Προχωρημένα Θέματα Βάσεων Δεδομένων

Ομαδοποίηση μεγάλου όγκου δεδομένων με εκτέλεση του αλγορίθμου μηχανικής μάθησης k-means με χρήση του προγραμματιστικού μοντέλου map reduce

---

Στοιχεία

Όνομα:	Γιαννακούλιας Γεώργιος
Αριθμός Μητρώου:	03115044
Ημερομηνία:	9 Μαρτίου 2020
Σχολή:	Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (ΕΜΠ)
Μάθημα:	Προχωρημένα Θέματα Βάσεων Δεδομένων (Ροή Λ)
Περιγραφή:	Το κείμενο αφορά την εξαμηνιαία εργασία του μαθήματος “Προχωρημένα Θέματα Βάσεων Δεδομένων”. Επέλεξα να υλοποιήσω ατομικά το 3ο Θέμα.

Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήσαμε είναι πραγματικά και αφορούν σε διαδρομές taxi στην Νέα Υόρκη. Οι δοθείσες διαδρομές των taxi έγιναν από τον Ιανουάριο εώς το Ιούνιο του 2015 και υπάρχουν διαθέσιμες online. Λόγω των περιορισμένων πόρων που η κάθε ομάδα έχει στη διάθεσή της, θα επεξεργαστούμε μόνο ένα υποσύνολο μεγέθους 2 GB. Τα δεδομένα αυτά περιέχουν 13 εκατομμύρια διαδρομές, που πραγματοποιήθηκαν το Μάρτιο του 2015 και μπορείτε να τα κατεβάσετε από εδώ. Στο συμπιεσμένο αρχείο που σας δίνουμε, περιλαμβάνονται δύο comma-delimited αρχεία κειμένου (.csv) που ονομάζονται: yellow_tripdata_1m.csv και yellow_tripvendors_1m.csv.
Το πρώτο αρχείο περιλαμβάνει όλη την απαραίτητη πληροφορία για μια διαδρομή. Το αρχείο των TripData έχει την εξής μορφή:

```
369367789289,2015-03-27 18:29:39,2015-03-27
19:08:28,-73.975051879882813,40.760562896728516,-73.847900390625,40.7326850 89111328,34.8
369367789290,2015-03-27 18:29:40,2015-03-27
18:38:35,-73.988876342773438,40.77423095703125,-73.985160827636719,40.76343 9178466797,11.16
```

Το πρώτο πεδίο αποτελεί το μοναδικό id μιας διαδρομής. Το δεύτερο (τρίτο) πεδίο την ημερομηνία και ώρα έναρξης (λήξης) της διαδρομής. Το τέταρτο και πέμπτο πεδίο το γεωγραφικό μήκος και πλάτος του σημείου επιβίβασης, ενώ το έκτο και έβδομο πεδίο περιλαμβάνουν το γεωγραφικό μήκος και πλάτος του σημείου αποβίβασης. Τέλος, το όγδοο πεδίο δείχνει το συνολικό κόστος της διαδρομής.
Το δεύτερο αρχείο που σας δίνεται περιέχει πληροφορία για τις εταιρίες taxi. Η μορφή του φαίνεται στο παρακάτω παράδειγμα:

```
369367789289,1
369367789290,2
```

Το πρώτο πεδίο αποτελεί το μοναδικό id μιας διαδρομής και το δεύτερο πεδίο το μοναδικό αναγνωριστικό μιας εταιρείας taxi (vendor).

Περιγραφή Εργασίας

Χρησιμοποιώντας τα δεδομένα που περιγράψαμε, στο τρίτο θέμα θέλουμε να βρούμε τις κεντρικές συντεταγμένες των top 5 περιοχών επιβίβασης πελατών. Για το σκοπό αυτό ζητείται να υλοποιηθεί ο αλγόριθμος ομαδοποίησης k-means (wiki), τον οποίο θα χρησιμοποιήσουμε για να ομαδοποιήσουμε τα σημεία επιβίβασης σε k=5 περιοχές (clusters) και να βρούμε το κέντρο των σημείων κάθε περιοχής.
Ο αλγόριθμος είναι επαναληπτικός και θεωρώντας 5 αρχικά κέντρα με προκαθορισμένες συντεταγμένες, γίνονται δύο βήματα σε κάθε επανάληψη:

1. Ανάθεση σημείων σε περιοχή: βρίσκουμε και αντιστοιχούμε κάθε σημείο επιβίβασης στο κοντινότερο από τα 5 κέντρα.
2. Ανανέωση κέντρων περιοχών: υπολογίζουμε τις συντεταγμένες των 5 νέων κέντρων ως τον μέσο όρο των συντεταγμένων των σημείων που ανατέθηκαν σε κάθε περιοχή.

Ο αλγόριθμος συγκλίνει όταν τα κέντρα δεν αλλάζουν πλέον από επανάληψη σε επανάληψη.
Φορτώστε αρχικά τα csv αρχεία που σας δίνονται στο HDFS. Για την υλοποίησή σας θεωρείστε ως αρχικά κέντρα τις συντεταγμένες επιβίβασης από τις 5 πρώτες γραμμές των δεδομένων μας και ότι χρειάζονται 3 μόνο επαναλήψεις του αλγορίθμου ώστε να επιτευχθεί σύγκλιση. Αποθηκεύσετε σε ένα αρχείο στο HDFS το αποτέλεσμα.

Σημείωση: Υπολογισμός απόστασης (Haversine).
Αν φ είναι το γεωγραφικό πλάτος και λ το 5 γεωγραφικό μήκος, τότε η απόσταση δύο σημείων δίνεται από τους τύπους:
$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2)$
$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$
$d = R \cdot c$, όπου R είναι η ακτίνα της Γης (6371m)

Μεθοδολογία

Σε πρώτη φάση στήσαμε το hdfs (distributed file system) και εγκαταστήσαμε το spark (general-purpose cluster-computing framework) στον υπολογιστή του okeanos που μας δόθηκε στα πλαίσια του μαθήματος, σύμφωνα με τις δοσμένες οδηγίες.
Ο αλγόριθμος k-means λειτουργεί εκτελώντας επαναληπτικά φάση ανάθεσης σημείων σε περιοχή και έπειτα την φάση επανυπολογισμού των συντεταγμένων των κέντρων, όπως περιγράφηκαν προηγουμένως.
Στην δική μας υλοποίηση θεωρούμε 5 συστάδες και σύγκλιση μετά την 3 επαναλήψεις.
Τα δεδομένα που έχουμε στην διάθεση μας περιέχουν εσφαλμένες καταχωρήσεις, συγκεκριμένα διαδρομές που ξεκινούν ή που τελειώνουν σε συντεταγμένες (0.0, 0.0). Πριν την έναρξη του αλγορίθμου φιλτράρουμε λοιπον τα δεδομένα μας.
Στην συνέχεια αρχικοποιούμε τις συντεταγμένες των κέντρων σύμφωνα με τις θέσεις επιβίβασης των πρώτων 5 διαδρομών.
Για τον υπολογισμό χρησιμοποιούμε το προγραμματιστικό μοντέλο map reduce.
Στην φάση map, για κάθε διαδρομή υπολογίζουμε το κοντινότερο κέντρο (κάθε κέντρο έχει ξεχωριστό id) και κάνουμε emit το id του κέντρου μαζί με τις συντεταγμένες του σημείου επιβίβασης.
Στην φάση reduce, κάνουμε reduce σύμφωνα με το κλειδί, συνεπώς για κάθε διαφορετικό κέντρο (με αντίστοιχο id) διαθέτουμε τις συντεταγμένες των αντίστοιχων σημείων επιβίβασης. Συνεπώς, υπολογίζουμε τον μέσο όρο τους και ανανεώνουμε τις συντεταγμένες του κέντρου.
Μετά την 3 επανάληψη της παραπάνω διαδικασίας έχουμε υπολογίσει τις ανανεωμένες τιμές των κέντρων επιβίβασης.
Για τον υπολογισμό έχουμε δημιουργήσει συναρτήσεις που

Ψευδοκώδικας

Συναρτήσεις:

- haversine: παίρνει ως όρισμα 2 ζεύγη συντεταγμένων (x,y) και υπολογίζει την απόσταση τους με χρήση της haversine formula
- filter: παίρνει ως όρισμα μια διαδρομή και υπολογίζει εάν είναι έγκυρη (δηλαδή δεν περιέχει 0.0 συντεταγμένες επιβίβασης ή αποβίβασης)
- closest: παίρνει ως όρισμα ένα σημείο και την λίστα των κέντρων (αναγνωριστικό και συντεταγμένες) και επιστρέφει το αναγνωριστικό του κοντινότερου κέντρου

Ψευδοκώδικας υλοποίησης του γενικού προγράμματος:

```
rides = read("yellow_tripdata_1m.csv")
filter(rides)
centers = head(rides, 5)
```

```
Repeat 3 times:
    mapped_rides = rides.map(closest(ride, centers), ride)
    centers = mean(mapped_rides.reduce())

print(centers)
```

Ψευδοκώδικας υλοποίησης του map:

```
map(value):
// value: (x,y) coordinates of start of ride
id = closest((x,y), centers)
emit(id, (x,y))
```

Ψευδοκώδικας υλοποίησης του reduce:

```
reduce(key, value):
// key: id of center
// value: [(x1,y1), (x2,y2), ...]
(sumX, sumY) = sum(value)
times = len(value)
result = ( sumX / times, sumY / times )
emit((x, result))
```

Αποτελέσματα

Μετά την εκτέλεση του προγράμματος προκύπτουν οι εξής συντεταγμένες για τα κέντρα των 5 συστάδων:

```
center with id 0 has coordinates (-74.00230011766776, 40.73169985176288)
center with id 1 has coordinates (-73.83759587792197, 40.71642357510716)
center with id 2 has coordinates (-73.99480097463035, 40.713152178744146)
center with id 3 has coordinates (-73.98811549038231, 40.74593646854986)
center with id 4 has coordinates (-73.9685119851316, 40.772067643696445)
```

Μπορεί κανείς να έχει πρόσβαση στα αρχεία του hdfs με τα παρακάτω λινκ:

- http://83.212.76.238:50070/explorer.html#/

όπου 83.212.76.238 η public ipv4 του υπολογιστή master.