Ερωτήσεις επανάληψης

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

1. Οι δείκτες παράβλεψης (skip pointers) δεν βελτιώνουν την απόδοση στην περίπτωση των ερωτημάτων διάζευξης (or).

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

2. Το stemming (περικοπή κλιτικών καταλήξεων) μπορεί να οδηγήσει σε λιγότερα αλλά με μεγαλύτερο μήκος (δηλαδή, με περισσότερες καταχωρήσεις) αντεστραμμένα ευρετήρια.

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

3. Η ληματοποίηση δεν επηρεάζει το μέγεθος του λεξικού.

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

4. Έστω δύο φράσεις με δύο λέξεις κάθε μία: η φράση t1 t2 και η φράση t3 t4.

Οι δυο φράσεις είναι το ίδιο συχνές, ενώ οι όροι t1 και t2 από τους οποίους αποτελείται η πρώτη φράση είναι πιο συχνοί από τους όρους t3 και t4 από τους οποίους αποτελείται η δεύτερη φράση.

Σε αυτή την περίπτωση, η χρήση biwords ενδείκνυται περισσότερο (έχει περισσότερα οφέλη) για τη φράση t1 t2 από ότι για τη φράση t3 t4 από ότι ή χρήση ανεστραμμένου ευρετηρίου θέσης (positional inverted index).

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

5. Θεωρείστε ότι έχουμε ευρετήρια πεδίου, όπου διατηρούμε ένα διαφορετικό ευρετήριο για κάθε πεδίο (για παράδειγμα ένα ευρετήριο για εμφανίσεις των όρων στο τίτλο ενός εγγράφου και ένα διαφορετικό ευρετήριο για τις εμφανίσεις του όρου στο ίδιο το έγγραφο).

Τα ευρετήρια αυτά είναι ιδιαίτερα χρήσιμα για παραμετρικά ερωτήματα.

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

6. Για τον υπολογισμό βαθμού ανά-έγγραφο (document-at-a-time) είναι προτιμότερη η διάταξη των εγγράφων στις λίστες καταχωρήσεων με βάση τη συχνότητα εμφάνισης του όρου στα έγγραφα.

Για καθένα από τα παρακάτω πείτε αν ισχύει ή όχι.
Θα χρησιμοποιήσουμε το chat
Κάντε like στο **Ναι** (ισχύει) ή στο **Όχι** (δεν ισχύει)

7. Τα word embeddings είναι ανεξάρτητα από τη συλλογή κειμένων που χρησιμοποιούμε για την εκπαίδευση.

Για καθένα από τα παρακάτω διαλέξτε **A**, **B** ή **C**.
Θα χρησιμοποιήσουμε το chat
Κάντε like στην απάντηση που θεωρείτε σωστή.

Θεωρείστε μια συλλογή με 1.000.000 έγγραφα όπου κάθε έγγραφο περιέχει συνολικά 5.000 όρους από τους οποίους 2.000 είναι διακριτοί (δηλαδή, διαφορετικοί μεταξύ τους), ενώ υπάρχουν συνολικά 800.000 διακριτοί όροι.

(1) Πόσους όρους περιέχει το λεξικό;
A. 500.000
B. 10.000.000
C. 800.000

(2) Πόσες είναι οι θέσεις του δυαδικού πίνακα όρων-εγγράφων
A. 5 x 10^9
B. 2 x 10^9
C. 8 x 10^11

(3) Ποιο είναι το ποσοστό των μη μηδενικών τιμών σε αυτόν τον πίνακα;
A. 0.25%
B. 0.4%
C. 0.625%

Για καθένα από τα παρακάτω διαλέξτε **A**, **B** ή **C**.
Θα χρησιμοποιήσουμε το chat
Κάντε like στην απάντηση που θεωρείτε σωστή.

Θεωρείστε μια συλλογή με 1.000.000 έγγραφα όπου κάθε έγγραφο περιέχει συνολικά 5.000 όρους από τους οποίους 2.000 είναι διακριτοί (δηλαδή, διαφορετικοί μεταξύ τους), ενώ υπάρχουν συνολικά 800.000 διακριτοί όροι.

(4) Πόσες είναι οι λίστες καταχωρήσεων του ανεστραμμένου ευρετηρίου;
A. 1.000.000
B. 800.000
C. 10.000.000

(5) Πόσες είναι οι καταχωρήσεις (postings) του ανεστραμμένου ευρετηρίου;
A. 5 x 10^9
B. 2 x 10^9
C. 8 x 10^11

Για καθένα από τα παρακάτω διαλέξτε **A**, **B** ή **C**.
Θα χρησιμοποιήσουμε το chat
Κάντε like στην απάντηση που θεωρείτε σωστή.

Θεωρείστε μια συλλογή με 1.000.000 έγγραφα όπου κάθε έγγραφο περιέχει συνολικά 5.000 όρους από τους οποίους 2.000 είναι διακριτοί (δηλαδή, διαφορετικοί μεταξύ τους), ενώ υπάρχουν συνολικά 800.000 διακριτοί όροι.

(6) Πόσες είναι οι λίστες καταχωρήσεων του ανεστραμμένου ευρετηρίου με πληροφορία θέσης (positional index);
A. 1.000.000
B. 800.000
C. 10.000.000

(7) Πόσες είναι οι καταχωρήσεις (postings) του ανεστραμμένου ευρετηρίου με πληροφορία θέσης (positional index);
A. 5 x 10^9
B. 2 x 10^9
C. 8 x 10^11

Για καθένα από τα παρακάτω διαλέξτε **A**, **B** ή **C**.
Θα χρησιμοποιήσουμε το chat
Κάντε like στην απάντηση που θεωρείτε σωστή.

Έστω μια συλλογή που περιέχει τα ακόλουθα έγγραφα:
d1: a b c
d2: a a f d
d3: a c d e c a
d4: b e a b
d5 a b d

(1) Το διάνυσμα που αναπαριστά κάθε έγγραφο με στάθμιση tf-idf θα έχει
A. 2 διαστάσεις
B. Δε ξέρουμε
C. 6 διαστάσεις

(2) Η τιμή στη θέση 3 του διανυσματικής αναπαράστασης του d5 είναι
A. 0
B. 1
C. 0.5