2ο Σύνολο Ασκήσεων

Καταληκτική Ημερομηνία Παράδοσης: Τετάρτη 24 Απριλίου 2013, στην αρχή του μαθήματος
Ύλη: Κεφάλαια 5 (εκτός 5.3.2), 6 (εκτός των 6.1 και 6.4.4), 7 και 8 από το βιβλίο
Μπορεί να γίνει προφορική εξέταση

Άσκηση 1
(α) Θεωρείστε μια συλλογή εγγράφων απλού κειμένου.
(α) Έστω ο συνολικός αριθμός εμφανίσεων λέξεων (τα tokens) είναι 20.000.000. Ποιο είναι το εκτιμώμενο μέγεθος λεξιλογίου (πλήθος όρων);
(β) Έστω ότι ο πιο συχνά εμφανιζόμενος όρο; εμφανίζεται 3.600.000 φορές. Πόσες φορές εκτιμάτε ότι θα εμφανίζεται ο 6ος πιο συχνά εμφανιζόμενος όρος;

Άσκηση 2
Θέλετε να σχεδιάσετε ένα ΣΑΠ που να βασίζεται στο διανυσματικό μοντέλο για μια συλλογή κειμένων συνολικού μεγέθους στο δίσκο 1 Gigabyte. Έστω ότι το μέσο μέγεθος των όρων που εμφανίζονται στα κείμενα είναι 10 χαραχτήρες και ότι το πλήθος των διαφορετικών όρων της συλλογής είναι 10.000. Υποθέστε 200.000 έγγραφα.
(α) Ποιο το αναμενόμενο (μέγιστο) μέγεθος του ανεστραμμένου ευρετηρίου για τη συλλογή αυτή;
(β) Ποιο το αναμενόμενο (μέγιστο) μέγεθος του ανεστραμμένου ευρετηρίου αν χρησιμοποιήσετε block addressing με μέγεθος block ίσο με 200 όρους; Τι ποσοστό μείωσης έχουμε, σε σχέση με το (α);
(γ) Αν έπρεπε το ευρετήριο να καταλαμβάνει το πολύ 1 MB (π.χ. για να χωράει στην κύρια μνήμη ενός κινητού τηλεφώνου) πως θα σχεδιάζατε το ανεστραμμένο ευρετήριο;
(δ) Αν έπρεπε να καταλαμβάνει το πολύ 100 K τι θα κάνατε;
(ε) Αν έπρεπε να καταλαμβάνει το πολύ 10 K τι θα κάνατε;

Άσκηση 3
Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 5 έγγραφα:
Έγγραφο 1: «New York Times»
Έγγραφο 2: «New Times»
Έγγραφο 3: «Financial Times»
Έγγραφο 4: «High High Times»
Έγγραφο 5: «New Financial Times»

(α) Δώστε τη διανυσματική αναπαράσταση του κάθε εγγράφου με βάρη TF-IDF (για ευκολία θεωρήστε ότι IDF=N/DF και όχι IDF=log(N/DF)). Θεωρείστε ότι η θέση του κάθε όρου στα διανύσματα γίνεται αλφαβητικά.
(β) Θεωρείστε το ερώτημα q1=«high financial». Υπολογίστε το TF-IDF διάνυσμα αυτού του ερωτήματος και δώστε την διάταξη των εγγράφων που θα επιστρέψει ένα σύστημα που βασίζεται στο διανυσματικό μοντέλο.
(γ) Θεωρείστε τα ερωτήματα q2 = «high AND financial, q3 = «high OR financial» και δώστε τις απαντήσεις που θα επιστρέψει ένα σύστημα που βασίζεται στο Boolean μοντέλο.
(δ) Ποια δυο έγγραφα είναι ποιο όμοια μεταξύ τους με βάση το διανυσματικό μοντέλο;

Άσκηση 4
Θέλετε να ενημερωθείτε για το πρόβλημα του ακατάλληλου αλογίσιου κρέατος σε διάφορα προϊόντα στην Ελλάδα. Για το σκοπό αυτό διατυπώνεται αυτήν την πληροφοριακή σας ανάγκη ως το παρακάτω ερώτημα:

horse meat scandal Greece

Δώστε αυτό το ερώτημα στο google, yahoo και bing.

Θεωρείστε τα πρώτα 10 αποτελέσματα του ερωτήματος (δηλαδή, την πρώτη σελίδα των αποτελεσμάτων).

(α) Υπάρχουν κοινά αποτελέσματα και πόσα.

(β) Αξιολογείστε τη συνάφεια της κάθε σελίδας στο αποτέλεσμα χρησιμοποιώντας μόνο την περίληψη (snippet). Η αξιολόγηση θα είναι δυαδική, δηλαδή χαρακτηρίστε κάθε σελίδα ως Σ (συναφή) ή Ν (μη συναφή). Συγκρίνετε τις 3 μηχανές αναζήτησης χρησιμοποιώντας 2 διαφορετικά κατάλληλα μέτρα αξιολόγησης. Εξηγείστε γιατί επιλέξατε αυτά τα 2 μέτρα και ποια από τις 3 μηχανές αναζήτησης είναι καλύτερη για αυτό το ερώτημα.

(γ) Επαναλάβετε το (β) αφού δείτε και τις σχετικές σελίδες και σχολιάστε αν άλλαξε κάτι. Σχολιάστε επίσης αν κάποια από τις 3 μηχανές έχει καλύτερες περιλήψεις.

(δ) Χαρακτηρίστε τη κάθε σελίδα στο αποτέλεσμα του google χρησιμοποιώντας μόνο την περίληψη (snippet) με τώρα δίνοντας ένα βαθμό από το 0 έως το 9 (0 μη συναφής – 9 πολύ συναφής). Υπολογίστε το NDCG.

Άσκηση 5

Θεωρείστε μια συλλογή αξιολόγησης που αποτελείται από 40 έγγραφα $\{d_1,\ldots,d_{40}\}$. Η συλλογή αξιολόγησης περιλαμβάνει ένα ερώτημα *q* για το οποίο γνωρίζουμε ότι τα έγγραφα της συλλογής που είναι συναφή είναι 5, συγκεκριμένα τα $\{d_1, d_{11}, d_{18}, d_{21}, d_{33}\}$. Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων S1, S2 και S3. Για το λόγο αυτό υποβάλλουμε σε κάθε σύστημα το ερώτημα *q* και λαμβάνουμε τις εξής απαντήσεις με διάταξη από τα πιο συναφή στο λιγότερο από τα αριστερά προς τα δεξιά:

$Ans(S1,q) = <d_{11}, d_4, d_{18}, d_2, d_{21}, d_{33}, d_9, d_7, d_8, d_6, d_1, d_5>$

$Ans(S2,q) = < d_9, d_7, d_5, d_6, d_{11}, d_4, d_8, d_2, d_1, d_{33}, d_{18}, d_{21}>$

$Ans(S3,q) = < d_{18}, d_{33}, d_{11}, d_1, d_5, d_2 >$

Συγκρίνετε τα τρία αυτά συστήματα ως προς τα εξής μέτρα:

(α) F-Measure,

(β) R-Ακρίβεια (R-Precision)

(γ) Σχεδιάστε τις καμπύλες ακρίβειας/ανάκλησης (P/R curves). Για κάθε σύστημα δώστε 2 γραφήματα: ένα που να απεικονίζει τα P/R σημεία όπως προκύπτουν από τις απαντήσεις, και ένα χρησιμοποιώντας κανονικοποιημένα επίπεδα ανάκλησης (standard recall levels).

(δ) Με βάση την παραπάνω ανάλυση σχολιάστε την καταλληλότητα των 3 συστημάτων και ποιο θα επιλέγατε και γιατί.