1η Σειρά Ασκήσεων

Μοντέλα Ανάκτησης
Ημερομηνία Παράδοσης: Δευτέρα 23 Μαρτίου 2009
Ομάδες των 2 ατόμων

Άσκηση 1

Θεωρείστε το παρακάτω κείμενα:

d1 : Ο κομήτης του Χάλλεϋ μας επισκέπτεται περίπου κάθε εβδομήντα έξι χρόνια.
d2 : Ο κομήτης του Χάλλεϋ πήρε το όνομά του από τον αστρονόμο Έντμοντ Χάλλεϋ.
d3 : Ένας κομήτης διαγράφει ελλειπτική τροχιά.
d4 : Ο πλανήτης Άρης έχει δύο φυσικούς δορυφόρους, το Δείμο και το Φόβο.
d5 : Ο πλανήτης Δίας έχει 63 γνωστούς φυσικούς δορυφόρους.
d6 : Ένας κομήτης έχει μικρότερη διάμετρο από ότι ένας πλανήτης.
d7 : Ο Άρης είναι ένας πλανήτης του ηλιακού μας συστήματος.

και το ερώτημα q = {κομήτης, Χάλεϋ}.

(α) Αναπαραστείστε τα κείμενα και τη ερώτηση χρησιμοποιώντας το λογικό (Boolean), διανυσματικό και πιθανοκρατικό μοντέλο και υπολογίστε την ομοιότητα τους με την ερώτηση. Θεωρείστε ως όρους μόνο τα ουσιαστικά και κύρια ονόματα.
(β) Αγνοώντας την απάντησή σας στο ερώτημα (α) πως θα κατατάσσατε εσείς τα έγγραφα.

Άσκηση 2

Έστω ότι έχουμε ένα μοντέλο ανάκτησης το οποίο βλέπει τα έγγραφα και τις επερωτήσεις ως σύνολα όρων. Συγκρίνετε τις ακόλουθες συναρτήσεις διαβάθμισης (ranking functions):

$R_1(d, q) = \frac{\lvert d \cap q \rvert}{\lvert q \cap d \rvert + \lvert q \setminus d \rvert}$	$R_3(d, q) = \frac{\lvert d \cap q \rvert}{\lvert d \setminus q \rvert + \lvert q \setminus d \rvert + \lvert d \cap q \rvert}$
$R_2(d, q) = \frac{\lvert d \cap q \rvert}{\lvert q \cap d \rvert + \lvert d \setminus q \rvert}$	$R_4(d, q) = 2 * \frac{\lvert d \cap q \rvert}{\lvert d \cup q \rvert}$

Άσκηση 3

Θεωρείστε ένα Σύστημα Ανάκτησης Πληροφοριών (ΣΑΠ) από μια μεγάλη συλλογή κειμένων. Θέλουμε να δώσουμε τη δυνατότητα χρήσης του ΣΑΠ μέσω κινητού τηλεφώνου. Για αυτό θέλουμε να ορίσουμε μια συνάρτηση διαβάθμισης η οποία να ευνοεί τα μικρά κείμενα, αφενός για να κρατήσουμε σε χαμηλά επίπεδα τον όγκο δεδομένων που θα μεταφέρονται και αφετέρου γιατί οι χρήστες κινητών τηλεφώνων προτιμούν τα μικρά κείμενα (εξαιτίας του μικρού μεγέθους της οθόνης). Θεωρείστε ότι οι επερωτήσεις των χρηστών είναι σάκοι λέξεων (bag of words). Σχεδιάστε μια συνάρτηση διαβάθμισης για το σκοπό αυτό για κάθε μια από τις παρακάτω περιπτώσεις
(α) Το ευρετήριο του ΣΑΠ έχει δυαδικά (0,1) βάρη (όπως για παράδειγμα το ευρετήριο του Boolean μοντέλου)
(β) Το ευρετήριο έχει βάρη TF-IDF.
Τεκμηριώστε τις προτάσεις σας (με αποδείξεις ή παραδείγματα).