hy-562 - Report for Assignment 1

Gkigkis Petros - A.M. 948 - gigis@csd.uoc.gr

Exercice 1:

Στο πρώτο MapReduce χρησιμοποιώ παρόμοια λογική με αυτή του παραδείγματος του WordCount.

Στο Map (Job 1):

Αφαιρώ τα tokens που δεν κάνουν match “[A-Za-z0-9]” (regex) και επιπλέον χρησιμοποιώ την toLowerCase συνάρτηση.

Στο Reduce (Job 1):

Εκτελείτε το reduce και επιπλέον αφαιρώ keys (words) που δεν είναι stopwords δηλαδή έχουν εμφανιστεί συνολικά λιγότερες απο 4000 φορές.

Αποθηκεύω τα παραγόμενα αρχεία στο “output/temp”.

Στο Map (Job 2):

Εναλλάσσω το key με το value για να εκμεταλλευτώ το ότι ο reducer ταξινομεί τα keys. Επιπλέον υλοποίησα μια class Comparator για να κάνω Descending sorting.

(Άλλος τρόπος είναι να κάνω * -1 τα keys εδώ και ξανά * -1 στο reducer για να έχω ταξινόμηση σε descending order.)

Στο Reduce (Job 1):

Τυπώνω τα k πιο συχνά words και αποθηκεύω όλα τα key, values στο “output“.

Μέσα στο φάκελο scripts υπάρχει script σε bash που μετατρέπει το output του hadoop στο ζητούμενο stopwords.csv.

Exercice 2a:

1. Ο μέσος χρόνος εκτέλεσης είναι 21,3 seconds.

2. Ο μέσος χρόνος εκτέλεσης είναι 15,1 seconds. Ο χρόνος έχει μειωθεί διότι με την χρήση του combiner το output των mapper ουσιαστικά γίνεται reduce “locally” στο κάθε machine με αποτέλεσμα λιγότερα key-values να στέλνονται πάνω από το δίκτυο καθώς επίσης και ο reducer έχει να κάνει λιγότερα reduce.

3. Ο μέσος χρόνος έχει μειωθεί και είναι στα 14,2 seconds. Βελτιώνεται ο χρόνος λόγω του ότι γίνονται write/read στο δίσκο λιγότερα δεδομένα. Επιπλέον μεταφέρονται λιγότερα (bits!bytes!mbytes!) πάνω από το δίκτυο.

4. Ο μέσος χρόνος αυξήθηκε στα 16, seconds. Παρατηρώ ότι ο χρόνος αυξήθηκε με την αύξηση του αριθμού των reducers. Ο βασικός λόγος που συνέβη αυτό είναι ότι τα δεδομένα πλέον compressed είναι μικρά σε μέγεθος και όταν γίνονται split σε 50 reducers μέχρι το hadoop να δεσμεύσει 50 “nodes” και να στείλει τα δεδομένα, τα απαραίτητα configurations και να περιμένει να εκτελέσουν όλοι οι reducer το task που τους ανατέθηκε παίρνει παραπάνω χρόνο. Ο αριθμός των reducers εξαρτάται από το μέγεθος των δεδομένων.

Exercice 2b:

1. done

2.

i) Υπάρχουν 45248 μοναδικές λέξεις. Αυτό φαίνεται από τον Counter στο MapReduce Framework “Reduce output records”.

ii) Βρήκα 15111 λέξεις που εμφανίζονται μόνο μία φορά. Για να βρω αυτές τις λέξεις μέτρησα στους reduces τα keys που έχουν μόνο ένα value. (Δηλαδή έχουν εμφανιστεί μια φορά)

iii) Τα words που βρίσκονται μονό σε ένα αρχείο είναι 31618.

Exercice 3:

Δημιούργησα έναν custom type για να επιστρέφω ως value ο οποίος είναι ένα ακόμα key-value (Text, IntWritable). Σε αυτόν τον custom type key = όνομα αρχείου και value το πόσες φορές συναντήσαμε το word.

Υλοποίησα και μία ξεχωριστή κλάση combiner όπου στον νέο τύπο που έφτιαξα προσθέτει τα number of occurrencies σε ένα συγκεκριμένο αρχείο.