Διαχείριση Περιεχομένου Παγκόσμιου Ιστού και Γλωσσικά Εργαλεία

Έτος 2012-2013 Εαρινό εξάμηνο

Ημερομηνία παράδοσης : 20-06-2013

Γενικά

Στόχος της άσκησης είναι η καταγραφή της μεταβολής στον χρόνο, της πολικότητας κειμένου που έχει αναρτηθεί στον Παγκόσμιο Ιστο γύρω από ένα θέμα. Στα πλαίσια της παρούσας άσκησης θα υλοποιήσετε ένα σύστημα το οποίο θα έχει την δυνατότητα να παρακολουθεί αναρτήσεις από διάφορους ιστοτόπους (forums, blogs, κλπ) και να εντοπίζει κείμενα (για παράδειγμα reviews προϊόντων) που σχετίζονται με το θέμα ενδιαφέροντος και να καταγράφει για κάθε ένα από αυτά την πολικότητας του, αξιοποιώντας το SentiWordNet. Η άσκηση χωρίζεται σε τρία μέρη.Συγκεκριμένα:

Μέρος Α

Εκπαιδεύστε ένα κατηγοριοποιητή της επιλογής σας προκειμένου να του δώσετε την δυνατότητα να αναγνωρίζει κειμενα που σχετίζονται με το θέμα ενδιαφέροντος που θα επιλέξετε. Για την εκπαίδευση, χρησιμοποιείτε κείμενα που είτε σχετίζονται με το θέμα ενδιαφέροντος που έχετε επιλέξει είτε όχι δημιουργώντας δύο κατηγορίες. Σαν χαρακτηριστικά εκπαίδευσης χρησιμοποιείστε τα θέματα (stems) των λέξεων που βρέθηκαν στο σώμα των κειμένων αυτών. Ιδανικά ο κατηγοριοποιητής θα πρέπει να έχει την δυνατότητα να αποφαίνεται για το εάν ένα κείμενο σχετίζεται με το θέμα ενδιαφέροντος ή όχι, λαμβάνοντας υπόψη μόνο το κείμενο αγκύρωσης(anchor text) ενός συνδέσμου προς το συγκεκριμένο κείμενο.

Μέρος Β

Κατασκευάστε ένα πρόγραμμα επιλεκτικής προσκόμισης ιστοσελίδων, το οποίο θα ελέγχει την ύπαρξη κειμένων που άπτονται της θεματολογίας ενδιαφέροντος, στην πρώτη σελίδα επιλεγμένων ιστοτόπων. Για κάθε ημέρα εκτέλεσης το πρόγραμμα θα έχει την δυνατότητα να εντοπίζει όλους τους υπερσυνδέσμους που οδηγούν σε κείμενα/σχόλια εντός του ιστότοπου και να απορρίπτει όλους τους υπερσυνδέσμους που οδηγούν εκτός αυτού. Επιπλέον θα πρέπει να απορρίπτονται σύνδεσμοι που έχει επεξεργαστεί το σύστημα σε προηγούμενη εκτέλεση. Στην συνέχεια θα αξιολογεί, χρησιμοποιώντας τον κατηγοριοποιητή που υλοποιήθηκε στο μέρος Α με είσοδο το κείμενο αγκύρωσης του συνδέσμου που οδηγεί στο άρθρο, το βαθμό εγγύτητάς του με το θέμα ενδιαφέροντος. Για κάθε κείμενο που θεωρείται σχετικό θα πρέπει να καταγράφονται οι εξής πληροφορίες:

- Το URL της ιστοσελίδας του άρθρου
- Το κείμενο του κυρίως άρθρου
- Η ημερομηνία ανακοίνωσης

Η έξοδος του συγκεκριμένου προγράμματος θα πρέπει να είναι μία λίστα αρχείων – ένα για κάθε ημερομηνία εκτέλεσης – κάθε ένα από τα οποία θα περιέχει μία λίστα εγγραφών. Η εγγραφή θα αναφέρεται σε κάθε κείμενο που ελέγχθηκε και βρέθηκε σχετικό με το θέμα ενδιαφέροντος και θα διατηρεί την ζητούμενη πληροφορία, όπως αυτή αναφέρεται στην παραπάνω λιστα.

Μέρος Γ

Για κάθε κείμενο (στην μορφή που έχει εξαγχθεί από το προηγούμενο στάδιο) που έχει βρεθεί σχετικό με το θέμα ενδιαφέροντος, σχολιάστε μορφοσυντακτικά το κυρίως κείμενο του άρθρου και εξάγετε τα επίθετα και τα επιρρήματα που μπορούν να βρεθούν στο WordNet, χρησιμοποιωντας τις βιβλιοθήκες του NLTK. Με την βοήθεια του SentiWordNet υπολογίστε το βαθμό πολικότητας για το κάθε επίθετο/επίρρημα. Στην συνέχεια υπολογίστε τον συνολικό βαθμό πολικότητας για κάθε κείμενο αθροίζοντας όλες τις επιμέρους πολικότητες.

Παρακολουθήστε για χρονικό διάστημα 10 ημερών, τουλάχιστον 3 διαφορετικές πηγές (ιστοσελίδες, forums, blogs) με κείμενο στην αγγλική γλώσσα.

Δημιουργήστε διαγράμματα στο οποία φαίνεται η αυξομείωση της βαθμολογίας της πολικότητας των δεδομένων ανά κατηγορία ενδιαφέροντος, συγκρίνοντας και τις διαφορετικές πηγές μεταξύ τους. Σχολιάστε τα αποτελέσματα. (αξονας x: χρόνος , αξονας y: βαθμολογία, ένα γράφημα για κάθε πηγή)

Τέλος, τα τρία μέρη θα πρέπει να εκτελούνται με ενιαίο τρόπο, και η έξοδος του ενός, να αποτελεί είσοδο για το επόμενο.

Σημειώσεις:

1. Η άσκηση να υλοποιηθεί σε ομάδες των 3 ατόμων και θα βαθμολογείται με το 80% του τελικού βαθμού(υπόλοιπο 20% από την προφορική εξέταση). Μπορεί να υλοποιηθεί και από ομάδες του ενός (60% άσκηση- 40% προφ. εξέταση)ή των δύο ατόμων(70% άσκηση- 30% προφ. εξέταση). Για την βαθμολόγηση της άσκησης θα ληφθούν υπόψιν τόσο η ολοκλήρωση με επιτυχία όλων των επιμέρους τμημάτων της άσκησης όσο και η αποτελεσματικότητα της προσέγγισής που επιλέξατε σε θέματα υλοποίησης.
2. Το θέμα ενδιαφέροντος το επιλέγετε εσεις. Καθορίζεται από το σύνολο των κειμένων που θα θεωρήσετε ως σχετικά και από το σύνολο των κειμένων που θα θεωρήσετε ως μη σχετικά.
3. Το σύστημα που θα υλοποιήσετε θα διαχειρίζεται κείμενα γραμμένα στην Αγγλική.
4. Οι επισκέψεις στο ίδιο ισότοπο θα πρέπει να πραγματοποιούνται με ήπιο τρόπο. Για παράδειγμα να αποφεύγετε να πραγματοποιείτε διαδοχικές αιτήσεις προς τον ίδιο εξυπηρετητή.
5. Η δομή των αρχείων εξόδου του συστήματος επιλεκτικής προσκόμισης θα καθοριστεί από κάθε ομάδα ξεχωριστά. Μπορείτε να χρησιμοποιήσετε είτε αρχεία xml, είτε κάποια άλλη μορφή διαχωρισμού της πληροφορίας – για παράδειγμα μία γραμμή ανά άρθρο και κάποιος ειδικός χαρακτήρας διαχωρισμού (πχ‘#’) δηλ: url#title#text#date

6. Προκειμένου να γνωρίζει το σύστημά σας ποιες διευθύνσεις url έχει επεξεργαστεί σε προηγούμενες εκτελέσεις, θα πρέπει να διατηρεί μία κατάλληλη «δομή» στην μνήμη η οποία θα αποθηκεύεται στον δίσκο μετά το τέλος της εκτέλεσης του προγράμματος ώστε να είναι διαθέσιμο σε επόμενη εκτέλεση. Επιλέξτε κατάλληλη δομή. Χρησιμοποιήστε είτε κάποια έτοιμη βιβλιοθήκη είτε γράψτε τις δικές σας συναρτήσεις. Το ζητούμενο είναι το σύστημά σας να μπορεί να διαχειριστεί τουλάχιστον 2000 διευθύνσεις url.
7. Μπορείτε να χρησιμοποιήσετε όποιο είδος κατηγοριοποιητή θέλετε. Τα παραδείγματα που θα ορίσουν το θέμα ενδιαφέροντος θα πρέπει να τα επιλέξετε μόνοι σας. Χρησιμοποιήστε τουλάχιστον 40-60 κείμενα ως παραδείγματα εκπαίδευσης. Το θέμα ενδιαφέροντος μπορείτε να το ορίσετε εσείς, επιλέγοντας κάποιο που να άπτεται μιας γενικής κατηγορίας. Πχ. κείμενα που σχετίζονται με το νέο μοντέλο κινητού τηλεφώνου χ.
8. Η γλώσσα υλοποίησης θα είναι η Python 2.7 σε συνδυασμό με το πακέτο NLTK.
9. Για την εξαγωγή των θεμάτων των λέξεων μπορείτε να χρησιμοποιησετε τον Porter stemmer που συμπεριλαμβάνεται στο πακέτο NLTK.
10. Η άσκηση μπορεί να παραδοθεί και τον Σεπτέμβριο.
11. Η παράδοση της άσκησης θα πραγματοποιηθεί ηλεκτρονικά αποστέλοντας την σε διεύθυνση που θα σας υποδειχθεί στο forum. Να έχετε και μία εκδοση του κώδικα κατά την προφορική εξέταση της άσκησης.
12. Για οποιαδηποτε άλλη απορία χρησιμοποιείστε το forum.

Παραδοτέα

1. Ο κώδικας για τα τρία μέρη ξεχωριστά. Ο κώδικας θα είναι σχολιασμένος
2. Τα αρχεία εξόδου που προέκυψαν από την εκτέλεση του μέρους Β
3. Μία πολύ σύντομη αναφορά που θα αναλύετε τις επιλογές σας καθώς και θα αναφέρετε τα συμπεράσματά σας από τον σχολιασμό των γραφημάτων που θα προκύψουν.