HADOOP Project
Ακαδημαϊκό Έτος 2017-2018

Υπεύθυνος Εργασίας:
Δημήτριος Τυχάλας, Υποψ. Διδάκτορας, <dtychala@csd.auth.gr>

Άσκηση 1η:

Χρησιμοποιώντας το dataset της Wikipedia που θα βρείτε στο /tmp/wikipedia-dataset/ του cluster υλοποιείστε τα ακόλουθα **χρησιμοποιώντας το Hadoop**:

(**a**) **Inverted Index** των κειμένων του dataset με την χρήση του Hadoop.

- *Είσοδος*: φάκελος input με πολλαπλά text αρχεία.
- *Έξοδος*: φάκελος output με αρχεία της μορφής (ενδεικτικά):

"word": (docId1, docId2, …)

Σημείωση: Ο φάκελος output θα βρίσκεται στο /user/<username>/output όπου <username> το δικό σας username που έχετε για την σύνδεση στο PDSG Cluster.

- Προβλήματα που θα αντιμετωπίσετε στο συγκεκριμένο βήμα:
  - Tokenization
  - Αφαίρεση xml tags
  - Αφαίρεση σημείων στίξης

(**b**) Φιλτράρισμα ασήμαντων λέξεων (άρθρων, προσδιορισμών κλπ). Μια naive μέθοδος για να αφαιρέσετε τις λέξεις αυτές είναι να επιτρέψετε μόνο λέξεις μεγαλύτερες από N γράμματα (π.χ. N >= 4).

(**c**) (**Προαιρετικό)** Εύρεση σημαντικότερων λέξεων σε ένα κείμενο σύμφωνα με την εξής λογική: Μία σημαντική λέξη εμφανίζεται πολλές φορές μέσα σε ένα κείμενο αλλά δεν εμφανίζεται σε άλλα. Από τα αποτελέσματα του βήματος b μπορείτε να βρείτε ποιές λέξεις αναφέρονται σε λίγα μόνο έγγραφα και από εκεί μπορείτε να εξάγετε συμπεράσματα για το ποιες είναι πιο σημαντικές ειδικά αν χρησιμοποιήσετε εκτός από το inverted index και το wordcount.

Άσκηση 2η:

Χρησιμοποιώντας τα αποτελέσματα της Άσκησης 1 και το Apache Mahout, υλοποιείστε μια εφαρμογή που χρησιμοποιεί τον K-Means Clustering αλγόριθμο του Mahout για να ομαδοποιήσει τα έγγραφα του dataset σε ομάδες ανάλογα με την θεματολογία τους.

Παραδοτέα:

- Ο κώδικας των εργασιών σε μορφή συμπιεσμένου αρχείου (zip, tar.gz, 7zip, rar κλπ.).
- Το εκτελέσιμο jar αρχείο που χρησιμοποιήσατε για να τρέξετε την εργασία

στο cluster.

- Ένα μικρό technical report που να περιγράφει την διαδικασία που ακολουθήσατε για την συγγραφή του κώδικα και την εκτέλεση του στο cluster. Επίσης θα πρέπει να περιλαμβάνει και τους φακέλους στους οποίους βρίσκονται τα αποτελέσματα από την εκτέλεση της εργασίας (δηλ. του φακέλους του HDFS).
- Για κάθε επικοινωνία παρακαλώ το subject του e-mail σας να είναι PDSG-Coursework.

Βιβλιογραφία:

1. Manning - Hadoop In Action 3rd Edition (early access) http://manning.com/lam/
2. Manning - Mahout In Action http://manning.com/owen/

Πηγές:

- Hadoop Tutorial https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
- Mahout https://mahout.apache.org/
- CDH https://www.cloudera.com/downloads/quickstart_vms/5-13.html
- gwtwiki - http://code.google.com/p/gwtwiki/