Αναφορά Εργασίας 2: Επεξεργασία Μεγάλων Δεδομένων με Hadoop MapReduce

Κωνσταντίνα Μαρίνα Μπλέτσα, ΑΕΜ 243

Υποεργασία 1 - Numeronyms

Ως numeronym ορίζεται το αλφαριθμητικό που σχηματίζεται από τον πρώτο και τον τελευταίο χαρακτήρα μιας λέξης, με το πλήθος των ενδιάμεσων χαρακτήρων να παρεμβάλλεται ανάμεσά τους. Η λογική του Mapper, όπως υλοποιήθηκε στην κλάση NumeronymMapper, περιλαμβάνει τον διαχωρισμό των λέξεων και τον καθαρισμό τους, ώστε να κρατηθούν μόνο οι αλφαβητικοί χαρακτήρες και να μετατραπούν σε πεζά, καθιστώντας την εφαρμογή case-insensitive. Επίσης, εφαρμόστηκε φίλτρο ώστε να αγνοούνται οι λέξεις με μήκος μικρότερο των τριών χαρακτήρων, καθώς και τα σημεία στίξης. Στο στάδιο του Reducer, αθροίστηκαν οι εμφανίσεις κάθε numeronym και το τελικό αποτέλεσμα φιλτραρίστηκε βάσει της παραμέτρου k, η οποία ορίζει το ελάχιστο πλήθος εμφανίσεων που απαιτείται για να συμπεριληφθεί ένα numeronym στην έξοδο. Τα αποτελέσματα της εκτέλεσης για την παράμετρο k=10 καταγράφηκαν στο αρχείο numeronym_results.txt. Η ανάλυση των δεδομένων εξόδου επιβεβαιώνει την ορθή λειτουργία του φίλτρου, καθώς όλες οι καταγραφές έχουν συχνότητα μεγαλύτερη ή ίση του 10. Το πιο συχνό numeronym που εντοπίστηκε είναι το "t1e" με 79.438 εμφανίσεις, το οποίο πιθανότατα αντιστοιχεί στο άρθρο "the" . Ακολουθούν τα "h1s" με 11.727 εμφανίσεις και "w1s" με 11.367 εμφανίσεις, τα οποία παραπέμπουν σε συχνές λέξεις όπως "his" και "was" αντίστοιχα.

Υποεργασία 2 - DNA

Η δεύτερη υποεργασία αφορούσε την ανάλυση βιολογικών δεδομένων και συγκεκριμένα μιας ακολουθίας DNA από το βακτήριο E. coli, η οποία αποτελείται από τα σύμβολα A, G, C και T. Στόχος ήταν η καταμέτρηση των συνεχόμενων 2-άδων, 3-άδων και 4-άδων βάσεων για κάθε γραμμή του αρχείου εισόδου ανεξάρτητα. Η υλοποίηση βασίστηκε στην κλάση DnaMapper, η οποία για κάθε γραμμή εκτελεί επαναληπτικούς ελέγχους για μήκη συμβολοσειρών n ίσον με 2, 3 και 4. Χρησιμοποιώντας τη μέθοδο του sliding window, ο αλγόριθμος εξάγει όλες τις πιθανές υπο-συμβολοσειρές εντός των ορίων της γραμμής, διασφαλίζοντας ότι δεν εξετάζονται ακολουθίες που εκτείνονται σε διαφορετικές γραμμές, όπως απαιτούσε η εκφώνηση. Στη συνέχεια, ο Reducer άθροισε τις εμφανίσεις για κάθε μοναδικό N-gram που εντοπίστηκε. Τα αποτελέσματα που προέκυψαν στο αρχείο dna_results.txt αναδεικνύουν τη συχνότητα εμφάνισης των διαφόρων μοτίβων DNA. Παρατηρείται ότι οι 2-άδες εμφανίζουν τις υψηλότερες συχνότητες λόγω του περιορισμένου αριθμού συνδυασμών, με το ζεύγος "GC" να κυριαρχεί με 378.417 εμφανίσεις, ακολουθούμενο από το "CG" με 341.676 εμφανίσεις . Στις 3-άδες, το μοτίβο "CGC" ξεχωρίζει με 112.398 εμφανίσεις, ενώ στις 4-άδες οι συχνότητες μειώνονται λόγω της αυξημένης πολυπλοκότητας. Για παράδειγμα, η 4-άδα "CAGC" εμφανίζεται 35.901 φορές, ενώ η "CTAG" είναι πολύ σπάνια μόνο 845 εμφανίσεις.