4η Σειρά Ασκήσεων

Λειτουργίες Κειμένου. Ανάδραση.
Ημερομηνία Παράδοσης: Άσκηση 1&2&3 Τρίτη 24 Νοεμβρίου 2009 στο μάθημα
Άσκηση 4 Τρίτη 1 Δεκεμβρίου 2009 στο μάθημα

Άσκηση 1 (ατομική)

Άσκηση 2.1 (του online βιβλίου)

Άσκηση 2 (ατομική)

Άσκηση 9.5 (του online βιβλίου)

Άσκηση 3 (ατομική)

Θεωρείστε τα παρακάτω κείμενα (έγγραφα):
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.

Θεωρείστε ότι επιλέγουμε τους όρους K ={κομήτης, Χάλλεϋ, πλανήτης, δορυφόρος, τροχιά, σύστημα}. Αναπαραστήστε τα κείμενα χρησιμοποιώντας το διανυσματικό μοντέλο (vector model).
(α) Θεωρείστε την ερώτηση q_0 = {κομήτης, Χάλλεϋ}. Χρησιμοποιώντας ψευτο-ανάδραση και τα 2 κορυφαία έγγραφα, δώστε τη νέα ερώτηση και το αποτέλεσμα της μετά την ανάδραση.
(β) Κατασκευάστε τον πίνακα αυτό-συσχέτισης (correlation matrix) και τον κανονικοποιημένο πίνακα αυτό-συσχέτισης (normalized correlation matrix) για το σύνολο όρων K και τα έγραφα της απάντησης της ερώτησης q_0. Επαναλάβετε το ερώτημα (α) προσθέτοντας έναν όρο στην ερώτηση με βάση τον κανονικοποιημένο πίνακα.
(γ) Επαναλάβετε το ερώτημα (β) χρησιμοποιώντας καθολική ανάλυση.

Άσκηση 3 (2 Ομάδες)

Σχετικά με το Lucene http://lucene.apache.org/java/docs/index.html.

Θα χρησιμοποιείστε το Lucene για να κατασκευάσετε ένα ευρετήριο για μια συλλογή εγγράφων. Τα έγγραφα σας μπορεί να είναι έγγραφα κειμένου (text). Η συλλογή σας θα πρέπει να περιέχει τουλάχιστον 10 έγγραφα.

Οδηγίες για την κατασκευή του ευρετηρίου θα δοθούν στη σελίδα του μαθήματος.