

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Θέματα Ιουνίου 2014.

1α) Έστω μια συλλογή που αποτελείται από τα ακόλουθα κείμενα

d1 «χαριτωμένο σκυλάκι»

d2 «μεγάλο σκυλί»

d3 «χαριτωμένο άσπρο γατάκι»

d4 «άσπρη γάτα»

Αν η συλλογή αυτή έχει υποστεί τεχνική της περιστολής (stemming) και έχει οργανωθεί με ένα σύστημα ανάκτησης πληροφορίας που βασίζεται στο διωνυμικό μοντέλο, δώστε την κατάταξη των κειμένων όταν στο σύστημα υποβληθεί η ερώτηση «μεγάλο άσπρο σκυλί» θεωρήστε επίσης ότι η μέθοδος στάθμισης tf-idf στο συγκεκριμένο σύστημα είναι «ltc.ltc».

1β)

1γ) Για ποιες από τις ακόλουθες ερωτήσεις οι δείκτες παράλειψης (skip pointers) είναι χρήσιμοι και για ποιες άχρηστοι; Εξηγήστε τις απαντήσεις σας.

x AND y, όπου x είναι συχνός όρος και y είναι σπάνιος

x AND y, όπου x και y και οι δυο σπάνιοι όροι

i. x OR y, όπου x είναι συχνός όρος και y είναι σπάνιος

ii. x OR y, όπου x και y και οι δυο σπάνιοι όροι

2α) Απόσταση edit amazing και analyzing

2β) Η ακόλουθη λίστα επιστρέφει ως αποτέλεσμα μια αναζήτηση σε ένα σύστημα ανάκτησης πληροφορίας. Τα σύμβολα R και N δείχνουν τα πραγματικά σχετικά και άσχετα κείμενα στη λίστα αυτή.

R R N R N N R R N N N R N N N N N R N N R N N N N N N R N R N

i. ποια είναι η ακρίβεια του συστήματος στο 20^ο κείμενο της λίστας (top 20)

ii. ποια είναι η ανάκληση στο 20^ο κείμενο της λίστας (top 20)

iii. ποια η τιμή του μέτρου F1 στο 20^ο κείμενο της λίστας (top 20)

iv. σχεδιάστε διάγραμμα ακρίβειας ανάκλησης

v. ποια είναι η ακρίβεια παρεμβολής (interpolated precision) για επίπεδο ανάκλησης 33%.

2γ) Υποθέστε ότι έχουμε μια μεγάλη συλλογή κειμένου. Μετά από έλεγχο των πρώτων 1.000.000 λέξεων βρήκαμε 30.000 διαφορετικές λέξεις. Πόσες περίπου διαφορετικές λέξεις θα πρέπει να περιμένουμε στις πρώτες 100.000 λέξεις της συλλογής; Γιατί;

3α) Θεωρήστε τον ακόλουθο ιστό σελίδων

- Δώστε την εξίσωση PageRank όταν η πιθανότητα τηλεμεταφοράς (teleporting) είναι α.
- Υπολογίστε τις βαθμολογίες των σελίδων στην α=0.5.

