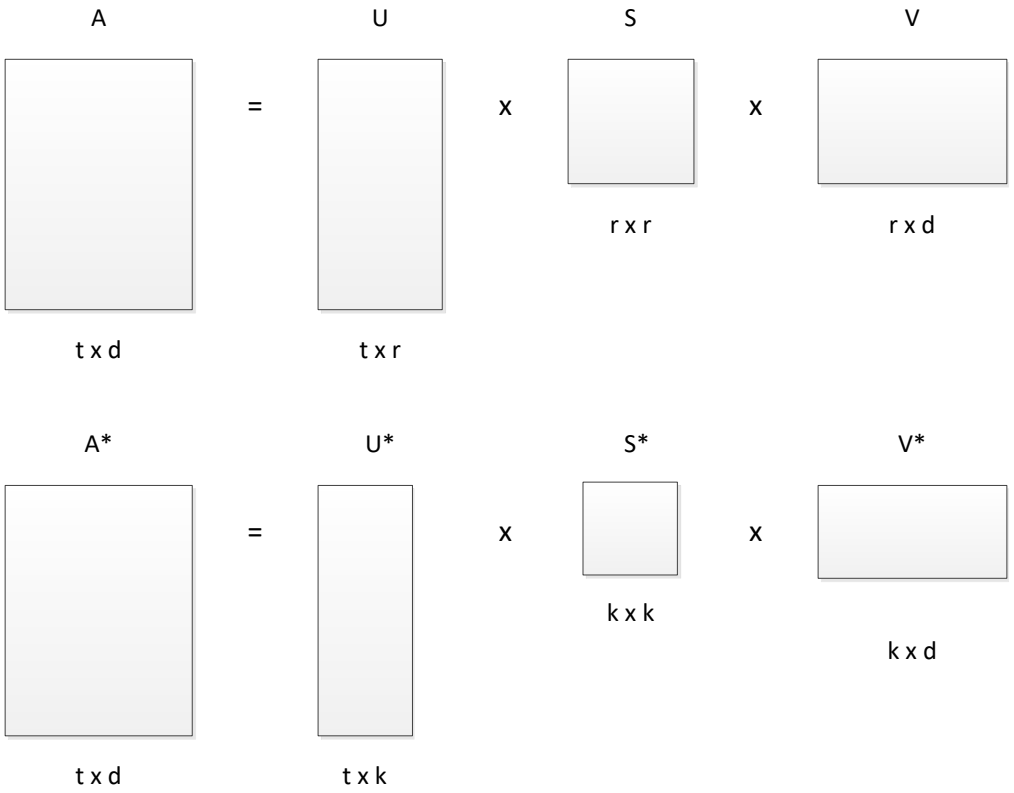
# Λανθάνουσα Σημασιολογική Δεικτοδότηση (Latent Semantic Indexing-LSI)

- Το Latent Semantic Indexing (LSI) είναι μια σημαντική τεχνική δεικτοδότησης και ανάκτησης.
- Χρησιμοποιεί τη μέθοδο Singular Value Decomposition (SVD) για να ανακαλύψει πρότυπα και συσχετίσεις μεταξύ των όρων και των εννοιών που περιέχονται σε μη δομημένες συλλογές κειμένων.
- Ο αρχικός πίνακας A, αναλύεται σε ένα γινόμενο τριών απλών πινάκων:

$$A=USV^T$$

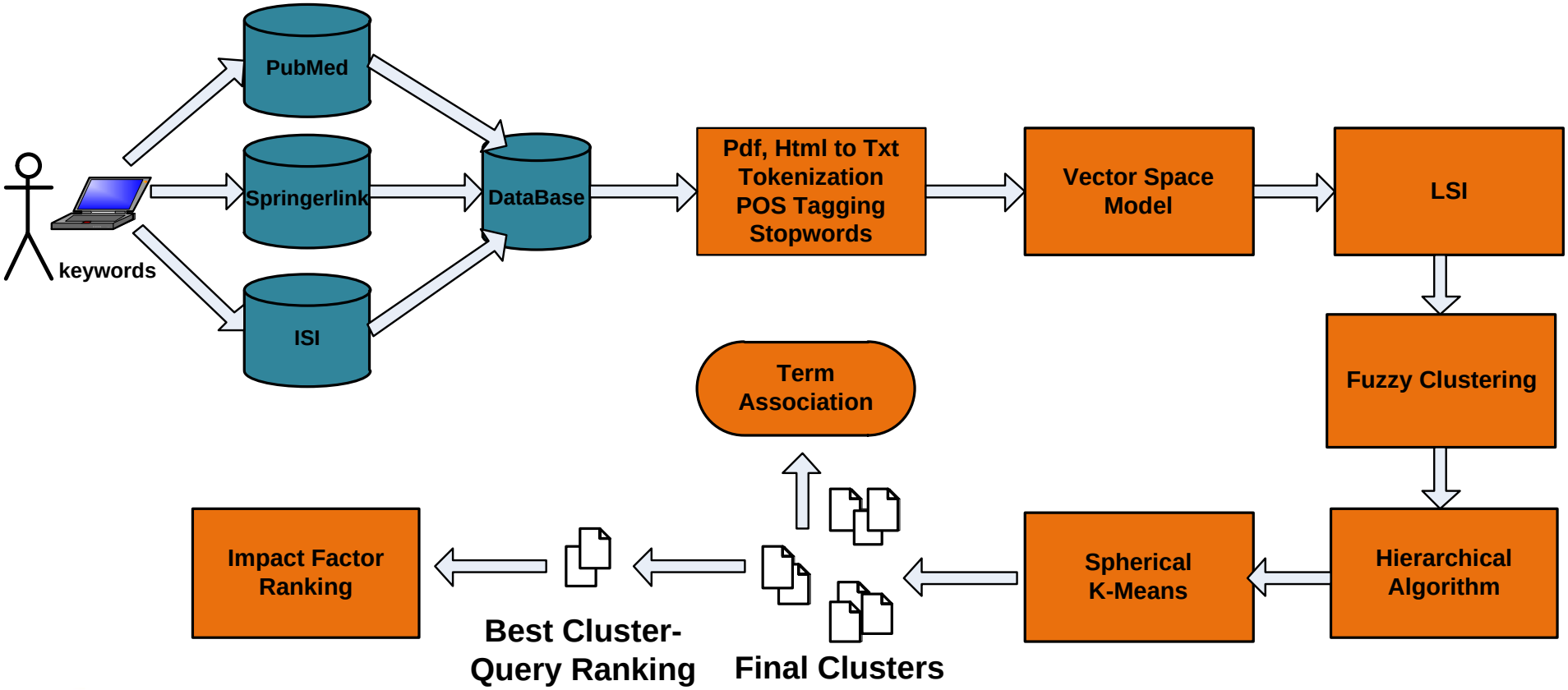
- Κρατάμε μόνο τις k μεγαλύτερες ιδιάζουσες τιμές σύμφωνα με ένα κατώφλι και παράγουμε τους πίνακες $U_k$, $S_k$, και $V_k$.

# Singular Value Decomposition (SVD)

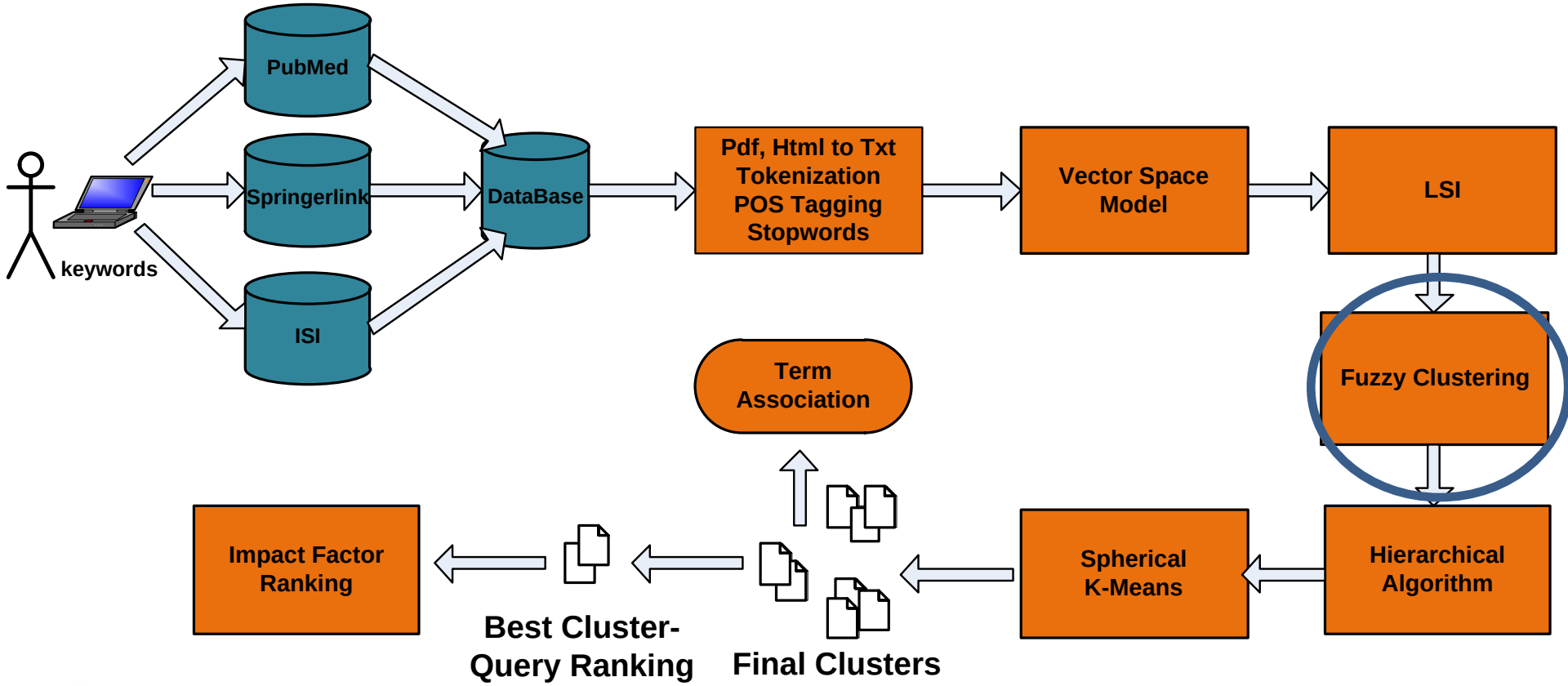



- Μειώνοντας το διανυσματικό χώρο σε k διαστάσεις, εξαλείφεται ο θόρυβος που προκαλεί κακή απόδοση στα συστήματα ανάκτησης πληροφορίας

# Βήμα 5


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

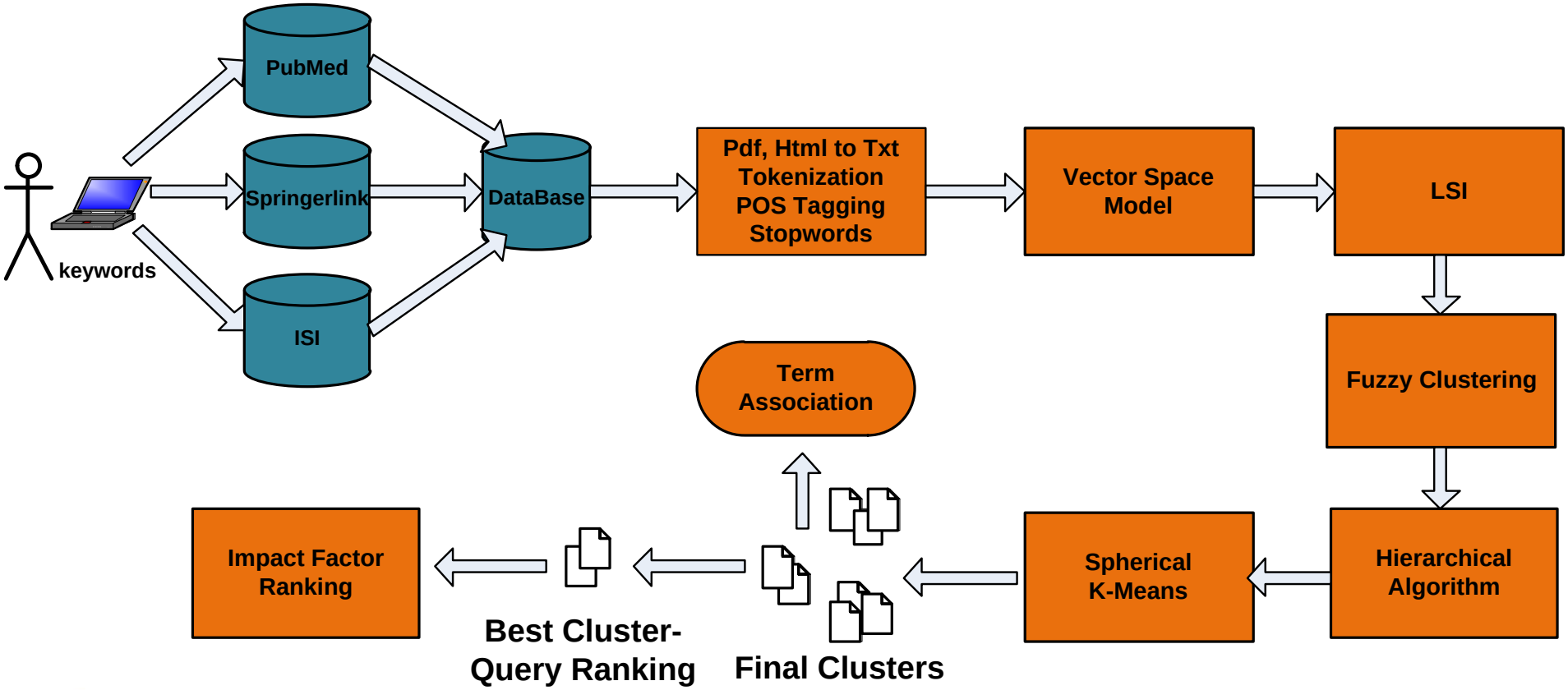
# Βήμα 5


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

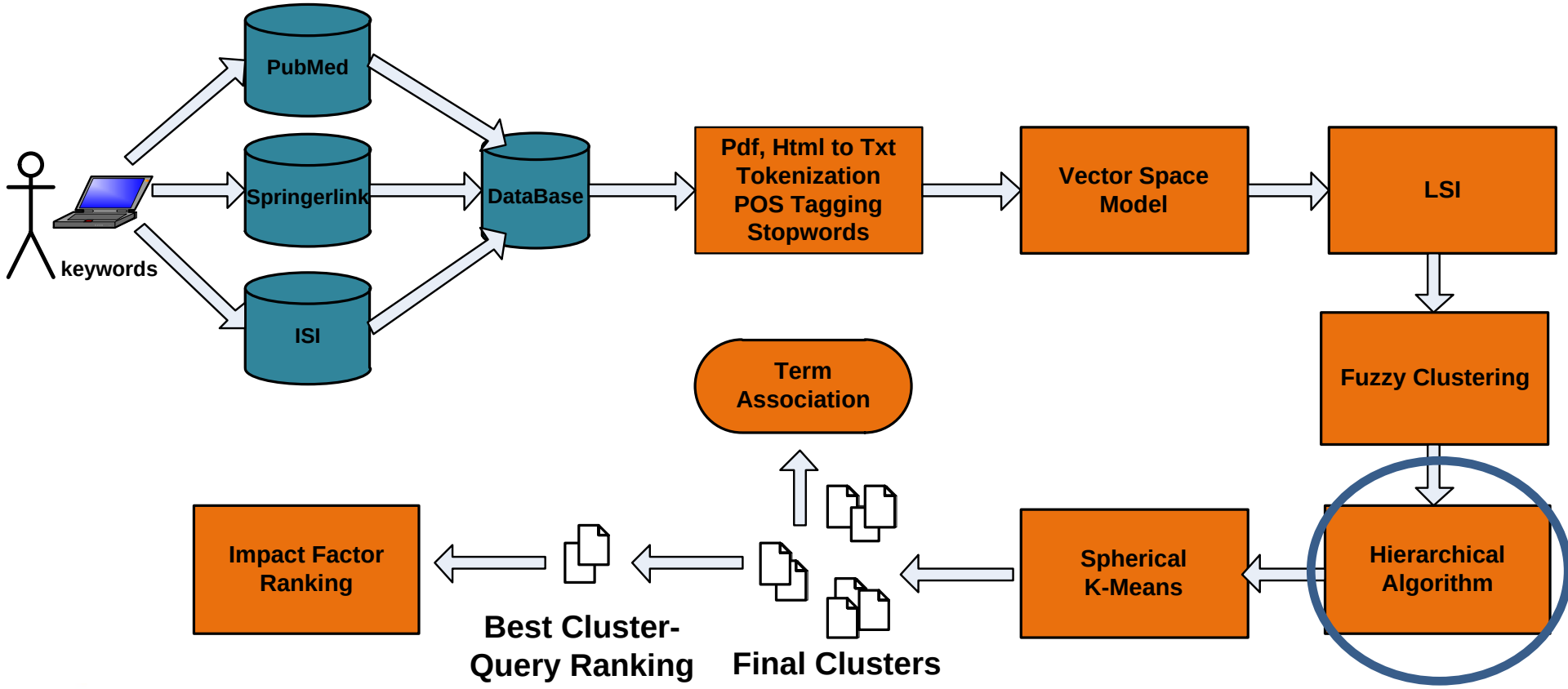
# Ασαφής Συσταδοποίηση (Fuzzy Clustering)

- Ερμηνεία των αποτελεσμάτων του LSI ως ένα είδος **Fuzzy Clustering**.
- Από τους πίνακες $V_k$ και $S_k$ που προέκυψαν από την SVD, υπολογίζουμε τον πίνακα $V_kS_k$, ο οποίος έχει n γραμμές και k στήλες.
- Ερμηνεύουμε τις k στήλες του ως ένα σύνολο από k **συστάδες** και τις n γραμμές του ως τα **κείμενα**.
- Κάθε στοιχείο (i,j) του πίνακα, όπου i είναι η γραμμή και j η στήλη, ορίζει τον **βαθμό συμμετοχής** του κειμένου i στη συστάδα j.
- Μετασχηματισμός σε **Crisp Clustering**: αντιστοιχίζοντας κάθε κείμενο στη συστάδα, στην οποία το κείμενο έχει το μεγαλύτερο βαθμό συμμετοχής, σύμφωνα με τον πίνακα.

# Βήμα 6

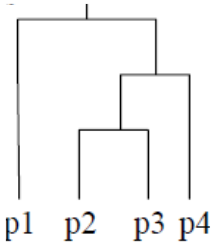

PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Βήμα 6

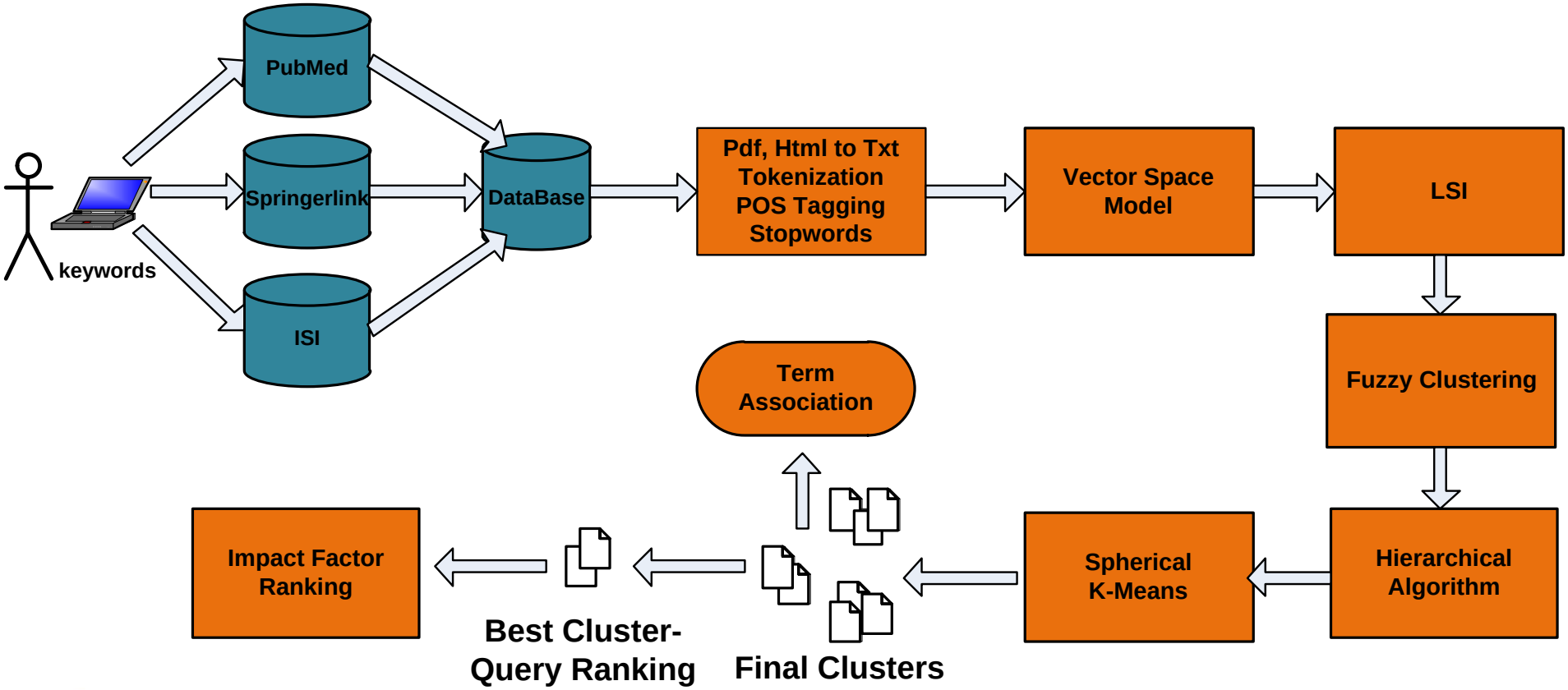

PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Ιεραρχικός Αλγόριθμος (Hierarchical Algorithm)

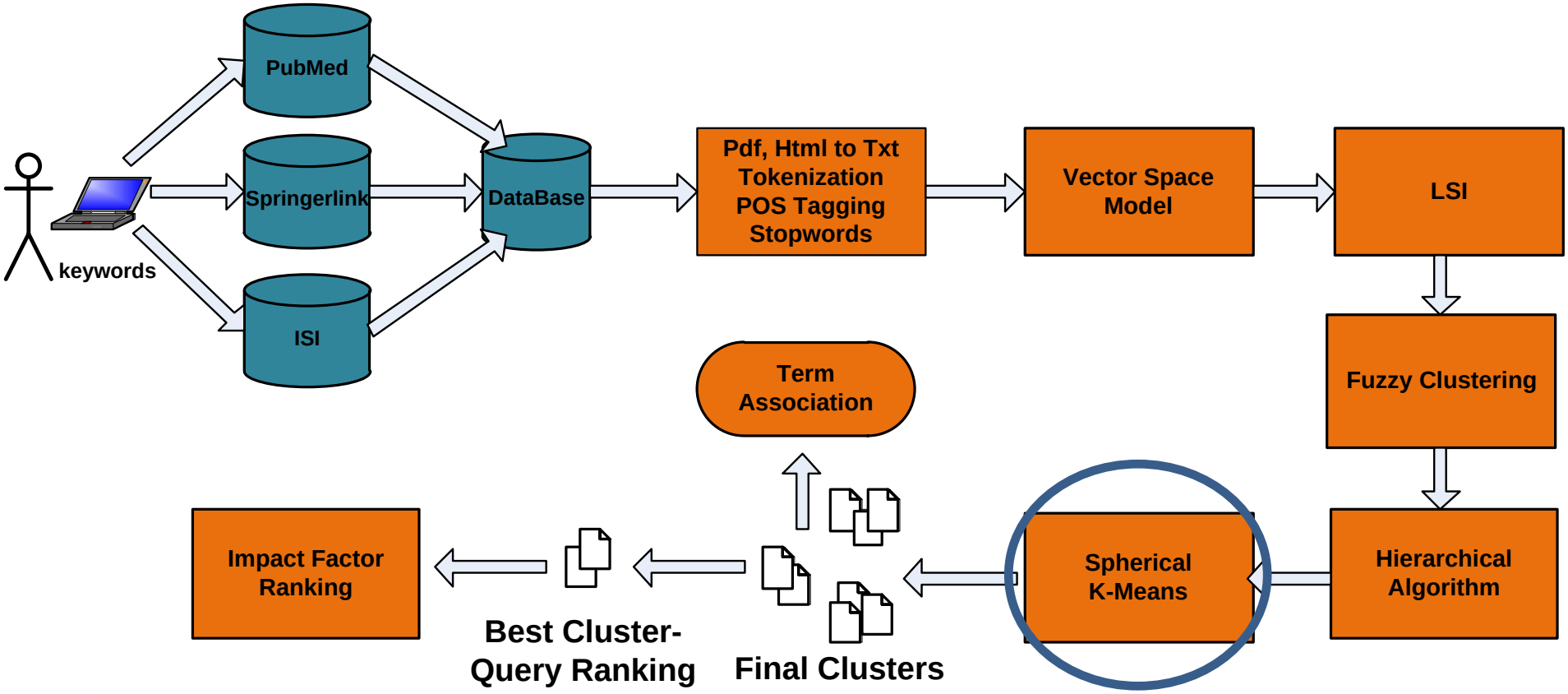
- **Μείωση** του αριθμού των συστάδων που προέκυψαν από το Fuzzy Clustering από k σε K.
- K είναι μια παράμετρος που παρέχεται από τον χρήστη. K είναι επίσης και ο αριθμός των αρχικών συστάδων που δίνονται ως είσοδος στον Spherical K-means, στο επόμενο βήμα.
- Ο Ιεραρχικός Συσσωρευτικός Αλγόριθμος, σε κάθε βήμα του, **ενώνει** τις δύο πιο όμοιες συστάδες, έως ότου ο αριθμός των συστάδων να είναι K.
- Δημιουργείται ένα δενδρόγραμμα:



# Βήμα 7


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Βήμα 7


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Spherical K-means Algorithm

- Μια από τις πιο γνωστές εκδοχές του K-means.
- Χρησιμοποιεί την **ομοιότητα συνημιτόνου** ως μετρική απόστασης.
- Καλή απόδοση σε μεγάλα σύνολα κειμένων.
- Κύριο μειονέκτημα του Spherical K-means είναι η τυχαία επιλογή των αρχικών συστάδων και των κέντρων τους.
- Για αυτόν το λόγο, ορίσαμε ως αρχικές συστάδες, τις συστάδες που προέκυψαν στο προηγούμενο βήμα του Ιεραρχικού Αλγορίθμου.
- Εφαρμογή του βελτιστοποιημένου "**Ping-Pong**" αλγόριθμου, ο οποίος αποτελείται από δύο βήματα:
  - ✓ Εφαρμογή του Spherical K-means και στην περίπτωση που αποτύχει, εφαρμογή του Kernighan-Lin.

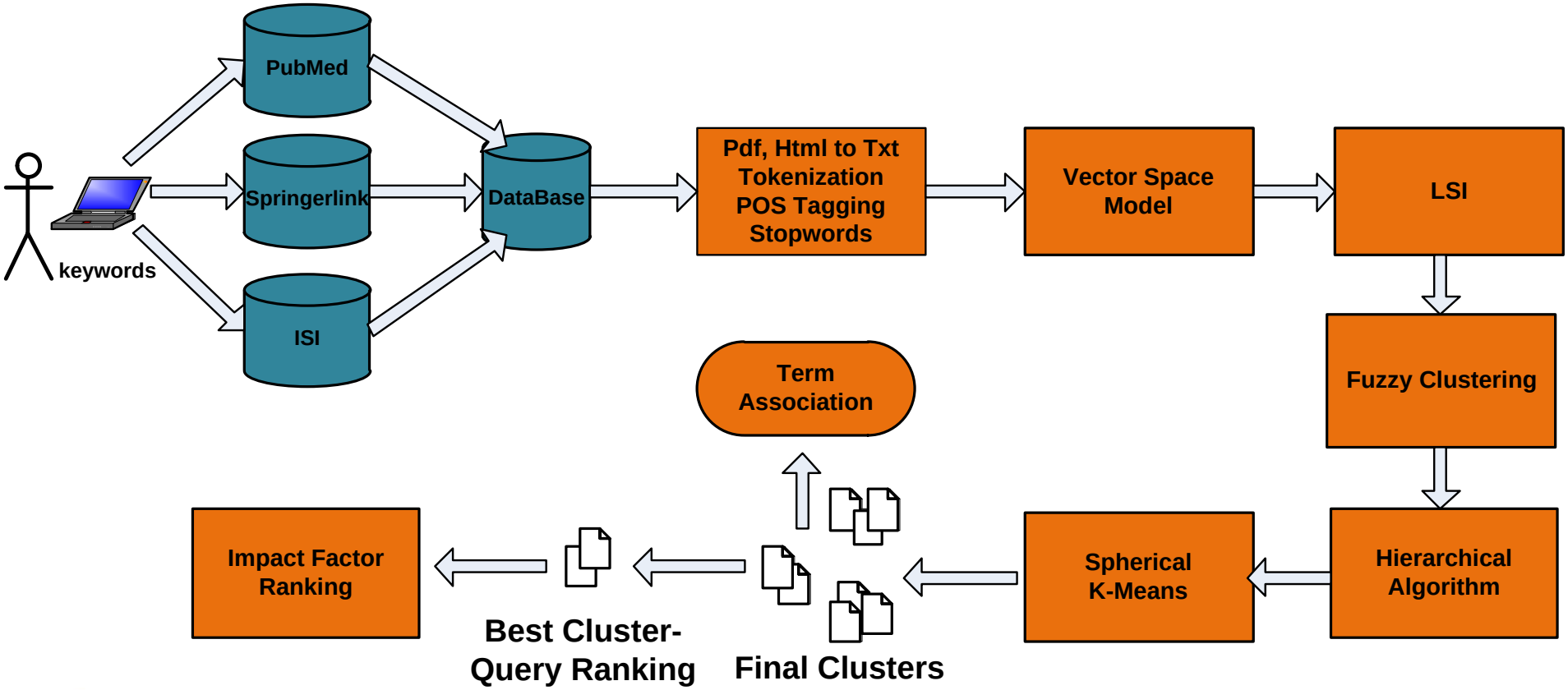
# Spherical K-means Algorithm

- Ξεκινάμε από μια αρχική διαμέριση και τα αντίστοιχα κέντρα των συστάδων.
- Για κάθε διάνυσμα κειμένου x βρίσκουμε το πιο κοντινό κέντρο με βάση την ομοιότητα συνημιτόνου και αντιστοιχίζουμε το κείμενο x στη συστάδα αυτή.
- Προκύπτει μια νέα διαμέριση.
- Υπολογίζουμε τα νέα κέντρα.
- Εάν αυξήθηκε η τιμή της αντικειμενικής συνάρτησης, επαναλαμβάνουμε. Διαφορετικά, σταματάμε.

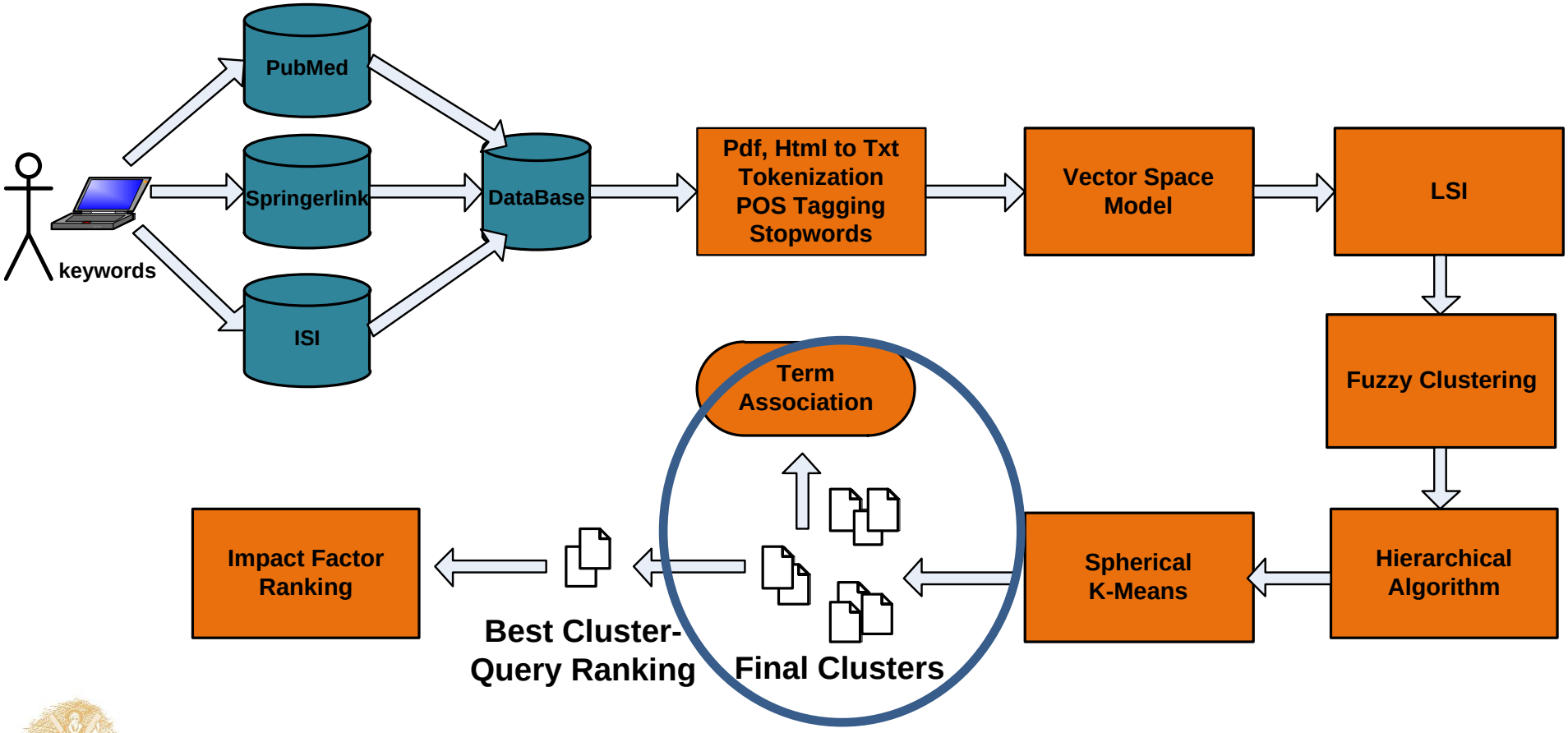
# First Variation & Ευρετική Kernighan-Lin

- Μια **first variation** επανάληψη μετακινεί ένα διάνυσμα από μια υπάρχουσα συστάδα σε μία άλλη.
- Δημιουργούμε όλες τις πιθανές μετακινήσεις και επιλέγουμε την μετακίνηση που μεγιστοποιεί την αντικειμενική συνάρτηση Q.
- Ένας τρόπος για να βελτιώσουμε την first variation είναι να επεκτείνουμε την τοπική αναζήτηση, αναζητώντας μια **αλληλουχία μετακινήσεων** αντί της μιας μόνο μετακίνησης.
- Η ιδέα αυτή υλοποιείται ακολουθώντας την ευρέως γνωστή ευρετική **Kernighan-Lin**.

# Βήμα 8


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

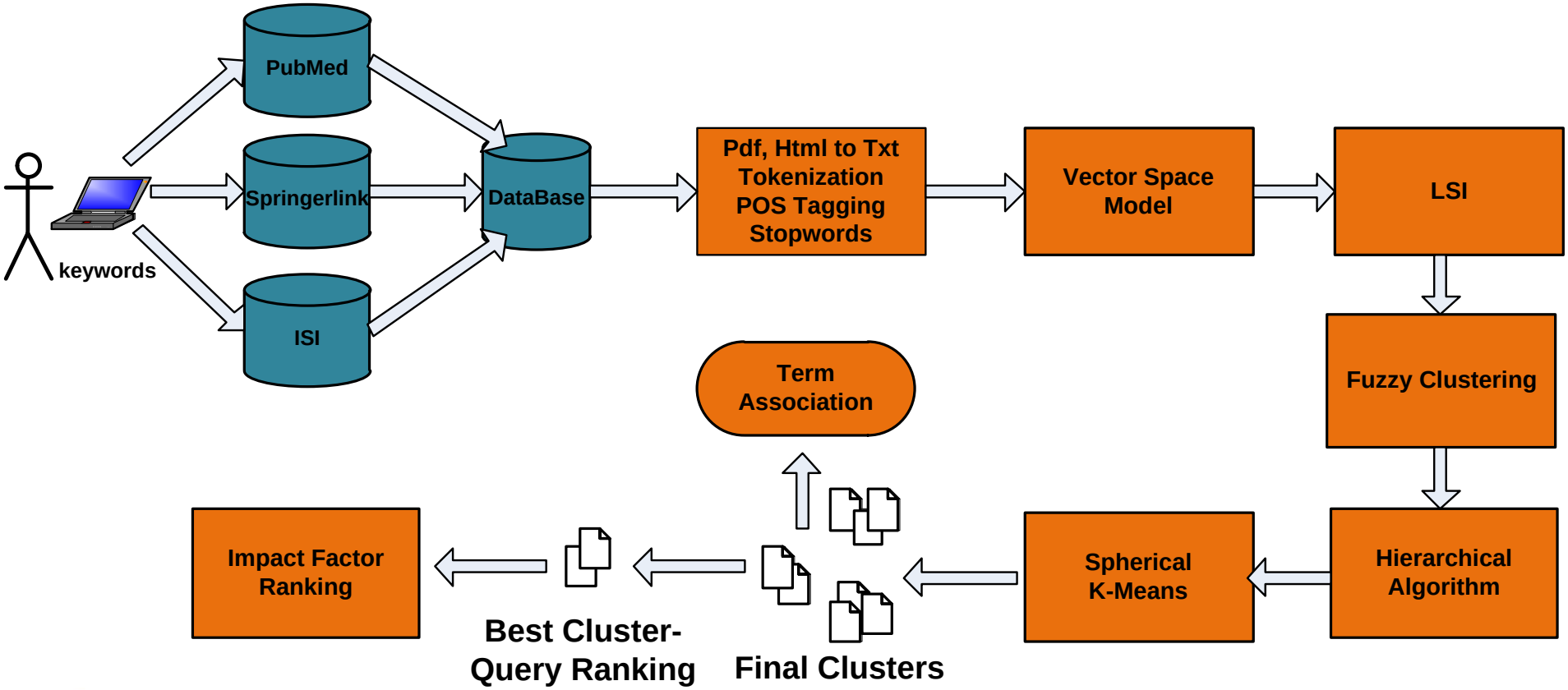
# Βήμα 8


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

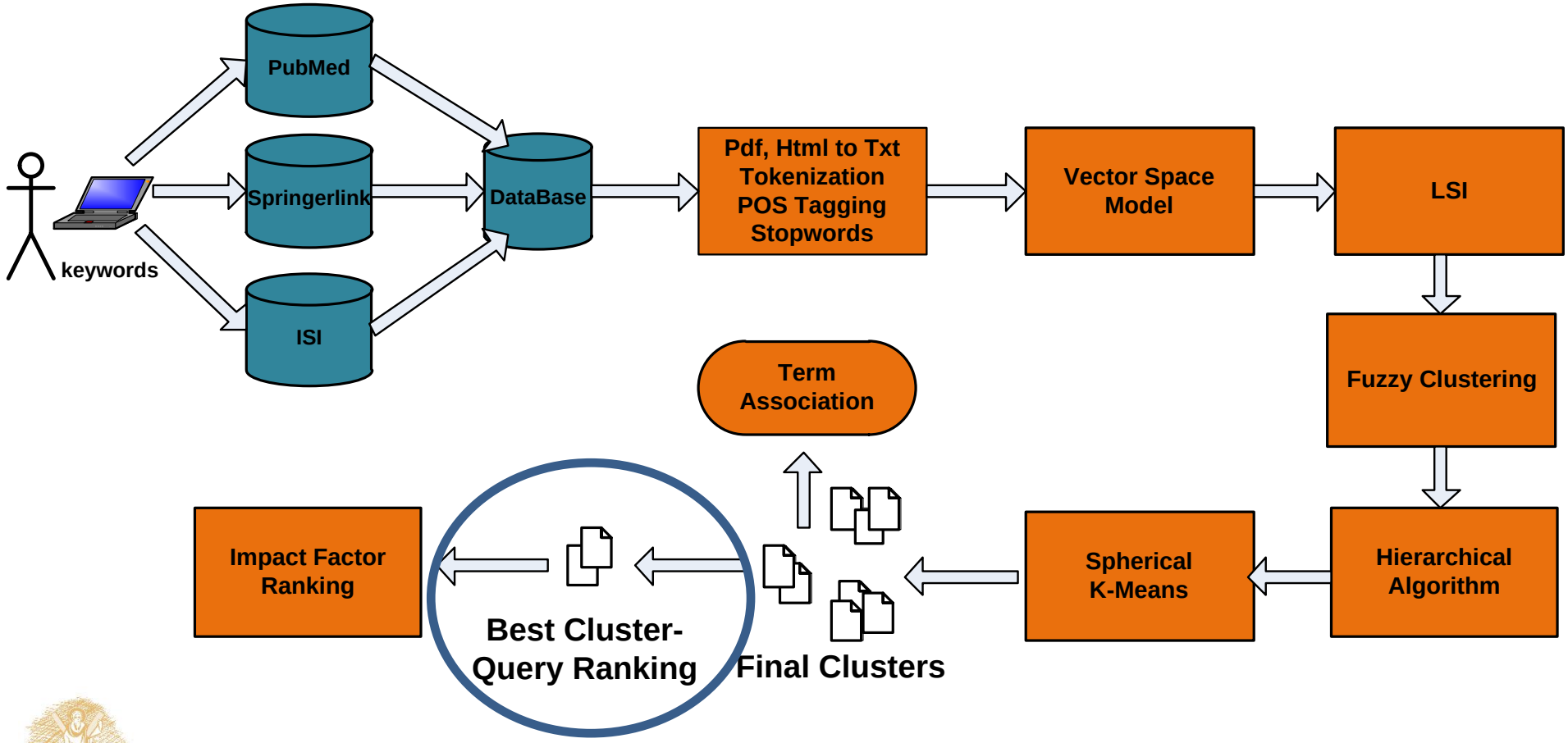
# Term Association

- Η ανακάλυψη συσχετίσεων μεταξύ όρων της βιοϊατρικής αποτελεί ένα από τα πιο προκλητικά προβλήματα του τομέα της βιοϊατρικής έρευνας, καθώς οι ερευνητές ενδιαφέρονται για την εξαγωγή συσχετίσεων μεταξύ των γονιδίων, πρωτεϊνών, ασθενειών και φαρμάκων.
- Στο βήμα αυτό έχουν ήδη προκύψει οι τελικές συστάδες και τα αντίστοιχα κέντρα τους.
- Θεωρούμε ως σχετικούς όρους, τους όρους των κέντρων (centroids) των τελικών συστάδων που έχουν βάρος μεγαλύτερο από ένα όριο.

# Βήμα 9


PubMed
Springerlink
ISI
DataBase
keywords
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

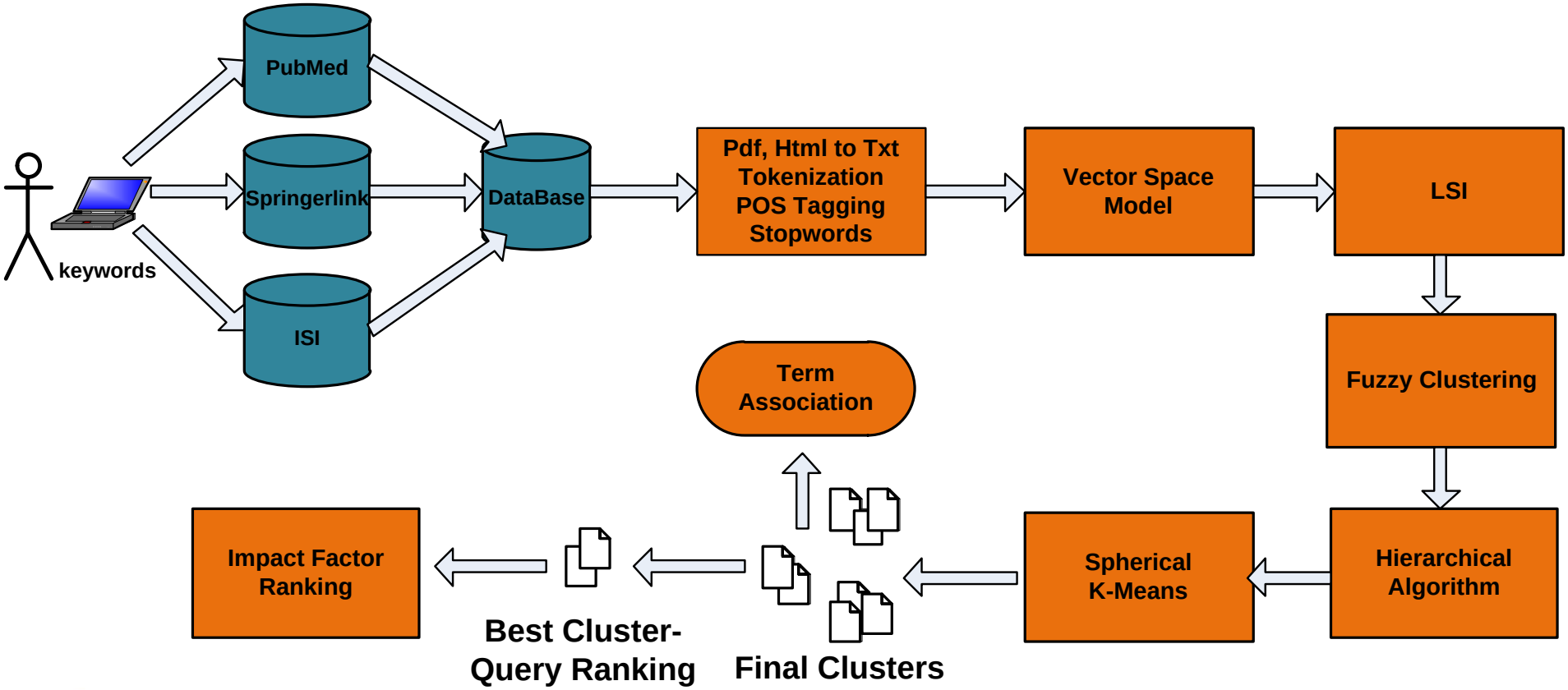
# Βήμα 9


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

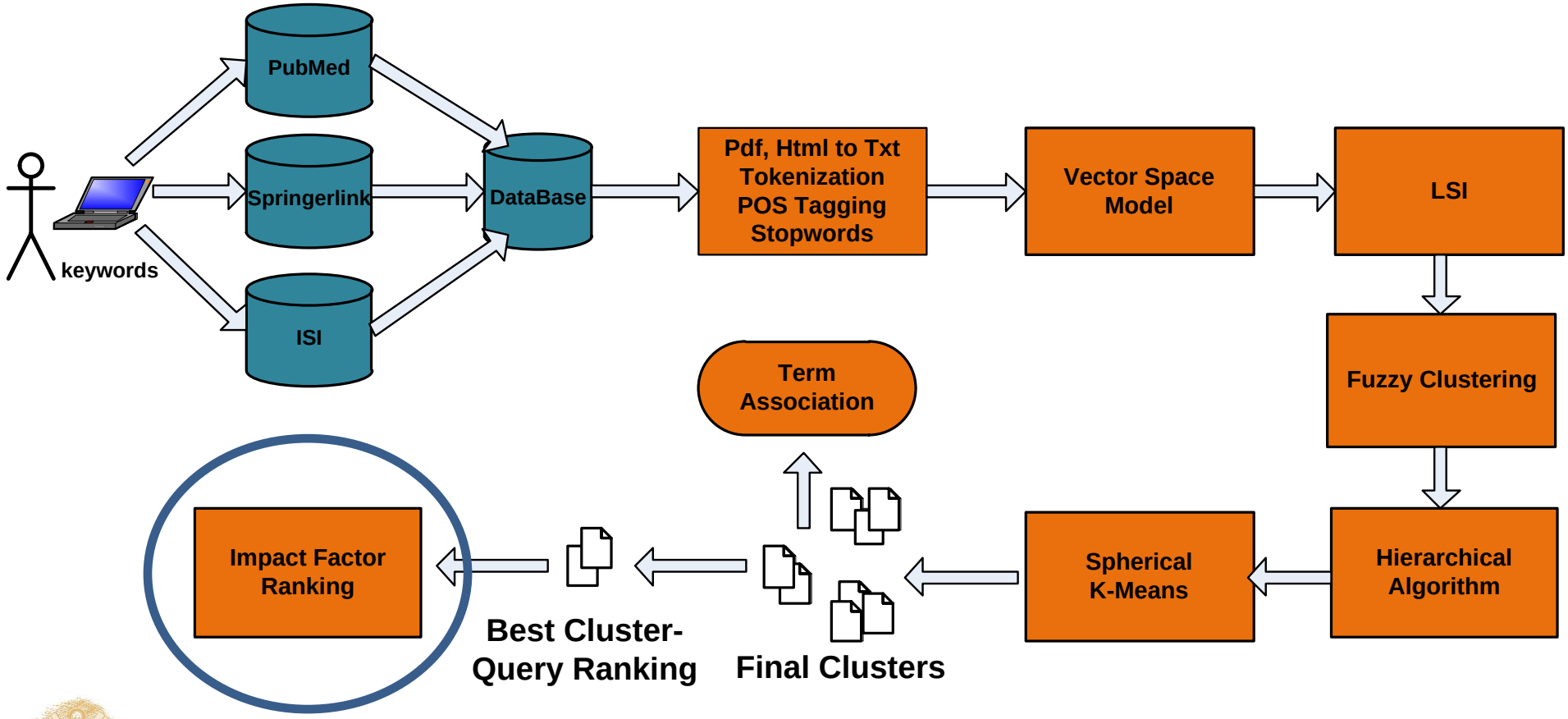
# Επιλογή «καλύτερης» συστάδας

- Αρχικά, επιλέγεται η «καλύτερη» συστάδα, δηλαδή η συστάδα που βρίσκεται πιο κοντά στο ερώτημα με βάση τη **μετρική συνημιτόνου**.
- Χρησιμοποιώντας τη μετρική της ομοιότητας συνημιτόνου, υπολογίζουμε την ομοιότητα του διανύσματος του ερωτήματος με το κέντρο (centroid) κάθε συστάδας και επιλέγουμε την συστάδα που παρουσιάζει την μεγαλύτερη ομοιότητα με το ερώτημα.
- Στη συνέχεια, ταξινομούμε τα κείμενα της καλύτερης συστάδας, με βάση την ομοιότητά τους με το ερώτημα, υπολογίζοντας την ομοιότητα συνημιτόνου του διανύσματος του ερωτήματος με τα διανύσματα των κειμένων της συστάδας.

# Βήμα 10


PubMed
Springerlink
ISI
keywords
DataBase
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Βήμα 10


PubMed
Springerlink
ISI
DataBase
keywords
Pdf, Html to Txt
Tokenization
POS Tagging
Stopwords
Vector Space
Model
LSI
Fuzzy Clustering
Hierarchical
Algorithm
Spherical
K-Means
Term
Association
Final Clusters
Best Cluster-
Query Ranking
Impact Factor
Ranking

# Impact Factor Ranking

- Πραγματοποιείται η ταξινόμηση των κειμένων της «καλύτερης» συστάδας με βάση το Impact Factor των κειμένων.

- Ο τρόπος με τον οποίο πραγματοποιείται η ταξινόμηση είναι ο εξής:
  - ✓ Έστω ένα κείμενο i που βρίσκεται στη θέση κ και ένα κείμενο j που βρίσκεται στη θέση κ-1, σύμφωνα με την προηγούμενη ταξινόμηση.
  - ✓ Υπολογίζουμε την ομοιότητα συνημιτόνου μεταξύ των κειμένων i και j και του ερωτήματος, έστω $sim_i$ και $sim_j$.
  - ✓ Όταν η απόλυτη διαφορά των $sim_i$ και $sim_j$ είναι μικρότερη από ένα όριο και το κείμενο j έχει μεγαλύτερο Impact Factor από το κείμενο i, τότε το κείμενο j τοποθετείται στην υψηλότερη (k) και το κείμενο i στη χαμηλότερη θέση k-1.

# Τρόποι Αναζήτησης

- Ο χρήστης έχει τη δυνατότητα να δώσει ένα **keyword** για αναζήτηση ή να επιλέξει ένα keyword από μια λίστα από προκαθορισμένα Topic βιολογικού περιεχομένου.
- Η εφαρμογή λειτουργεί ως **μέσο συμπιεσμένης αποθήκευσης** των προηγούμενων ερωτημάτων του χρήστη.
- Η εφαρμογή επιστρέφει αποτελέσματα για **παρόμοιες αναζητήσεις** που έχουν πραγματοποιεί στο παρελθόν και τα αποτελέσματα τους είναι αποθηκευμένα στη βάση δεδομένων του συστήματος.
- Η εφαρμογή εξάγει χρήσιμες **συσχετίσεις** μεταξύ των ερωτημάτων και των βιολογικών όρων της βιβλιογραφίας.

# Τέλος Ενότητας

# Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Πατρών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

# Σημειώματα

# Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

# Σημείωμα Αναφοράς



# Σημείωμα Αδειοδότησης

# Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.