



Τεχνητή Νοημοσύνη

8η διάλεξη (2026-27)

Ίων Ανδρουτσόπουλος

<http://www.aueb.gr/users/ion/>

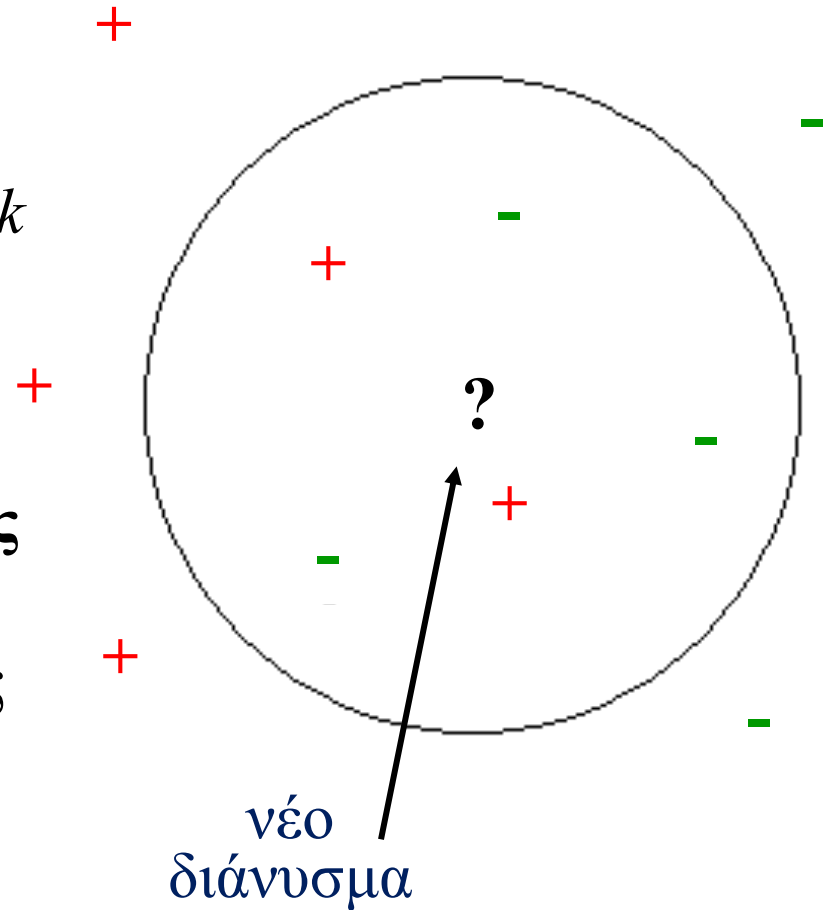
Οι διαφάνειες αυτής της διάλεξης βασίζονται σε ύλη των βιβλίων (α) *Artificial Intelligence – A Modern Approach* των S. Russel και P. Norvig, 2^η και 4^η έκδοση, Prentice Hall, 2003 και 2020, (β) *Τεχνητή Νοημοσύνη* των Βλαχάβα κ.ά., 3^η έκδοση, Β. Γκιούρδας Εκδοτική, 2006 και (γ) *Machine Learning* του T. Mitchell, McGraw-Hill, 1997. Τα περισσότερα σχήματα των διαφανειών βασίζονται σε αντίστοιχα σχήματα των διαφανειών που συνοδεύουν τα πρώτα δύο βιβλία.

Τι θα ακούσετε σήμερα

- Αλγόριθμος των k -κοντινότερων γειτόνων.
- Μη επιβλεπόμενη ομαδοποίηση με τον k -means.
- Αφελείς ταξινομητές Bayes.
- Δεδομένα εκπαίδευσης, ανάπτυξης, ελέγχου.

Αλγόριθμος k κοντινότερων γειτόνων (k -NN)

- Κατά την **εκπαίδευση** απλά **αποθηκεύουμε** τα διανύσματα εκπαίδευσης.
- Κατά τη **χρήση**, βρίσκουμε στα παραδείγματα εκπαίδευσης τους k κοντινότερους **γείτονες** του νέου διανύσματος (π.χ. $k = 5$).
- Κατατάσσουμε το νέο διάνυσμα στην κατηγορία της **πλειοψηφίας** των γειτόνων (εδώ «-»).
- Σε **προβλήματα παλινδρόμησης** (regression) που οι **αποκρίσεις** είναι **πραγματικοί αριθμοί**, επιστρέφουμε π.χ. τη **μέση τιμή** των γειτόνων.



Μέτρο απόστασης

- Παραδείγματα μέτρων απόστασης:
 - **Ευκλείδεια απόσταση**
 - ή για δίτιμες ιδιότητες, σε πόσες θέσεις διαφέρουν τα δύο διανύσματα.

$$d(\vec{x}_i, \vec{x}_j) \equiv \sum_{r=1}^m \delta(x_{ir}, x_{jr})$$

$$\delta(x, y) \equiv \begin{cases} 0, & \text{αν } x = y \\ 1, & \text{διαφορετικά} \end{cases}$$

- Βελτίωση: οι διαφορές σταθμίζονται ανάλογα με το **πόσο σημαντική** είναι η **ιδιότητα** (**attribute weighting**):

$$d(\vec{x}_i, \vec{x}_j) \equiv \sum_{r=1}^m IG(C, X_r) \cdot \delta(x_{ir}, x_{jr})$$

Παράδειγμα

- Έχουν αποθηκευτεί τα ακόλουθα διανύσματα εκπαίδευσης:

$\langle 1, 0, 0, 1, \mathbf{1} \rangle$

$\langle 1, 1, 0, 1, \mathbf{1} \rangle$

$\langle 0, 0, 1, 1, \mathbf{0} \rangle$

$\langle 0, 0, 1, 0, \mathbf{0} \rangle$

Ο τελευταίος αριθμός είναι η κατηγορία.

- Φτάνει **νέο** διάνυσμα:

$\langle 1, 1, 0, 0, ? \rangle$

- Σε **ποια κατηγορία** κατατάσσεται το νέο διάνυσμα;
 - Για $k = 3$ και με ισοβαρείς ιδιότητες.

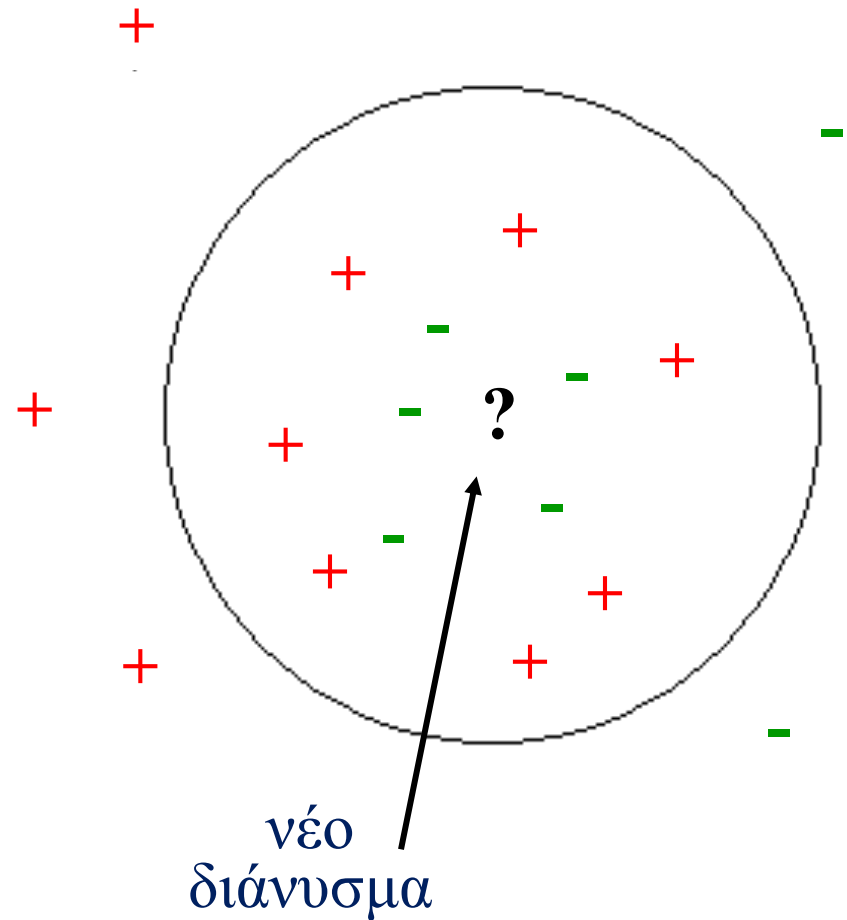
Βάρος απόστασης (distance weighting)

Περαιτέρω βελτίωση:

- Η **ψήφος** κάθε γείτονα $\vec{x}_i$ έχει **βάρος** π.χ. αντιστρόφως ανάλογο της **απόστασής** του από το νέο διάνυσμα $\vec{x}$.
- Δηλαδή αφού βρούμε τους k κοντινότερους γείτονες, πολλαπλασιάζουμε την ψήφο κάθε γείτονα $\vec{x}_i$ με:

$$1/d(\vec{x}_i, \vec{x})$$

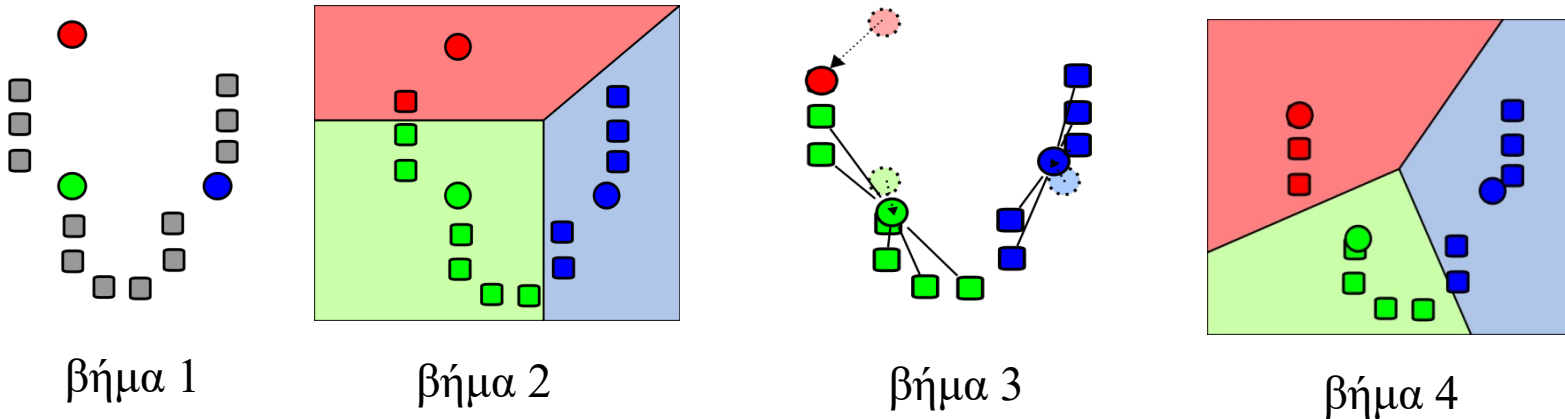
- Κάνει **λιγότερο σημαντική την επιλογή του k** . Οι μακρινοί δεν μετράνε πολύ.



Χαρακτηριστικά k -NN

- Πολύ **μικρό υπολογιστικό κόστος εκπαίδευσης**.
 - Απλά αποθηκεύουμε τα παραδείγματα εκπαίδευσης.
- Αλλά **μεγάλες απαιτήσεις μνήμης**.
 - Αποθηκεύουμε όλα τα παραδείγματα εκπαίδευσης.
- Και **μεγάλο υπολογιστικό κόστος κατά την αποκρίση**.
 - Χρειάζεται να υπολογίζουμε κάθε φορά τις αποστάσεις από όλα τα παραδείγματα εκπαίδευσης.
 - Υπάρχουν **προσεγγίσεις** που **μειώνουν** πολύ αυτό το **κόστος**.
- Μπορεί να προσεγγίσει **οποιαδήποτε συνάρτηση**.
 - Όχι μόνο π.χ. γραμμικές συναρτήσεις.
- **Μη παραμετρική μέθοδος μάθησης**.
 - Δεν μαθαίνουμε τιμές **σταθερού πλήθους παραμέτρων**. Όσο αυξάνονται τα παραδείγματα, περισσότερη μνήμη.

Ομαδοποίηση με τον k -means



- **Ξεκίνα** με k **τυχαία κέντρα βαρών** (ένα για κάθε ομάδα).
 - **Χρωμάτισε**: θεώρησε ότι κάθε στιγμιότυπο ανήκει στην ομάδα με το κοντινότερο κέντρο βάρους.
 - Υπολόγισε το **νέο κέντρο βάρους** κάθε ομάδας.
 - **Επαναλήψεις** ως σύγκλιση...
- Στην πράξη χρησιμοποιείται πολύ. Πρέπει να ξέρουμε το k .
- Προσπαθεί να ελαχιστοποιήσει τις **συνολικές αποστάσεις** των στιγμιότυπων **από τα κέντρα βαρών** των ομάδων τους.
- Μπορεί να παγιδευτεί σε τοπικό ελάχιστο. Ευαίσθητος στα αρχικά κέντρα βαρών. **Επανεκκινήσεις** με διαφορετικά αρχικά κέντρα βαρών.

Παράδειγμα αλγορίθμου
μη επιβλεπόμενης
μάθησης

Αφελείς ταξινομητές Bayes (Naive Bayes)

- Παράδειγμα: εισερχόμενο μήνυμα, παριστάνεται ως:

$$\vec{X} = \langle X_1, X_2, \dots, X_m \rangle = \langle 0, 1, \dots, 1 \rangle$$

- Συνάρτηση ταξινόμησης:

$$h(\vec{X}) = 1, \text{ ανν } P(C = 1 | \vec{X}) > P(C = 0 | \vec{X})$$

- Με το θεώρημα του Bayes (εδώ $c = 0$ ή $c = 1$):

$$P(C = c | \vec{X}) = \frac{P(C = c) \cdot P(\vec{X} | C = c)}{P(\vec{X})}$$

Πρέπει να εκτιμηθούν κατά την εκπαίδευση οι πιθανότητες όλων των συνδυασμών $x_1, x_2, \dots, x_m | c$. Πάρα πολλοί (2^m). Και πολλοί είναι σπάνιοι στα δεδομένα μας, θα παίρνουμε μηδενικές πιθανότητες.

Παραδοχή ανεξαρτησίας

- Οι αφελείς ταξινομητές Bayes κάνουν την παραδοχή ότι οι τιμές των $X_1, \dots, X_m$ είναι **ανεξάρτητες δεδομένης της τιμής της C** .
 - Συνήθως δεν ισχύει, αλλά στην πράξη καλά αποτελέσματα.

$$P(\vec{X} = \langle x_1, x_2, \dots, x_m \rangle \mid C = c) =$$

$$P(X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_m = x_m \mid C = c) \approx$$

$$P(X_1 = x_1 \mid C = c) \cdot \dots \cdot P(X_m = x_m \mid C = c) =$$

$$\prod_{i=1}^m P(X_i = x_i \mid C = c)$$

Αφελείς ταξινομητές Bayes – συνέχεια

- Τότε:

$$P(C = 1 | \vec{X}) = \frac{P(C = 1) \cdot \prod_{i=1}^m P(X_i = x_i | C = 1)}{P(\vec{X})}$$

$$P(C = 0 | \vec{X}) = \frac{P(C = 0) \cdot \prod_{i=1}^m P(X_i = x_i | C = 0)}{P(\vec{X})}$$

- Τώρα όλες οι πιθανότητες μπορούν να εκτιμηθούν εύκολα από τα παραδείγματα εκπαίδευσης.
- Οι παρονομαστές δεν μας χρειάζονται, γιατί είναι ίδιοι.

Εξομάλυνση πιθανοτήτων

- Προσπαθούμε να εκτιμήσουμε την $P(X_i = x_i \mid C = c)$.
 - 1^η προσέγγιση: Σε πόσα μηνύματα εκπαίδευσης κατηγορίας c εμφανίζεται (για $X_i = 1$) ή όχι ($X_i = 0$) η λέξη της X_i ;
 - Τι γίνεται όμως αν το $X_i = x_i$ δεν συμβαίνει σε κανένα μήνυμα εκπαίδευσης της c ; **Μηδενική εκτίμηση**.
 - Μηδενίζεται και ολόκληρο το $\prod_{i=1}^m P(X_i = x_i \mid C = c)$
 - Μηδενική $P(C = c \mid \bar{X})$ λόγω μόνο μιας ιδιότητας.
- Ένας τρόπος εξομάλυνσης: **εκτιμήτρια Laplace**.
 - Θεωρούμε κατά την εκτίμηση της $P(X_i = x_i \mid C = c)$ ότι υπάρχουν δύο ακόμη ψευτο-μηνύματα εκπαίδευσης κατηγορίας c : ένα που περιέχει τη λέξη της X_i και ένα που δεν την περιέχει.
 - Επομένως +1 στον αριθμητή της εκτίμησης, +2 στον παρονομαστή.
 - Γενικότερα, κατά την εκτίμηση τυχαίας μεταβλητής με k δυνατές τιμές, +1 στον αριθμητή και + k στον παρονομαστή.

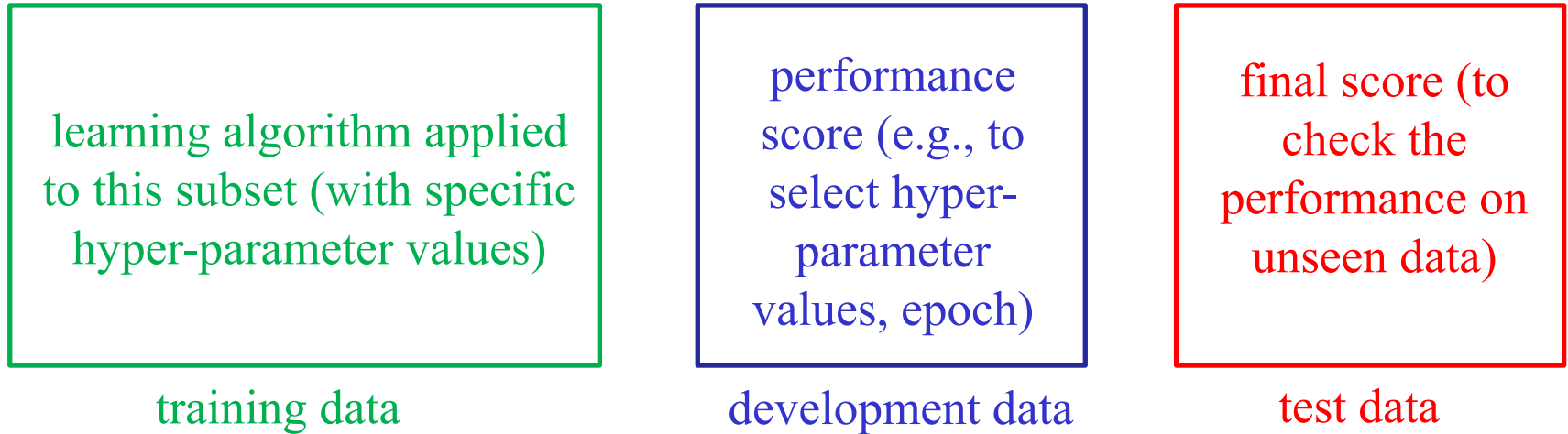
Χαρακτηριστικά Naive Bayes

- Πολύ μικρό υπολογιστικό κόστος:
 - $O(mN)$ κατά την εκπαίδευση για την εκτίμηση των πιθανοτήτων $P(X_i|C)$,
 - $O(m)$ κατά την κατάταξη για τον υπολογισμό του γινομένου των $P(X_i|C)$,
 - όπου N το πλήθος των παραδειγμάτων εκπαίδευσης και m το πλήθος των ιδιοτήτων.
- Πολύ μικρές απαιτήσεις μνήμης:
 - $O(m)$ για την αποθήκευση των εκτιμήσεων των $P(X_i|C)$.
- Παραμετρική μέθοδος μάθησης.
 - Μαθαίνουμε τιμές σταθερού πλήθους παραμέτρων.
 - Παράμετροι: όλες οι πιθανότητες $P(X_i|C)$.

Training, development, test data

- **Training data:**
 - We **train our model** on these data.
- **Development data:**
 - Used to **tune hyper-parameters** (e.g., k value in k -NN), select best **training epochs** (in neural networks) etc.
 - **If we make these choices by evaluating on test data, we indirectly train our model on the test dataset!**
- **Test data:**
 - Used for the final **evaluation of our model**, to see how well it performs on **unseen data**.

Training, development, test data



- In competitions, the **test data may not be publicly available**.
- We may have to use the **development data as test data**.
- A **small subset** of the **training data** may have to be “**held out**” as development data (e.g., for hyper-parameter tuning).
 - Often called “**validation data**”.
 - This **reduces** the size of the **training set**.
 - And a **small development set, may not be representative**.

Cross-validation

- **Instead of holding out data** from the training data:
 - Divide the training data into **n parts** (e.g., 5), often **preserving class ratios** (e.g., positives/negatives) in all parts (“stratified”).
 - Perform **n iterations** (folds) to obtain a **score** (e.g., accuracy) for a particular **combination of hyper-parameter values**.
 - In each iteration, use a **different part as validation data** (to obtain a score) and the **other $n - 1$ parts as training data**.
 - **Average** the score (e.g., accuracy) **over iterations** to obtain a score for the particular combination of hyper-parameter values.



Θα πρέπει να μελετήσετε τις διαφάνειες αυτής της διάλεξης και τις ασκήσεις μελέτης της (βλ. έγγραφα μαθήματος στο e-class) πριν από την επόμενη διάλεξη και πριν από το αντίστοιχο εργαστήριο. Γράψτε μόνοι σας τις λύσεις των ασκήσεων μελέτης (ακόμα και αν δίνονται οι λύσεις). Επίσης, γράψτε μόνοι σας τους ορισμούς, αλγορίθμους, παραδείγματα εκτέλεσης των αλγορίθμων και λοιπούς υπολογισμούς των διαφανειών. Ανατρέξτε στη βιβλιογραφία για πρόσθετες εξηγήσεις και γνώσεις (ιδιαίτερα όσοι ενδιαφέρεστε ιδιαίτερα για την TN).

Βιβλιογραφία

- Russel & Norvig (4^η έκδοση): ενότητες 19.7.1, 20.2.2.
- Βλαχάβας κ.ά: ενότητες 18.7, 18.8, υπο-ενότητα «Αλγόριθμος k-μέσων» της ενότητας 18.13.2.
- Aggarwal: σχετική ύλη υπάρχει στα κεφάλαια 6 και 9.
- Poole & Mackworth: σχετική ύλη υπάρχει στο κεφάλαιο 7.

Βιβλιογραφία – συνέχεια

- Στα εργαστήρια θα μάθετε και για το scikit-learn (Python).
 - <https://scikit-learn.org/>
- Δείτε προαιρετικά και το εισαγωγικό βιβλίο «A Course in Machine Learning» του Hal Daumé III.
 - Διατίθεται ελεύθερα. Βλ. <http://ciml.info/>
- Για μεθόδους πολύ γρήγορης προσεγγιστικής ανάκτησης κοντινότερων γειτόνων, δείτε π.χ.:
 - <http://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces/>

Βιβλιογραφία – συνέχεια

- Υπάρχουν πολλές μορφές του αφελούς ταξινομητή Bayes.
 - Η μορφή που εξετάσαμε χρησιμοποιεί **δίτιμες (Boolean) ιδιότητες** και λέγεται «**πολυμεταβλητή μορφή Bernoulli**» (multivariate Bernoulli Naive Bayes).
 - Η **πολυωνυμική (multinomial)** μορφή του NB μπορεί να λάβει υπόψη της και τις **συχνότητες των λέξεων** σε κάθε κείμενο κατά την κατάταξη κειμένων. Για περισσότερες πληροφορίες, βλ. http://www.aueb.gr/users/ion/docs/ceas2006_paper.pdf.
 - Περισσότερες πληροφορίες για τις μορφές του NB και άλλα θέματα που αναφέραμε (π.χ. κέρδος πληροφορίας) παρέχονται στο κεφάλαιο 13 του βιβλίου «An introduction to Information Retrieval» των C.D. Manning, P. Raghavan και H. Schütze. Διατίθεται ελεύθερα: <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>

