ΠΡΟΧΩΡΗΜΕΝΑ ΘΕΜΑΤΑ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ
Εξαμηνιαία Εργασία Εισαγωγή στο MapReduce

Παρακάτω ακολουθεί σύντομη περιγραφή της μεθοδολογίας που εφαρμόστηκε σε κάθε άσκηση.

1) α) Εφόσον ενδιαφερόμαστε μόνο για την ώρα έναρξης της διαδρομής και για την διάρκεια αυτής , από την είσοδο κρατάμε την ώρα έναρξης και βρίσκουμε την διάρκεια σε λεπτά (end_time - start_time). Έπειτα αθροίζουμε τις διάρκειες με βάση την ώρα έναρξης και κρατάμε το πλήθος αυτών.Τέλος βρίσκουμε τον μέσο όρο ανά ώρα έναρξης διαδρομης .

β) Σε αυτήν την περίπτωση χρειαζόμαστε τα vendor_id, route_id και cost καθε route . Κάνουμε join τους δύο πίνακες στο route_id και έτσι έχουμε για κάθε διαδρομή το κόστος της και το vendor_id της . Κρατάμε τα 2 τελευταία και με κλειδί το vendor_id κάνουμε reduce τα κόστη κρατώντας το max για κάθε vendor_id .

2) Κρατάμε τα coords του σημείου έναρξης κάθε διαδρομής . Έπειτα θέτουμε ως αρχικά centroids τα 5 πρώτα σημεία και τέλος εφαρμόσουμε τον αλγόριθμο K-Means θεωρώντας ότι συγκλίνει με 3 επαναλήψεις . Ο αλγόριθμος με Machine Learning σε κάθε επανάληψει εντοπίζει (βελτιωμένα) τα centroids .

3) Για να βρούμε το rank κάθε κόμβου-σελίδας εργαζόμαστε ως εξής : 1.Από το input κρατάμε τον κόμβο και το outbound link , και τα ομαδοποιούμε με βάση τον κόμβο. 2.Δημιουργούμε έναν πίνακα ranks που περιέχει το rank κάθε κόμβου (αρχικοποιημένο στο 0,5). 3.Έπειτα για κάθε κόμβο , βρίσκουμε την συνεισφορά του σε κάθε κόμβο που συνδέεται με ένα από τα outbound links του . 4.Έτσι , έχουμε ζεύγη (κόμβος, συνεισφορά από έναν κόμβο που δείχνει σε αυτόν) των οποίων αθροίζουμε τις τιμές με βάση τον κόμβο . Έχοντας κατασκευάσει τον όρο του αθροισματος προσθέτουμε την σταθερα 1 - d / N και βρίσκουμε την νέα τιμή του rank κάθε κόμβου. Επαναλαμβάνουμε τα βήματα 3 και 4 .

4) Αποθηκεύουμε τα δεδομένα των πινάκων και τα οργανώνουμε σε (key,value) ζεύγη με κλειδί για τα στοιχεία του πρώτου πίνακα την στήλη κάθε στοιχείου και value την tuple (γραμμή, τιμή) ενώ για τον B έχουμε key την γραμμή και value (στήλη, τιμή) . Στη συνέχεια κάνουμε join τα στοιχεία και ως key χρησιμοποιείται η στήλη για τα στοιχεία του A και η γραμμή για τα στοιχεία του B . Τα ζεύγη που δημιουργουνται πολλαπλασιαζονται μεταξυ τους και σαν κλειδί έχουν την θέση στον πίνακα-γινόμενο (γραμμή = γραμμή του στοιχείου του α , στήλη = στήλη του στοιχείου του β).Τέλος, αθροίζουμε τα στοιχεία με το ίδιο κλειδί και κρατάμε την θέση τους . Οπότε έχουμε για κάθε θέση του πίνακα-γινόμενο την τιμη της .