UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

Ανάκληση Πληροφορίας
Δημήτριος Κατσαρός

Ημέρα ανακοίνωσης: Friday, March 27, 2020

**Πρόβλημα-01**

Ένα μέτρο ομοιότητας μεταξύ δυο διανυσμάτων είναι η Ευκλείδεια απόσταση μεταξύ τους

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}$$

Δεδομένου ενός ερωτήματος q και των εγγράφων d1, d2, . . ., μπορούμε να διατάξουμε (rank) τα έγγραφα di με βάση την αυξανόμενη Ευκλείδεια απόστασή τους από το q. Δείξτε (με μαθηματική απόδειξη) ότι εάν q και di είναι όλα κανονικοποιημένα σε μοναδιαία διανύσματα (unit vectors), τότε η διάταξη που επιβάλλεται από την Ευκλείδεια απόσταση είναι πανομοιότυπη με αυτή που επιβάλλεται από την ομοιότητα συνημιτόνου (cosine similarity).

**Λύση-01**

$$\sum(q_i - w_i)^2 = \sum q_i^2 - 2\sum q_i w_i + \sum w_i^2 = 2(1 - \sum q_i w_i)$$

Επομένως:

$$\sum(q_i - v_i)^2 < \sum(q_i - w_i)^2 \Leftrightarrow 2(1 - \sum q_i v_i) < 2(1 - \sum q_i w_i) \Leftrightarrow \sum q_i v_i > \sum q_i w_i$$

**Πρόβλημα-02**

Δίνεται η ακολουθία bits: 0001’ 0000’ 0101’ 0000’ 0100’ 0000’ 0000’ 0001’ 0001’ 1111’ 1111’ 1111’ (οι απόστροφοι δεν είναι μέρος της ακολουθίας, έχουν εισαχθεί για να διευκολύνουν την αναγνωσιμότητα) που είναι κωδικοποιημένη και προήλθε από την κωδικοποίηση κατά Group VarInt μιας posting λίστας ενός όρου. Να αποκαταστήσετε την αρχική posting list εάν γνωρίζετει ότι αυτό που έχει κωδικοποιηθεί είναι τα κενά (gap encoding) και ότι ο πρώτος αριθμός της ακολουθίας που θα αποκωδικοποιηθεί δεν είναι κενό, αλλά το πρώτο docID της posting list.

**Λύση-03**

Είναι η λίστα: 80, 400, 431, 686. [Το πρώτο ID είναι το 80, και τα κενά είναι: 320, 31, 255.]

**Πρόβλημα-03**

Θεωρήστε την συλλογή των κάτωθι 4 εγγράφων (ένα έγγραφο σε κάθε γραμμή):
d1. John gives a book to Mary
d2. John who reads a book loves Mary
d3. who does John think Mary love?
d4. John thinks a book is a good gift
Αυτά τα έγγραφα υφίστανται προ-επεξεργασία με χρήση stop-word list και ενός stemmer. Ο προκύπτων index χτίζεται έτσι ώστε να επιτρέπει την απάντηση vector space-based ερωτημάτων.

- Δώστε την αναπαράσταση του index.

- Εστιάζουμε σε 3 terms που ανήκουν στο dictionary, δηλαδή στα book, love και Mary. Υπολογίστε την αναπαράσταση ως διάνυσμα με χρήση (tf–idf) για τα 4 έγγραφα της συλλογής (αυτά τα έγγραφα πρέπει να κανονικοποιηθούν με την Ευκλείδια κανονικοποίηση). [Για το tf χρησιμοποιήστε απλά τον αριθμό εμφανίσεων, ενώ για το idf τον λογάριθμο του κλάσματος που παρουσιάσαμε στις διαλέξεις και αναφέρεται στο βιβλίο σας.]
- Θεωρήστε το ερώτημα 'love Mary'. Δώστε τα αποτελέσματα μιας διαβαθμισμένης ανάκτησης (ranked retrieval) για το ερώτημα. Ποιο έγγραφο είναι το πιο σχετικό;

**Λύση-04**

| term $t$ | $N/df_t$ | $\rightarrow$ | d1 : $tf_{t,d_1}$ | d2 : $tf_{t,d_2}$ | d3 : $tf_{t,d_3}$ | d4 : $tf_{t,d_4}$ |
|---|---|---|---|---|---|---|
| book | 4/3 | $\rightarrow$ | d1 :1 | d2 :1 | d4 :1 | |
| gift | 4/1 | $\rightarrow$ | d4 :1 | | | |
| give | 4/1 | $\rightarrow$ | d1 :1 | | | |
| good | 4/1 | $\rightarrow$ | d4 :1 | | | |
| John | 4/4 | $\rightarrow$ | d1 :1 | d2 :1 | d3 :1 | d4 :1 |
| love | 4/2 | $\rightarrow$ | d2 :1 | d3 :1 | | |
| Mary | 4/3 | $\rightarrow$ | d1 :1 | d2 :1 | d3 :1 | |
| read | 4/1 | $\rightarrow$ | d2 :1 | | | |
| think | 4/2 | $\rightarrow$ | d3 :1 | d4 :1 | | |

$$v(\vec{d_1}) = \begin{pmatrix} (book) & \frac{1\times log(4/3)}{D_1} \\ (love) & 0 \\ (Mary) & \frac{1\times log(4/3)}{D_1} \end{pmatrix} \text{ where } D_1 = \sqrt{(log(4/3))^2 + 0 + (log(4/3))^2}$$

$$v(\vec{d_2}) = \begin{pmatrix} (book) & \frac{1\times log(4/3)}{D_2} \\ (love) & \frac{1\times log(4/2)}{D_2} \\ (Mary) & \frac{1\times log(4/3)}{D_2} \end{pmatrix} \text{ where } D_2 = \sqrt{(log(4/3))^2 + (log(4/2))^2 + (log(4/3))^2}$$

$$v(\vec{d_3}) = \begin{pmatrix} (book) & 0 \\ (love) & \frac{1\times log(4/2)}{D_3} \\ (Mary) & \frac{1\times log(4/3)}{D_3} \end{pmatrix} \text{ where } D_3 = \sqrt{0 + (log(4/2))^2 + (log(4/3))^2}$$

$$v(\vec{d_4}) = \begin{pmatrix} (book) & \frac{1\times log(4/3)}{D_4} \\ (love) & 0 \\ (Mary) & 0 \end{pmatrix} \text{ where } D_4 = \sqrt{(log(4/3))^2 + 0 + 0}$$

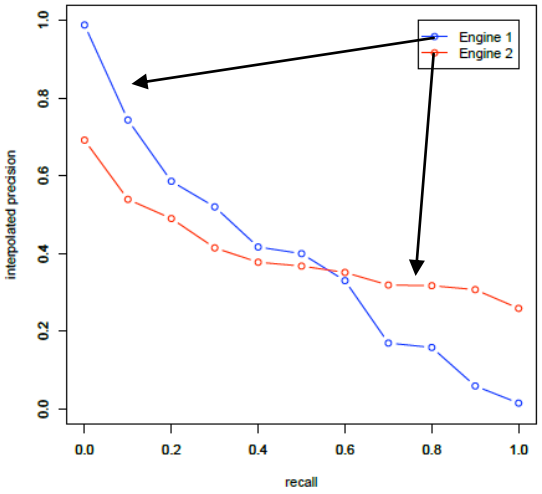
$$v(\vec{q}) = \begin{pmatrix} (book) & 0 \\ (love) & 1 \\ (Mary) & 1 \end{pmatrix}$$

Πιο σχετικό έγγραφο: d3

$$s(q, d_3) = \frac{log(4/2)}{D_3} + \frac{log(4/3)}{D_3}$$

**Πρόβλημα-04**

- Το πλαϊνό σχήμα δείχνει τις interpolated precision-recall καμπύλες για δυο μηχανές αναζήτησς που κάνουν index επιστημονικά άρθρα (για ένα συγκεκριμένο θέμα). Δεν υπάρχει καμία διαφορά μεταξύ των μηχανών, εκτός από το πώς κάνουν την διαβάθμιση των σχετικών άρθρων. Φανταστείτε ότι είστε κάποιος επιστήμονας που ψάχνει για όλα τα δημοσιευμένα άρθρα πάνω στο συγκεκριμένο αντικείμενο. Δεν θέλετε να χάσετε κανένα σχετικό άρθρο. Ποια από τις δυο μηχανές θα προτιμούσατε να χρησιμοποιήσετε, και γιατί;




**Λύση-05**

Θα επιλέγαμε την δεύτερη, γιατί παρόλο που και οι δυο βρίσκουν όλα τα σχετικά άρθρα (έχουν recall=1.0) εντούτοις η δεύτερη τα βρίσκει πιο γρήγορα (στα μεγάλα recall levels έχει υψηλότερο precision).

**Πρόβλημα-06**

Η χρήση του inverse document frequency (IDF) στο vector model για την ανάκληση πληροφορίας μπορεί να οδηγήσει στην ακόλουθη ανωμαλία: Μπορεί να υπάρχουν έγγραφα D και E και συλλογές A και B, όπου και οι δυο A και B περιέχουν και το D και το E, και ένα ερώτημα Q τέτοιο ώστε:

- Εάν το Q τίθεται στο πλαίσιο της συλλογής A, το D να κρίνεται ως πιο σχετικό ως προς το Q παρά το E.
- Εάν το Q τίθεται στο πλαίσιο της συλλογής B, το E να κρίνεται ως πιο σχετικό ως προς το Q παρά το D.

Εξηγήστε πώς μπορεί να συμβεί αυτό.

**Λύση-07**

Υποθέστε ότι το ερώτημα Q αποτελείται από δυο όρους, X και Y, όπου:

- ο X είναι συχνός στο έγγραφο D, αλλά σπάνιος στο έγγραφο E,

και

- ο Y είναι συχνός στο έγγραφο E, αλλά σπάνιος στο έγγραφο D

και

- ο X είναι σπάνιος στα έγγραφα της συλλογής A, αλλά κοινός στην συλλογή B,

και

- ο Y είναι κοινός στην συλλογή A, αλλά σπάνιος στην συλλογή B.

Τότε ο IDF για το X στην A είναι μεγάλος, και ο IDF για το Y στην A είναι μικρός, και συνεπώς ο X θα έχειμεγαλύτερο βάρος από ότι ο Y για ερωτήματα επί της συλλογής A, έτσι το D θα κατατάσσεται υψηλότερα από ότι το E για ερωτήματα επί της συλλογής A. Το αντίστροφο είναι αληθές για ερωτήματα επί της συλλογής B.