

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάριος Δ. Δικαϊάκος, Αναπληρωτής Καθηγητής

ΕΠΛ660: Ανάκτηση Πληροφοριών και Μηχανές Αναζήτησης

<http://www.cs.ucy.ac.cy/courses/epl660>

ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΕΡΓΑΣΙΑ 3η

Ημερομηνία ανακοίνωσης: 17/3/2010

Ημερομηνία Παράδοσης: **31/3/2010**

Στόχος

Να εξοικειωθούν οι φοιτητές με το Map-Reduce προγραμματιστικό μοντέλο.

Εκφώνηση

Το Map-Reduce είναι ένα framework λογισμικού που χρησιμοποιείται για την ανάπτυξη εφαρμογών που τρέχουν σε παράλληλα συστήματα βασισμένα σε μεγάλες συστοιχίες υπολογιστών (clusters). Το Map-Reduce μπορεί να χρησιμοποιηθεί για εφαρμογές που απαιτούν επεξεργασία τεράστιων όγκων δεδομένων, συνήθως σε κλίμακα πολλών terabytes. Το Map-Reduce χρησιμοποιείται αυτή την στιγμή από πλήθος μεγάλων εταιριών πληροφορικής όπως οι Google, Yahoo!, Amazon, Facebook, IBM για την αποδοτική διαχείριση των τεράστιων όγκων δεδομένων που διαθέτουν. Στα πλαίσια αυτής της εργασίας θα γίνει χρήση του Hadoop, μίας ανοικτού λογισμικού υλοποίησης του Map-Reduce.

Αυτή η εργασία περιλαμβάνει την **υλοποίηση ενός πληροφοριακού συστήματος ανάκτησης πληροφορίας για ψηφιακή βιβλιοθήκη**, όπως το Gutenberg Project. Αρχικά θα χρειαστεί να δημιουργήσετε ένα αντεστραμμένο ευρετήριο. Στη συνέχεια θα γράψετε ένα πρόγραμμα ερωτήσεων το οποίο θα δέχεται ένα σύνολο από όρους.

1. Κατασκευή αντεστραμμένου ευρετηρίου

Να υλοποιηθεί το αντεστραμμένο ευρετήριο με τη χρήση του Hadoop. Το πρόγραμμά σας πρέπει να διαθέτει τις ακόλουθες λειτουργικότητες:

- Εισαγωγή αρχείου κειμένου και δημιουργία αντεστραμμένου ευρετηρίου για τους όρους του αρχείου. Στη διαδικασία αυτή να χρησιμοποιήσετε απαλοιφή λέξεων αποκλεισμού (stop words).

- Πρόσθεση αρχείου κειμένου στο αντεστραμμένο ευρετήριο και ενημέρωση του ευρετηρίου.

2. Ερωτήματα

Να υλοποιηθεί ένα πρόγραμμα ερωτημάτων, το οποίο να υποστηρίζει **μόνο Boolean ερωτήματα (and, or, not)**. Ένα ερώτημα αποτελείται από έναν ή περισσότερους όρους αναζήτησης (keywords). Να υλοποιηθεί ένα απλό περιβάλλον διεπαφής χρήστη (όπως για παράδειγμα η διεπαφή του Google).

Οδηγίες

- Να υλοποιήσετε την άσκηση σε **ομάδες των 2 ατόμων**.
- Για την εισαγωγή αρχείου κειμένου χρησιμοποιείτε τα αρχεία που βρίσκονται στη διεύθυνση <http://www.cs.ucy.ac.cy/courses/EPL660/exercises/Shakespeare.zip>
- Μία λίστα με λέξεις αποκλεισμού μπορείτε να την κατεβάσετε από τη διεύθυνση: <http://www.cs.ucy.ac.cy/courses/EPL660/exercises/stoplist.txt> Αν θέλετε μπορείτε να δημιουργήσετε τη δικιά σας λίστα με λέξεις αποκλεισμού.
- Υλοποιείτε το πρόγραμμά σας με χρήση JAVA.

Χρήσιμες αναφορές

- Hadoop. <http://hadoop.apache.org/>
- MapReduce tutorial: http://hadoop.apache.org/common/docs/current/mapred_tutorial.html
- MapReduce. <http://labs.google.com/papers/mapreduce.html>.
- Project Gutenberg. <http://www.gutenberg.org/>.
- The Hadoop Distributed File System: Architecture and Design. <http://hadoop.apache.org/core/docs/current/hdfs design.html>.
- Tutorial #1: Hadoop and MapReduce: <http://www.adms.ethz.ch/teaching/lectures/spring08/NIS/assignment1/tutorial1.pdf>.
- D. Jeff and G. Sanjay. MapReduce: Simplified Data Processing on Large Clusters. In Sixth Symposium on Operating System Design and Implementation (OSDI'04), San Fransisco, CA, December 2004.
- D. Jeff and G. Sanjay. MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, 51(1), 2008.
- J. Zobel. Inverted Files for Text Search Engines. ACM Computing Surveys, 38(2), 2006.