HADOOP
Project Ακαδημαϊκό Έτος 2017-2018

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

Technical report

Τραχανίδου Ελένη
2620

❖ Εισαγωγικά:

Ο κώδικας υλοποιήθηκε σε γλώσσα python. Αναφέρεται στην Άσκηση 1α) *Inverted Index ,* b) *Φιλτράρισμα ασήμαντων λέξεων*

Ολοκληρώθηκαν πλήρως τα κομμάτια:

- Tokenization
- Αφαίρεση xml tags
- Αφαίρεση σημείων στίξης
- *Φιλτράρισμα ασήμαντων λέξεων*
- Αποθήκευση του ονόματος αρχείου που βρέθηκε η λέξη

❖ Διαδικασία στο cluster:

Η εκτέλεση των αρχείων έγινε με την εντολή:
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/**hadoop-streaming-2.7.5.jar** -file /home/trachanid/hadoop/mapper.py -mapper mapper.py -file /home/trachanid/hadoop/reducer.py -reducer reducer.py -input /user/tmp/wikipedia-dataset/* -output /user/trachanid/inverted_index/wikipedia-dataset-output

❖ Διαδικασία υλοποίησης κώδικα:

Η βασική λογική πίσω από την διαδικασία είναι αυτή της MapReduce.

- **Map stage** : Εδώ γίνεται επεξεργασία των δεδομένων αρχείων γραμμή προς γραμμή. Και αποθηκεύονται στο σύστημα του Hadoop (HDFS).
- **Reduce stage** : Εδώ γίνεται η επεξεργασία των δεδομένων του mapper. Τέλος δημιουργείται ένα set από output,τα οποία αποθηκεύονται στο HDFS.

Το αρχείο mapper.py χωρίζει σε λέξεις τα κείμενα, αφαιρώντας όλα τα σημεία στίξης και τις ετικέτες. Τα αποτελέσματα γράφονται σε ένα dictionary και αυτό με τη σειρά του σε ένα ενδιάμεσο αρχείο.

Το αρχείο reducer.py διαβάζει το dictionary από το ενδιάμεσο αρχείο και αποθηκεύει το περιεχόμενο σε ένα νέο dictionary. Εκεί με επανάληψη των κλειδιών και των αντίστοιχων τιμών τους, υπολογίζεται η συχνότητα εμφάνισης των λέξεων στα αρχεία συνολικά , και εμφανίζονται τα αρχείο στα οποίο βρέθηκε.

❖ Αποτελέσματα:

Τα αποτελέσματα της προσομοίωσης βρίσκονται στο φάκελο:
/user/trachanid/inverted_index/wikipedia-dataset-output/part-00000

★ Επειδή έστειλα δεύτερη και τελική φορά τα αρχεία, τα δεδομένα από το φάκελο wikipedia-dataset είχαν κατέβει. Ωστόσο μπορείτε να δείτε κάποια αποτελέσματα ελέγχου, στο φάκελο:
/user/trachanid/inverted_index/output

Η εκτέλεση των αρχείων έγινε με την εντολή:
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.5.jar -file /home/trachanid/mapper1.py -mapper mapper1.py -file /home/trachanid/reducer1.py -reducer reducer1.py -input /user/trachanid/inverted_index/input/* -output /user/trachanid/inverted_index/output