PRINCIPAL COMPONENT ANALYSIS (PCA) KAI HIERACHICAL CLUSTER ANALYSIS (HCA)

Γιώργος Μικρός
ΕΚΠΑ ~ University of Massachusetts, Boston

Ανάλυση Κύριων Συνιστωσών

- Η Ανάλυση Κύριων Συνιστωσών (ΑΚΣ) (Principal Component Analysis – PCA) είναι μία μέθοδος η οποία μετασχηματίζει γραμμικά ένα μεγάλο αριθμό μεταβλητών σε ένα νέο σύνολο μεταβλητών που μεταξύ τους δεν υπάρχουν συσχετίσεις.
- Οι νέες μεταβλητές ονομάζονται κύριες συνιστώσες (principal components).
- Από αυτές μας αρκεί η εξέταση των πρώτων δύο ή τριών για να εξηγήσουμε την ποικιλία των δεδομένων μας και να εντοπίσουμε σημαντικές κανονικότητες ή ομαδοποιήσεις.

Προϋποθέσεις

- Για να υπολογιστεί η ΑΚΣ χρειάζεται να συλλεχθούν μετρήσεις από πολλές μεταβλητές, οι οποίες σχετίζονται εννοιολογικά.
- Στην υφομετρία για παράδειγμα μπορούμε να χρησιμοποιήσουμε την συχνότητα εμφάνισης των 100 πιο συχνών λέξεων σε ένα ΗΣΚ δύο συγγραφέων.
- Η κλασική μορφή αναπαράστασης αυτών των μετρήσεων γίνεται σε πίνακες όπου οι στήλες αντιπροσωπεύουν τις μεταβλητές (λέξεις) και οι σειρές αντιπροσωπεύουν τα κείμενα στα οποία μετρήθηκαν.

Κειμενική αναπαράσταση σε διανυσματική μορφή

Αρχείο κειμένου	Κειμενικό Γένος	Συγγραφέας	και	ο	του	την	είναι
1.txt	Email	A	18	17	13	7	3
2.txt	Email	B	27	21	13	9	2
3.txt	Ανάρτηση σε ιστολόγιο	A	19	16	15	6	4
4.txt	Ανάρτηση σε ιστολόγιο	Άγνωστος	28	20	17	9	1

Ερμηνευτική δύναμη της ΑΚΣ

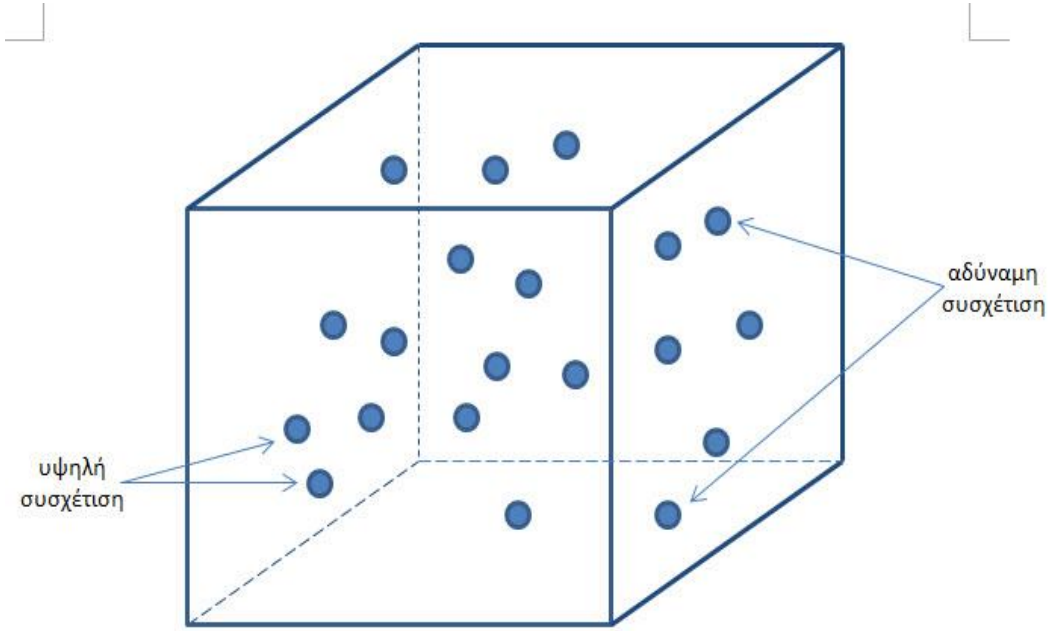
- Η ΑΚΣ θα μας βοηθήσει να αποφασίσουμε αν οι σχέσεις που εμφανίζουν οι συχνότητες των πιο συχνών λέξεων στα κείμενα, σχετίζεται με τον συγγραφέα των κειμένων, αλλά και το κειμενικό γένος.
- Ο στόχος της ΑΚΣ είναι η ερμηνεία της διακύμανσης (variance) που εμφανίζουν τα πολυδιάστατα δεδομένα μας. Για να το πετύχει αυτό η ΑΚΣ εξάγει συνιστώσες, δηλαδή μέρος της διακύμανσης από την συνολική διακύμανση των δεδομένων.
- Η κάθε κύρια συνιστώσα είναι στην ουσία ένας γραμμικός συνδυασμός των αρχικών μεταβλητών με βάρη (weighted linear combination), τα οποία αντιπροσωπεύουν τον βαθμό συσχέτισης της μεταβλητής με την συνιστώσα.
- Για να μην υπάρξει πρόβλημα με την χρήση διαφορετικών κλιμάκων στις μεταβλητές ή με την διαφορά συχνότητας που παρουσιάζουν οι υψίσυχνες από της λιγότερο συχνές λέξεις, είναι σημαντικό να κανονικοποιήσουμε τις συχνότητες των λέξεων, να μετατρέψουμε δηλαδή την κλίμακα μέτρησής τους σε μία άλλη, όπου ο μέσος όρος είναι 0 και η τυπική απόκλιση είναι 1, να τις μετατρέψουμε δηλαδή σε z τιμές.

Πίνακας συχνοτήτων λέξεων όπου οι συχνότητες των λέξεων έχουν μετατραπεί σε τυπικές τιμές.

Αρχείο κειμένου	και	ο	του	την	είναι
1.txt	-0.956	-0.630	-0.783	-0.500	0.387
2.txt	0.765	1.050	-0.783	0.833	-0.387
3.txt	-0.765	-1.050	0.261	-1.167	1.162
4.txt	0.956	0.630	1.306	0.833	-1.162

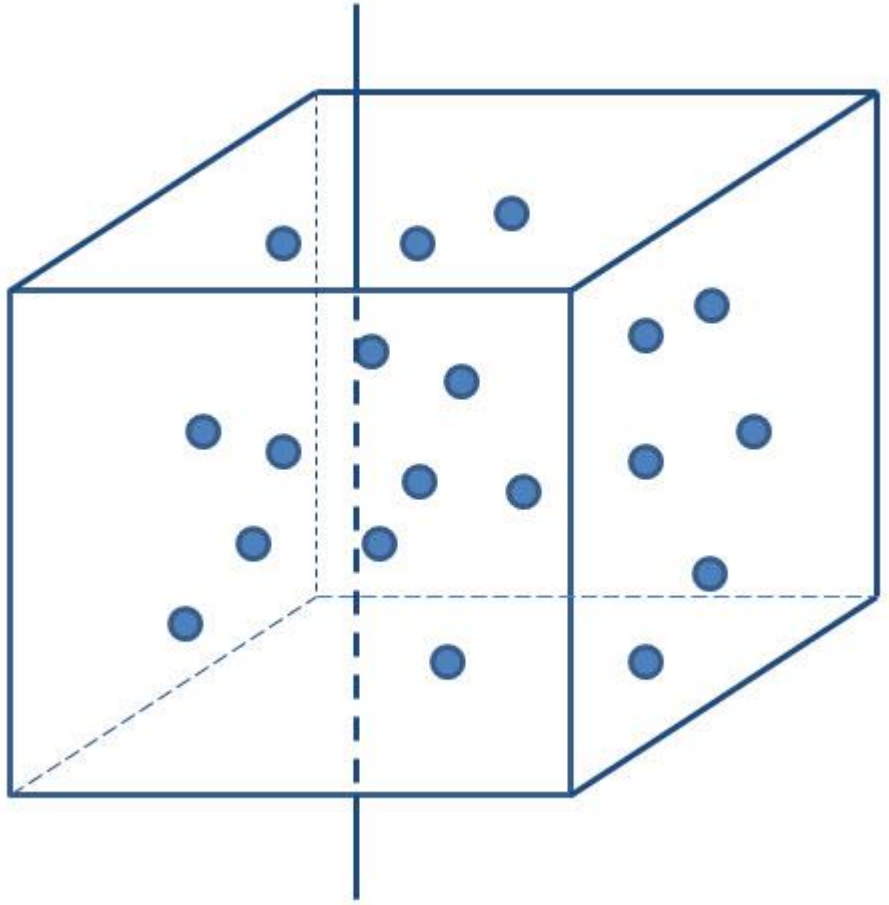
Εξαγωγή Κυρίων Συνιστωσών

- Φανταστείτε ότι οι μεταβλητές μας (συχνότητες των πιο συχνών λέξεων) βρίσκονται σε έναν τρισδιάστατο χώρο (κύβος) και η θέση τους αντιπροσωπεύει τον βαθμό συσχέτισης μεταξύ τους.
- Για να εξαγάγουμε τις κύριες συνιστώσες αυτού του χώρου θα πρέπει να ακολουθήσουμε τους παρακάτω κανόνες:
  - Η εξαγωγή, δηλαδή η προσαρμογή συνιστωσών στον πολυδιάσταστο χώρο γίνεται με την προσαρμογή μίας συνιστώσας κάθε φορά.
  - Κάθε συνιστώσα θα πρέπει να περάσει μέσα από το κέντρο βάρους αυτού του χώρου.
  - Μόλις η πρώτη συνιστώσα περάσει στον χώρο, όλες οι επόμενες θα πρέπει να διασταυρώσουν την πρώτη ή άλλες που ήδη βρίσκονται στον χώρο με γωνία 90^{o}.
  - Όταν μία συνιστώσα εισέρχεται στον πολυδιάστατο χώρο θα πρέπει να εξηγήσει όλη την διακύμανση που είναι εφικτό να συλλάβει, δηλαδή να προσανατολιστεί έτσι ώστε να ικανοποιηθεί το κριτήριο της βέλτιστης προσαρμογής (best fit criterion).
  - Αν υπάρχουν ήδη συνιστώσες στον χώρο, τότε μια νέα συνιστώσα θα πρέπει να εξηγήσει διακύμανση που δεν έχει ήδη εξηγηθεί από τις προηγούμενες συνιστώσες.



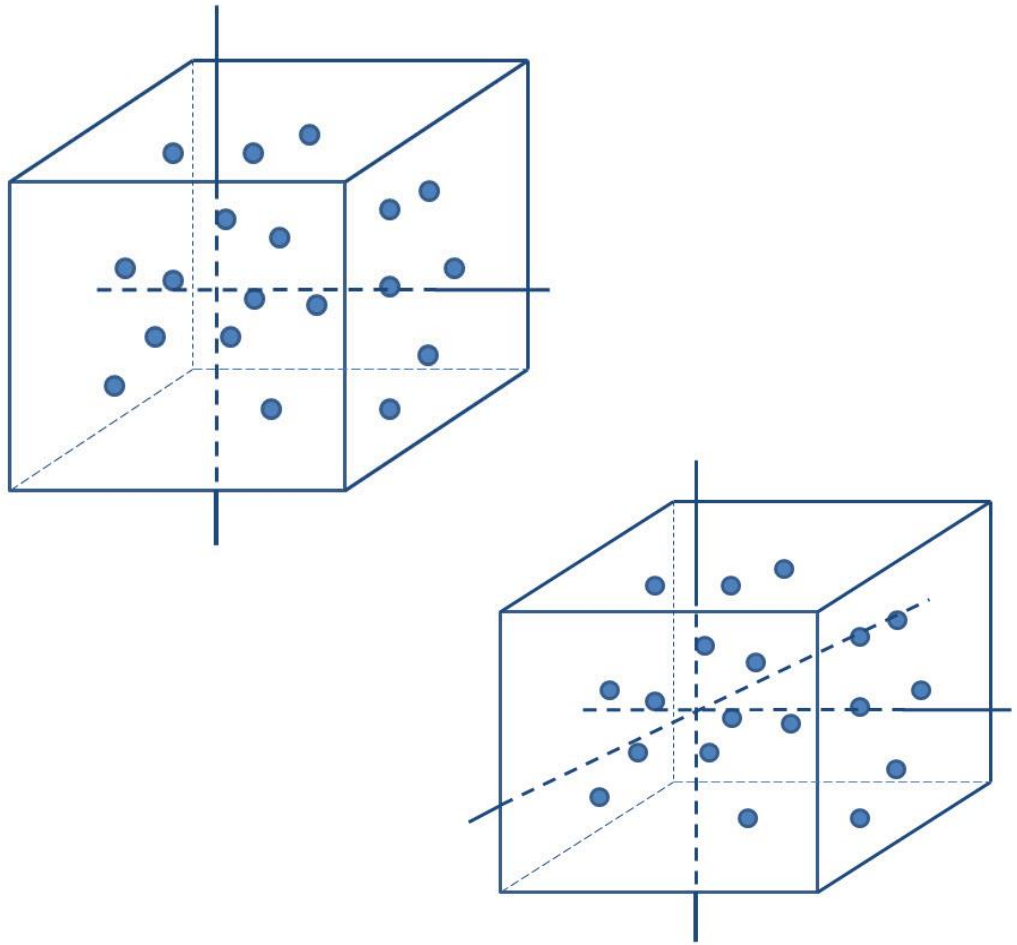
Πρώτη Κύρια Συνιστώσα

- Η εξαγωγή της πρώτης κύριας συνιστώσας στον χώρο αυτό σχετίζεται με την τοποθέτηση ενός άξονα στον χώρο σε οποιαδήποτε γωνία έτσι ώστε να περνά κοντά από όλα τα σημεία – μεταβλητές. Η θέση αυτή είναι μοναδική και προκύπτει από την επίλυση μιας γραμμικής εξίσωσης όπου το άθροισμα των τετραγώνων των αποστάσεων των μεταβλητών από τον άξονα είναι το ελάχιστο (μέθοδος ελάχιστων τετραγώνων – least squares method).
- Η τοποθέτηση της πρώτης κυρίας συνιστώσας έγινε με τρόπο που επιτρέπει την ερμηνεία του μεγαλύτερου δυνατού ποσοστού διακύμανσης των δεδομένων. Το ποσοστό αυτό δε, θα είναι πάντα μεγαλύτερο από οποιαδήποτε επόμενη τοποθέτηση συνιστώσας.



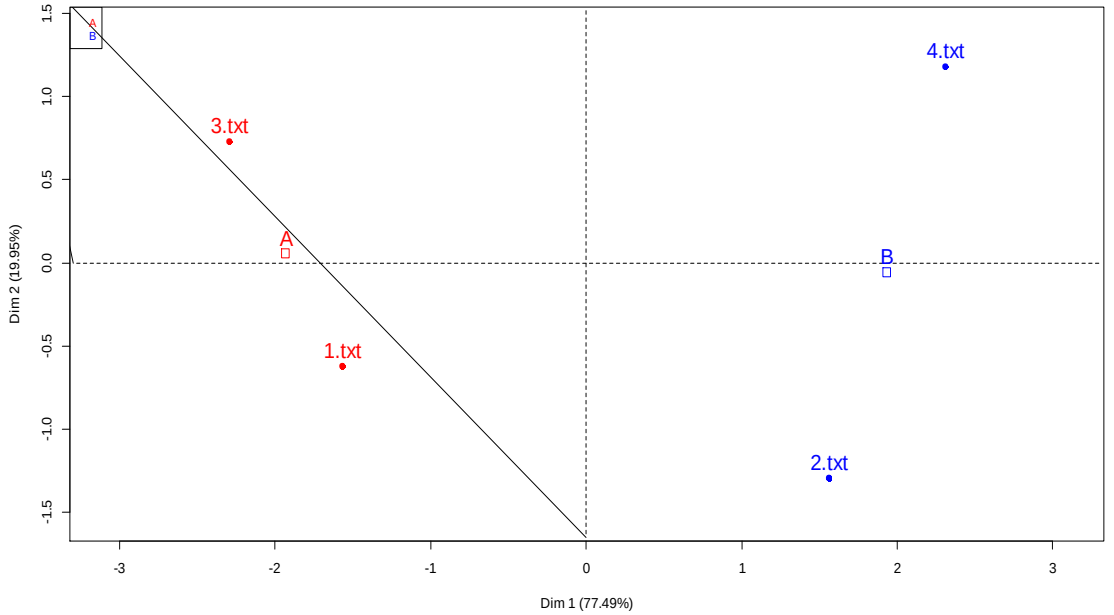
Επόμενες Κύριες Συνιστώσες

- Η δεύτερη Κύρια Συνιστώσα θα πρέπει να τέμνει την πρώτη σε γωνία 90^o, δηλαδή θα πρέπει να είναι ορθογώνια σε αυτήν γιατί οι δύο συνιστώσες θα πρέπει να είναι ανεξάρτητες μεταξύ τους και να μην εμφανίζουν κάποια συσχέτιση (Κανόνας 3). Και στην περίπτωση της δεύτερης συνιστώσας, η βέλτιστη θέση θα καθοριστεί από την μέθοδο των ελαχίστων τετραγώνων. Βασική προϋπόθεση είναι η συνιστώσα να περνάει από το κέντρο βάρους του χώρου (Κανόνας 2) και θα πρέπει να είναι όσο το δυνατόν πιο κοντά σε όλες τις μεταβλητές παίρνοντας υπ' όψιν ωστόσο, ότι η πρώτη κύρια συνιστώσα έχει λάβει την καλύτερη θέση και επομένως εξηγεί το μεγαλύτερο ποσοστό διακύμανσης των δεδομένων.
- Η τρίτη κύρια συνιστώσα πρέπει επίσης να περάσει από το κέντρο του επιπέδου και να είναι ορθογώνια στις δύο προηγούμενες κύριες συνιστώσες. Θα πρέπει επίσης να καταλάβει την καλύτερη δυνατή θέση (πιο κοντινή στα σημεία – μεταβλητές). Η απόσταση αυτή θα είναι η τρίτη καλύτερη, γιατί η πιο καλή καταλήφθηκε από την πρώτη κύρια συνιστώσα και η δεύτερη πιο καλή από την δεύτερη κυρία συνιστώσα. Γενικά, μπορούμε να συνεχίσουμε την εξαγωγή συνιστωσών, αλλά η οπτικοποίησή τους δεν θα είναι εφικτή.



Γραφική αναπαράσταση της ΑΚΣ

Κείμενα	Σκορ στην Κύρια Συνιστώσα 1	Σκορ στην Κύρια Συνιστώσα 2
1.txt	1,359	-0,5375
2.txt	-1,3511	-1,12
3.txt	1,99	0,6341
4.txt	-1,9979	1,0234



Ερμηνεία

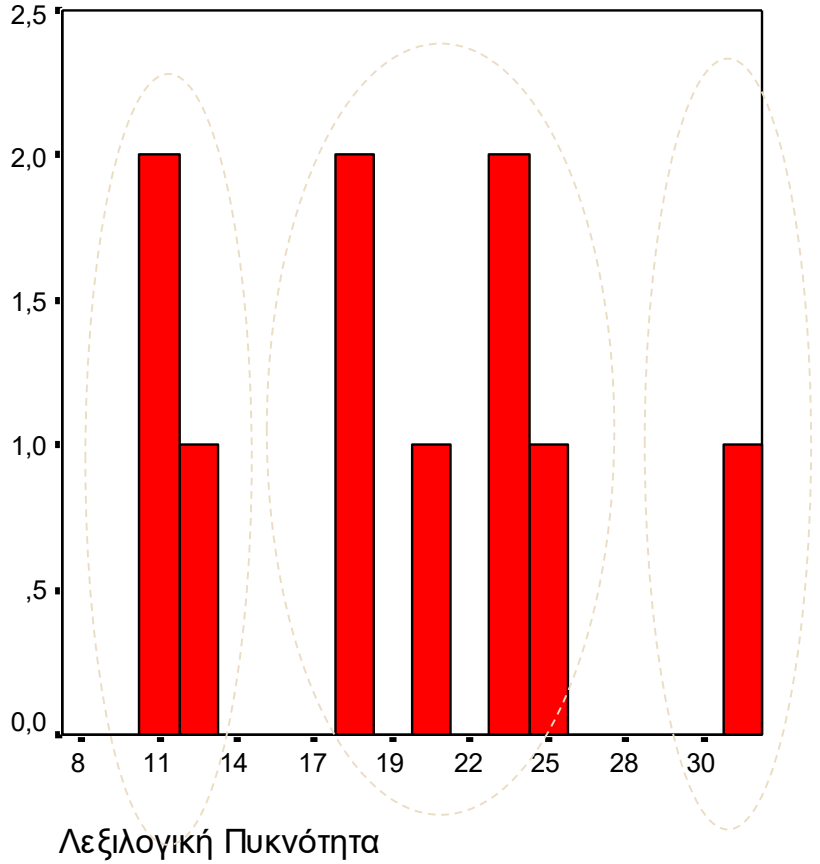
- Το παραπάνω διάγραμμα είναι το σημαντικότερο ερμηνευτικό εργαλείο στην αποκάλυψη της πατρότητας αγνώστων κειμένων. Ο οριζόντιος άξονας αντιστοιχεί στην πρώτη κύρια συνιστώσα, η οποία ερμηνεύει το 77,49% της διακύμανσης των δεδομένων μας. Όπως είναι εμφανές, τα κείμενα 1.txt και 3.txt, συγγραφέας των οποίων είναι ο Α, βρίσκονται στην αριστερή πλευρά του διαγράμματος (αρνητικές τιμές στα σκορ των κειμένων στην πρώτη κύρια συνιστώσα). Αντίθετα, το κείμενο 2.txt που ανήκει στον Β βρίσκεται στην δεξιά πλευρά (θετικές τιμές στα σκορ των κειμένων στην πρώτη κύρια συνιστώσα). Η πατρότητα του 4.txt μπορεί να προσδιοριστεί οπτικά. Το συγκεκριμένο κείμενο βρίσκεται στην δεξιά πλευρά (έχει θετικό σκορ στην πρώτη κύρια συνιστώσα), γεγονός που μαρτυρά ότι ανήκει στο υφομετρικό προφίλ του συγγραφέα Β. Επομένως η συγγραφική πατρότητα σχετίζεται με την θέση των κειμένων στην πρώτη κύρια συνιστώσα.
- Ο κάθετος άξονας αντιστοιχεί στην δεύτερη κύρια συνιστώσα η οποία ερμηνεύει μικρότερη διακύμανση στα δεδομένα μας (19,95%). Η θέση των κειμένων μας αποκαλύπτει ότι η συγκεκριμένη κύρια συνιστώσα σχετίζεται με το κειμενικό γένος. Πιο συγκεκριμένα, τα κείμενα 1.txt και 2.txt βρίσκονται στην κάτω πλευρά του διαγράμματος ενώ τα κείμενα 3.txt και 4.txt βρίσκονται στην πάνω πλευρά του διαγράμματος. Από αυτή την κατανομή συμπεραίνουμε ότι οι θετικές τιμές στα σκορ των κειμένων στην δεύτερη συνιστώσα αντιστοιχούν στις αναρτήσεις σε ιστολόγια (κείμενα 3.txt και 4.txt), ενώ αρνητικές τιμές στα σκορ αντιστοιχούν σε emails.

Ανάλυση Συστάδων (cluster analysis)

- Η ΑΣ κατηγοριοποιεί ένα πλήθος παρατηρήσεων σε δύο ή περισσότερες αμοιβαία αποκλειόμενες ομάδες στηριζόμενη σε συνδυασμούς αριθμητικών μεταβλητών. Ο σκοπός της ΑΣ είναι να εντοπίσει ένα σύστημα που οργανώνει τις παρατηρήσεις σε ομάδες.
- Για παράδειγμα θα μπορούσαμε να διερευνήσουμε το κατά πόσο κάποιοι υφομετρικοί δείκτες (type/token ratio, μέσο μήκος λέξης, μέσο μήκος πρότασης κ.ά.) θα μπορούσαν να διακρίνουν μια σειρά από κείμενα και να τα κατατάξουν θεματικά.
- Μια σημαντική ιδιότητα της ΑΣ είναι ότι κατηγοριοποιεί τις παρατηρήσεις σε άγνωστες ομάδες.

Μια απλή ΑΣ

- Σε περιπτώσεις με μια ή δύο μεταβλητές μια απλή επισκόπηση των δεδομένων χρησιμοποιώντας ιστόγραμμα συχνότητας ή διάγραμμα διασποράς είναι αρκετή για να διαμορφώσουμε μια άποψη για τις δυνατές ομαδοποιήσεις.
- Στην περίπτωση αυτή η διάκριση σε ομάδες των κειμένων βάση της μέτρησης της λεξιλογικής πυκνότητας είναι σχεδόν προφανής.



Ο πίνακας εγγύτητας (proximities matrix)

- Η ΑΣ έχει ως αφετηρία με έναν πίνακα δεδομένων όπου τα δείγματα (συνήθως άνθρωποι στις κοινωνικές επιστήμες) είναι σειρές και οι παρατηρήσεις κωδικοποιούνται ως στήλες. Από την αρχή ο πίνακας που δημιουργείται περιλαμβάνει τιμές που είναι μετρήσεις εγγύτητας ή διαφοροποιήσεως μεταξύ δύο παρατηρήσεων.

Κειμενικό θέμα	***Λεξιλογική πυκνότητα***
Οικονομικά	11
Πολιτικά	11
Ανθρωπιστικά	13
Νομικά	18

	Οικο-νομικά	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	0	2	7
Πολιτικά	0	0	2	7
Ανθρωπιστικά	2	2	0	5
Νομικά	7	7	5	0

Υπολογίζοντας τις αποστάσεις

- Τα δεδομένα του πίνακα θα περιγραφούν χρησιμοποιώντας το γράμμα «Α». Η απόσταση γράφεται ως δείκτης στο Α. Έτσι η A_{34} περιγράφει την τομή των Ανθρωπιστικών και των Νομικών κειμένων.
- Η απόσταση υπολογίζεται με την απόλυτη τιμή της διαφοράς των δύο κειμένων. Για παράδειγμα η A_{34}, μεταξύ ανθρωπιστικών και νομικών κειμένων θα είναι |13-18| ή 5. Αν συμπληρώσουμε τον πίνακα εγγύτητας με τον τρόπο αυτό θα έχουμε τον διπλανό πίνακα.
- Ένας δεύτερος τρόπος υπολογισμού του πίνακα εγγύτητας είναι η χρήση των τετραγωνισμένων διαφορών. Π.χ. η απόσταση A_{34} θα γινόταν $(13-18)^2$ ή 25. Η συγκεκριμένη μέτρηση έχει το πλεονέκτημα ότι είναι συναφής με πολλές άλλες στατιστικές μετρήσεις όπως είναι η διακύμανση.

	Οικο-νομικά	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	0	2	7
Πολιτικά	0	0	2	7
Ανθρωπιστικά	2	2	0	5
Νομικά	7	7	5	0

	Οικο-νομικά	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	0	4	49
Πολιτικά	0	0	4	49
Ανθρωπιστικά	4	4	0	25
Νομικά	49	49	25	0

Πολυπαραγοντικές αποστάσεις

- Όταν για κάθε δείγμα έχουμε παραπάνω από μια μετρήσεις τότε θα πρέπει να βρεθεί ένας τρόπος για να συνδυαστούν σε έναν πίνακα οι επιμέρους πίνακες εγγύτητας που δημιουργούνται για κάθε μέτρηση. Συνήθως οι επιμέρους πίνακες αθροίζονται σε έναν όπως στο διπλανό παράδειγμα:

Type/token ratio	*Οικο-νομικά*	*Πολιτι-κά*	*Ανθρω-πιστικά*	*Νομι-κά*
Οικονομικά	0	64	81	100
Πολιτικά	64	0	1	4
Ανθρωπιστικά	81	1	0	1
Νομικά	100	4	1	0

+

Λεξική πυκνότητα	*Οικο-νομικά*	*Πολιτι-κά*	*Ανθρω-πιστικά*	*Νομι-κά*
Οικονομικά	0	0	4	49
Πολιτικά	0	0	4	49
Ανθρωπιστικά	4	4	0	25
Νομικά	49	49	25	0

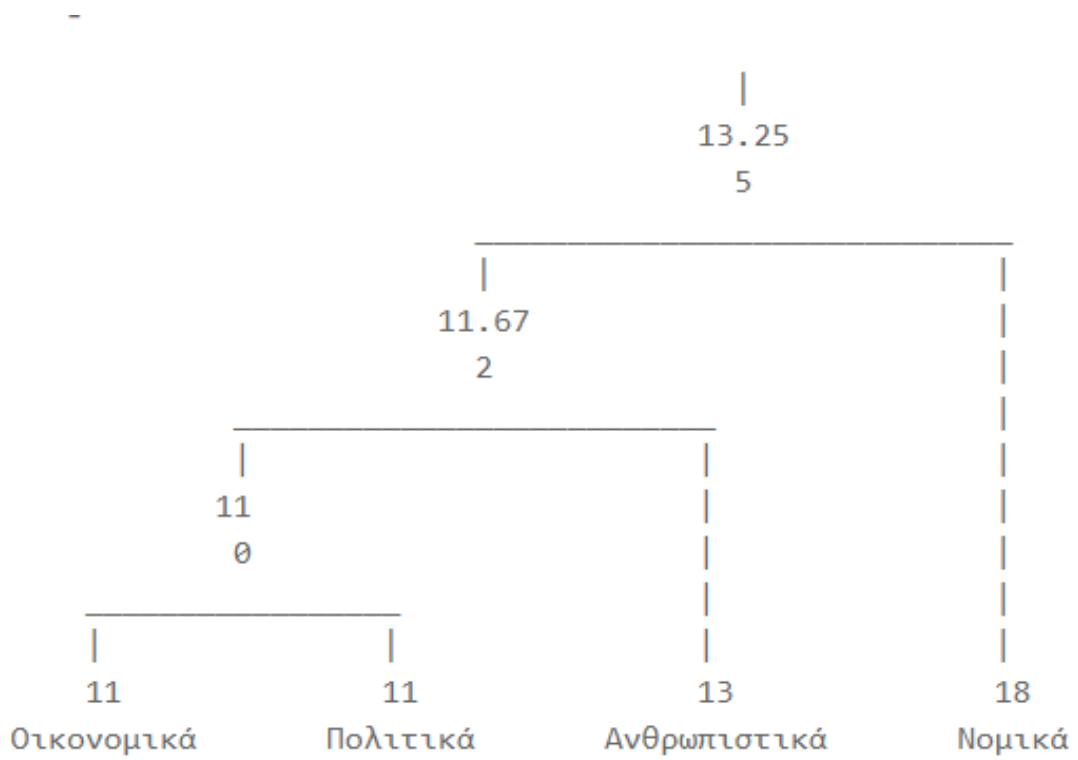
=

Σύνολο	*Οικο-νομικά*	*Πολιτι-κά*	*Ανθρω-πιστικά*	*Νομι-κά*
Οικονομικά	0	64	85	149
Πολιτικά	64	0	5	53
Ανθρωπιστικά	85	5	0	26
Νομικά	149	53	26	0

Η χρήση των αποστάσεων για την ομαδοποίηση των δειγμάτων

- Το επόμενο στάδιο μετά την μέτρηση των αποστάσεων είναι η διάκριση των δειγμάτων σε ομάδες βάσει των αποστάσεών τους.
- Αν ο αριθμός των ομάδων είναι γνωστός από πριν χρησιμοποιείται μια «επίπεδη» μέθοδος. Βάσει αυτής τα δείγματα αποδίδονται σε κάποια ομάδα στηριζόμενα σε κάποιο αρχικό κριτήριο. Υπολογίζεται ο μέσος όρος για κάθε ομάδα. Εν συνεχεία ανακατάσσονται τα δείγματα σε ομάδες βάσει τις ομοιότητας του δείγματος στο μέσο όρο της ομάδας. Αυτή η διαδικασία επαναλαμβάνεται αναδρομικά μέχρι όλα τα δείγματα να συμμετάσχουν σε κάποια ομάδα. *Αυτή η μέθοδος ονομάζεται και «k-means cluster analysis».*
- *Οι μέθοδοι ιεραρχικής συσταδοποίησης (hierarchical clustering methods) δεν απαιτούν προηγούμενη γνώση του αριθμού των ομάδων. Οι βασικότερες μέθοδοι είναι η διαιρετική (divisive) και η συσσωρευτική (agglomerative).*
  - *Οι διαιρετικές τεχνικές ξεκινούν προϋποθέτοντας μια ομάδα την οποία την διαιρούν σε υποομάδες συνεχόμενα μέχρι το κάθε δείγμα να αποτελεί το τελικό κλαδί μιας υποομάδες. Οι συσσωρευτικές τεχνικές ξεκινούν από κάθε δείγμα το οποίο περιγράφει μια υποομάδα και με συνεχόμενες συγχωνεύσεις φτάνουμε σε μια ομάδα.*
  - *Και στις δύο περιπτώσεις οι σχετικές τεχνικές περιγράφονται με δενδρόγραμμα ή δίτιμο δένδρο (binary tree). Τα δείγματα εμφανίζονται ως τελικοί κόμβοι στο δενδρόγραμμα, ενώ το μήκος των κλάδων δείχνει την απόσταση μεταξύ των υποομάδων που ενώνονται.*

Δενδρόγραμμα με αποστάσεις



Μέθοδοι συσταδοποίησης

- Απλή διασύνδεση (simple linkage): υπολογίζει την απόσταση μεταξύ των δύο υποομάδων ως την ελάχιστη απόσταση μεταξύ δύο μελών των αντίθετων ομάδων.
- Πλήρη διασύνδεση (complete linkage): υπολογίζει την απόσταση ανάμεσα στις δύο υποομάδες ως την μέγιστη απόσταση μεταξύ οποιωνδήποτε μελών στις υποομάδες.
- Μέση διασύνδεση (average linkage): υπολογίζει την απόσταση ανάμεσα στις υποομάδες ως τον μέσο όρο μεταξύ των δύο υποομάδων.

Η αποστάσεις Δέλτα (Delta distances)

- Delta, όπως ορίστηκε από τον Burrows (2002), είναι μια μετρική απόστασης. Περιγράφει την απόσταση μεταξύ ενός κειμένου και μιας ομάδας κειμένων. Αυτή η ομάδα μπορεί να θεωρηθεί ως η αναπαράσταση του ύφους μιας περιόδου, ενός κειμενικού γένους μιας συγκεκριμένης περιόδου. Ο Burrows μέσα από αυτή την προσέγγιση θεωρεί το ύφος ως απόκλιση από την νόρμα (Rosengren, 1972).
- Η Delta είναι στην ουσία η απόσταση μεταξύ των διανυσματικών αναπαραστάσεων των κειμένων σε έναν χώρο υψηλής διαστασιμότητας όπου κάθε λέξη (ή άλλο γλωσσικό χαρακτηριστικό) που εξετάζεται αντιστοιχεί σε μία από τις διαστάσεις αυτού του χώρου.

Υπολογισμός των Delta

- Η κειμενική αναπαράσταση που χρησιμοποιεί ο Burrows είναι ένα 'bag of words' μοντέλο, δηλ. μετράμε πόσο συχνά εμφανίζεται η κάθε λέξη σε κάθε κείμενο.
- Οι μετρήσεις των λέξεων μετατρέπονται σε σχετικές συχνότητες για να λάβουν υπόψη τους τα διαφορετικά κειμενικά μεγέθη. Για περαιτέρω επεξεργασία επιλέγονται οι ν πιο συχνές λέξεις (nMFW).
- Στην διανυσματική αναπαράσταση, η κάθε λέξη πλέον αντιστοιχεί σε μια διαφορετική διάσταση. Οι λεξικές συχνότητες των κειμένων αποτυπώνονται σε ένα πίνακα κειμένων - λέξεων (documents words matrix).

Υπολογισμός των Delta

- Ο Burrows (2002) τυποποιεί (standardizes) τις λεξικές συχνότητες, δηλ. κανονικοποιεί τις συχνότητες έτσι ώστε σε όλο το corpus, ο μέσος όρος κάθε λέξεις να είναι 0 και η τυπική απόκλιση 1 (η κανονικοποίηση αυτή ονομάζεται και μετατροπή σε 'z-score'). Αυτό μειώνει την επίδραση τον λέξεων που εμφανίζονται με τις μεγαλύτερες τιμές. Αφού οι λεξικές συχνότητες ακολουθούν το νόμο του Zipf (Zipf 1935), η απόσταση θα επηρεαζόταν κυρίως από λίγες πολύ υψίσυχνές λέξεις.
- Μόλις κανονικοποιηθούν τα κειμενικά διανύσματα (document vectors), υπάρχουν συγκεκριμένοι τρόποι για να υπολογίσουμε την απόσταση μεταξύ δύο κειμένων που αναπαριστώνται από τα διανύσματα u και v αντίστοιχα.
- Οι Delta αποστάσεις, στην ουσία είναι αποστάσεις Manhattan. Ωστόσο, ύστερες ερευνητικές προσπάθειες χρησιμοποίησαν και άλλες αποστάσεις με καλύτερα αποτελέσματα (π.χ. Cosine Distance).

Αποστάσεις

- Η απόσταση 'Manhattan' (η L1 norm του διανύσματος διαφοροποίησης) αθροίζει τις απόλυτες αποστάσεις της κανονικοποιημένης συχνότητας κάθε λέξης μεταξύ των δύο κειμένων.
- Η 'Ευκλίδεια απόσταση' (η L2 norm του διανύσματος διαφοροποίησης) υπολογίζει την απόσταση ως ευθεία γραμμή μεταξύ των διανυσμάτων
- Η 'απόσταση συνημίτονου' (cosine distance) αντιστοιχεί στην γωνία φ μεταξύ των διανυσμάτων (η οποία στην ουσία είναι ισοδύναμη με την κανονικοποίηση των διανυσμάτων ως προς το μήκος και τον υπολογισμό της Ευκλίδειας απόστασης σε αυτά).