Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Προχωρημένα θέματα βάσεων δεδομένων.

Εξαμηνιαία εργασία, θέμα 3ο: Machine Learning - Ομαδοποίηση δεδομένων με εκτέλεση του k-means αλγόριθμου.

Ονοματεπώνυμο: Παπακωνσταντίνου Πολύβιος
Αριθμός μητρώου: 03114892

Ημερομηνία: 10/03/2020

Εισαγωγή.

Στο συγκεκριμένο θέμα θα βασιστούμε στον αλγόριθμο k-means για MapReduce στο dataset που μας δίνεται από το cslab, προκειμένου να βρούμε τις συντεταγμένες των top 5 περιοχών επιβίβασης επιβατών ταξί.

Για το σκοπό αυτό έχουμε «στήσει» τοπικά στο μηχάνημα μας 2 virtual machines, σύμφωνα με τις οδηγίες του εργαστηρίου, τα οποία θα παίξουν το ρόλο master και slave αντίστοιχα. Θα χρησιμοποιήσουμε σύστημα HDFS για να αποθηκεύσουμε τα input και output files του κώδικα μας, καθώς και το Spark API για τον αλγόριθμο MapReduce που θα υλοποιήσουμε. Για τον κώδικα του προγράμματος μας επιλέξαμε τη γλώσσα Python.

Περίληψη κώδικα.

Αρχικά, εισάγουμε το dataset που κατεβάσαμε από το cslab και κρατάμε από τα δεδομένα μόνο τις $4^η$ και $5^η$ στήλες, καθώς για το συγκεκρίμενο θέμα μας ενδιαφέρουν μόνο τα σημεία επιβίβασης. Κάνουμε επίσης και έναν ακόμη έλεγχο προκειμένου να παραλείψουμε σημεία με συντεταγμένες [0.0, 0.0], καθώς πρόκειται για missing values που θα αλλοιώσουν τα αποτελέσματα του κώδικα μας.

Στη συνέχεια, σύμφωνα με τον αλγόριθμο k-means, τρέχουμε επαναληπτικά για 3 φορές: για κάθε σημείο [x, y] υπολογίζουμε την απόσταση Haversine από τα 5 κέντρα και κρατάμε index για το ποιο κέντρο είναι πιο κοντά στο εν λόγω σημείο. Προκειμένου τώρα να ανανεώσουμε το κάθε κέντρο, υπολογίζουμε τον μέσο όρο των κοντινότερων σημείων του. Όταν υπολογίσουμε όλα τα νέα κέντρα, επαναλαμβάνουμε τη διαδικασία.

Παραθέτουμε τα τελικά αποτελέσματα:

-73.83747289194628	40.716328670235036
-73.96851068547537	40.77206591259295
-73.99479622767157	40.713172559751065
-74.00242408772402	40.73170488692425
-73.98812567171869	40.74591444317248

Τα αρχεία που φτιάξαμε/δημιουργήθηκαν για την εκτέλεση του MapReduce, καθώς και ο ψευδοκώδικας, περιέχονται στα zip files όπως ζητήθηκε.
Σημειώνεται πως παραλήφθεικε το link για το hdfs site, αφού η διαδικασία έγινε τοπικά χωρίς public ip.