Obrada skupa podataka GeoLife Microsoft

Prvi projekat iz predmeta sistemi za obradu i analizu velike količine podataka

Ana Stojanović 1087

Sadržaj

Tema

Izabrati jedan od navedenih velikih mobilnih (prostorno-vremenskih) i IoT skupova podataka koji se odnose na kretanje ljudi i vozila u pametnim gradovima (Smart Cities), download-ovati, prečistiti, eventualno re-formatirati i postaviti podatke na HDFS.

Skup podataka: GeoLife - https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide

Razviti samostalnu aplikaciju korišćenjem Apache Spark framework-a koja uključuje sledeće funkcionalnosti:

1. Odrediti broj odgovarajućih pojava atributa/događaja na određenoj lokaciji u datom vremenu, koji zadovoljavaju određeni uslov.
2. Naći minimalne, maksimalne, srednje vrednosti (standardne devijacije) određene(-ih) atributa na zadatoj lokaciji i vremenu

Izvršiti evaluaciju i analizu performansi ove aplikacije u lokalnom/pseudo distribuiranom modu na klasteru Spark docker container-a.

Osnovni detalji implementacije projekta

Projekat je rađen u Python programskom jeziku i korišćen je PySpark.

Urađene su analize:

1. Nalaženje svih korišćenih prevoznih stredstava u celom skupu podataka
2. Analiza prevoza korisnika - Najduži i najkraći pojedinačni put koji je korisnik prešao po prevoznom sredtsvu, prosečan pređeni put, broj korišćenja određenog prevoznog sredstva
3. Dnevno korišćenje prevoznih stredstava jednog korisnika - Minimalna, maksimalna i prosečna dnevna upotreba prevoznih sredstava

Detaljniji opis svake analize se nalazi u sekciji "Analize" ovog dokumenta.

Aplikacija je testirana u lokalnom modu, spark standalone modu i na klasteru docker kontejnera. Više informacija o pokretanju aplikacije se nalazi u posebnom dokumentu.

Opis skupa podataka GeoLife

GeoLife je skup podataka koje su prikupljali naučnici iz Microsoft Research Asia centra. Sadrži trajektorije koje je u određenom vremenskom periodu skupljalo 182 korisnika. Svaka trajektorija sadrži sekvencu GPS koordinata (latituda, longituda, visina) kao i datum i vreme. Većina tačaka je semplovana na svakih 5-10 sekundi, sa distancom od 5-10 metara.

Pored trajektorija, korisnici su vodili evidenciju o korišćenom prevoznom sredstvu.

Ovaj skup podataka se sastoji od raznovrsnih putanja, koje uključuju različite načine kretanja i različita prevozna sredstva. Većina podataka je kreirana u Pekingu u Kini. Odredjeni korisnici su skupljali podatke samo par nedelja, dok su drugi skupljali podatke par godina.

Format GeoLife skupa podataka

Data

```
├── 118
│   ├── labels.txt
│   └── Trajectory
│       ├── 20070512171956.plt
│       ├── 20070519065023.plt
```

Skup podataka čine direktorijumi, gde svaki direktorijum predstavlja jednog korisnika označenog celobrojnim identifikatorom.

Svaki korisnik ima listu trajektorija koja se nalaze u folderu Trajectory. Trajektorije su u formatu .plt, i obično obuhvataju vremenski period od jednog do dva dana. Ime trajektorije označava vreme prve tačke u toj trajektoriji.

Opciono, za svakog korisnika postoji tekstualni fajl sa imenom labels.txt, koji sadrži način kretanja korisnika. Jedan labels.txt fajl pokriva celokupno kretanje jednog korisnika.

Izgled plt fajla

Trajektorija je u formatu sličnom csv formatu. Linije 1-6 se mogu odbaciti.

Linije sa smislenim podacima sadrže sledeća polja:

- Latituda u decimalnom formatu
- Longituda u decimalnom formatu
- Odbaciti polje, uvek 0
- Visina

- Datum s decimalnim delom, koji predstavlja broj sekundi koji je prošao od 12/30/1899
- Datum kao string
- Vreme kao string

Svi datumi su u GMT vremenskoj zoni.

```
Geolife trajectory
WGS 84
Altitude is in Feet
Reserved 3
0,2,255,My Track,0,0,2,8421376
0
39.984702,116.318417,0,492,39744.1201851852,2008-10-23,02:53:04
39.984683,116.31845,0,492,39744.1202546296,2008-10-23,02:53:10
39.984686,116.318417,0,492,39744.1203125,2008-10-23,02:53:15
39.984688,116.318385,0,492,39744.1203703704,2008-10-23,02:53:20
39.984655,116.318263,0,492,39744.1204282407,2008-10-23,02:53:25
39.984611,116.318026,0,493,39744.1204861111,2008-10-23,02:53:30
39.984608,116.317761,0,493,39744.1205439815,2008-10-23,02:53:35
39.984563,116.317517,0,496,39744.1206018519,2008-10-23,02:53:40
39.984539,116.317294,0,500,39744.1206597222,2008-10-23,02:53:45
39.984606,116.317065,0,505,39744.1207175926,2008-10-23,02:53:50
39.984568,116.316911,0,510,39744.120775463,2008-10-23,02:53:55
39.984586,116.316716,0,515,39744.1208333333,2008-10-23,02:54:00
39.984561,116.316527,0,520,39744.1208912037,2008-10-23,02:54:05
```

Izgled labels.txt fajla

labels.txt se sastoji od vremena i labele, gde su sva polja razdvojena "tab" ili "whitespace" karakterom.

Tekstualni fajl koji sarži sledeća polja:

- Početni datum
- Početno vreme
- Krajnji datum
- Krajnje vreme
- Labela

Svi datumi su u GMT vremenskoj zoni.

```
Start Time      End Time        Transportation Mode
2007/06/26 11:32:29     2007/06/26 11:40:29     bus
2008/03/28 14:52:54     2008/03/28 15:59:59     train
2008/03/28 16:00:00     2008/03/28 22:02:00     train
2008/03/29 01:27:50     2008/03/29 15:59:59     train
2008/03/29 16:00:00     2008/03/30 15:59:59     train
2008/03/30 16:00:00     2008/03/31 03:13:11     train
2008/03/31 04:17:59     2008/03/31 15:31:06     train
2008/03/31 16:00:08     2008/03/31 16:09:01     taxi
2008/03/31 17:26:04     2008/04/01 00:35:26     train
2008/04/01 00:48:32     2008/04/01 00:59:23     taxi
2008/04/01 01:00:22     2008/04/01 01:08:13     walk
2008/04/01 03:46:35     2008/04/01 03:54:28     taxi
2008/04/01 04:15:38     2008/04/01 11:06:16     train
```

Učitavanje trajektorija

Za čitanje .plt fajlova je korišćen dataframe API. Budući da je .plt fajl u formatu redova radvojenih zarezom, može se pročitati kao csv fajl.

Svi .plt fajlovi iz Trajectories foldera se učitavaju i pretvaraju u format:

```
| -- user_id: string
| -- trajectory_id: string
| -- date: date
| -- datetime: timestamp
| -- latitude: double
| -- longitude: double
```

User id i trajectory id se dobijaju izvlačenjem identifikatora iz putanje svakog fajla korišćenjem regex-a. Datum, vreme i koordinate se konvertuju u odgovarajući tip podataka i čitaju iz .plt fajla. Filtriraju se podaci koji imaju nevalidne kordinate, datum ili korisnički id.

Učitavanje labela

Korišćen je DataFrame api, i labels.txt fajl je učitan kao csv fajl koji ima tab karakter između kolona. Dataset se pretvara u format:

```
|-- l_user_id: string
|-- label_id: long
|-- start: timestamp
|-- end: timestamp
|-- label: string
```

Label id je generisani podatak koji služi da razlikuje dve uzastopne labele istog tipa.
Izbacuju se redovi sa nevalidnim podacima.

Spajanje labela i trajektorija

Trajektorije i labele se spajaju u jedan dataframe, korišćenjem dataframe funkcije join.

Join se radi na osnovu korisničkog id-ja i vremena.

```
df = df.join(dfl, [df.user_id == dfl.l_user_id, df.datetime.between(dfl.start, dfl.end)], "inner")
```

Konačni format skupa podataka nad kojim se rade analize

```
| -- user_id: string
| -- trajectory_id: string
| -- date: date
| -- datetime: timestamp
| -- latitude: double
| -- longitude: double
| -- label_id: long
| -- start: timestamp
| -- end: timestamp
| -- label: string
```

Arhitektura projekta

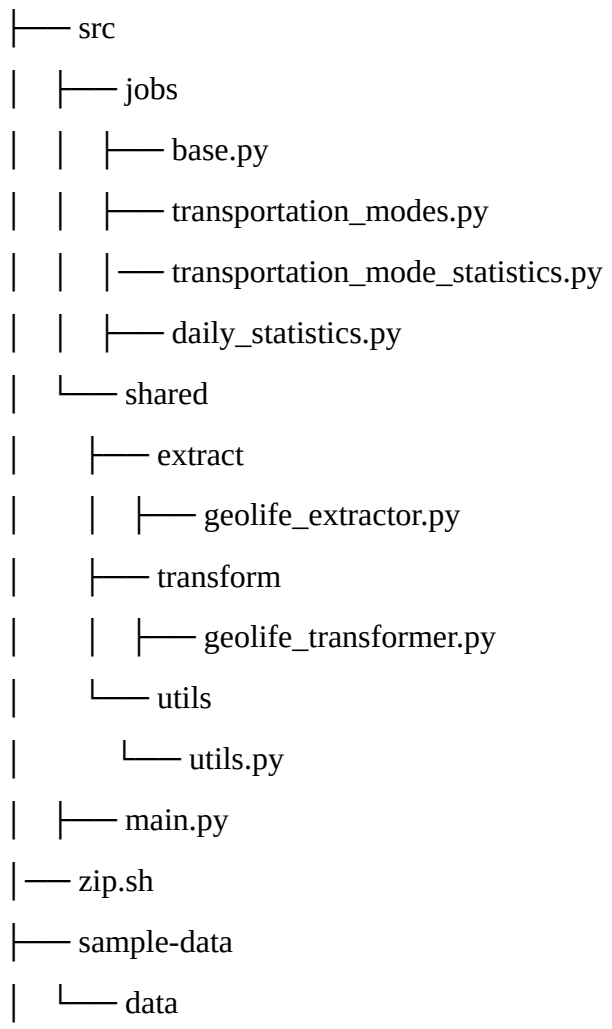
Izvorni kod projekta se sastoji od python paketa jobs i shared i izvršnog programa main.py.

Paket jobs sadrži sve urađene analize, gde se jedna analiza nalazi u jednom fajlu.

Paket shared sadrži kod koji je zajednički za više analiza i to:

- extract - učitavanje geolife skupa podataka
- transform - transformacije koje su zajedničke za više analiza
- utils - utility metode

Takođe postoji i folder sample-data, koji sadrži mali podskup dataset-a, i skripta zip.sh, koja služi za pakovanje svih fajlova u zip format.



Analize

Nalaženje korišćenih prevoznih sredstava za sve korisnike

Ova analiza se može naći u paketu jobs -> transportation_modes

Analiza učitava sve labele iz skupa podataka i nalazi jedinstven skup.

Korišćene su pyspark funkcije select i distinct.

```
Reading dataset
+---------+--------+-------------------+-------------------+-----+
|l_user_id|label_id|              start|                end|label|
+---------+--------+-------------------+-------------------+-----+
|      068|       0|2008-09-14 13:03:08|2008-09-14 13:13:13| bike|
|      068|       1|2008-09-15 02:37:31|2008-09-15 06:08:51| bike|
|      068|       2|2008-09-15 10:17:05|2008-09-15 10:29:55| bike|
|      068|       3|2008-09-15 12:30:30|2008-09-15 12:50:38| bike|
|      068|       4|2008-09-16 03:13:38|2008-09-16 03:17:03| bike|
+---------+--------+-------------------+-------------------+-----+
only showing top 5 rows

+--------+
|   label|
+--------+
|airplane|
|    bike|
|     bus|
|     car|
|  subway|
+--------+
only showing top 5 rows
```

Statistika korišćenja prevoznih sredstava

Ova analiza se može naći u paketu jobs -> transportation_mode_statistics

Ova analiza izračunava statistiku vezanu za prevozna sredstva koja je korisnik koristio. Analiza sadrži sledeće statistike:

- Koliko puta je korisnik koristio određeno prevozno sredstvo
- Najkraća, najduža i prosečna razdaljina koji je korisnik prešao određenim prevozom
- Najkraće, najduže i prosečno vreme korišćenja prevoznog sredstva

Analiza smatra da jedan red u fajlu labels.txt predstavlja jednan kontinualni put tim prevoznim sredsvom.

Prvi korak je učitavanje skupa podataka, i vršenje join-a nad trajektorijama i labelama.

```
+-------+--------------+----------+-------------------+---------+----------+----------+-------------------+-------------------+-----+
|user_id| trajectory_id|      date|           datetime| latitude| longitude|  label_id|              start|                end|label|
+-------+--------------+----------+-------------------+---------+----------+----------+-------------------+-------------------+-----+
|    010|20080405160011|2008-04-05|2008-04-05 16:00:11|34.532742|106.241397|8589934623|2008-04-05 16:00:00|2008-04-06 01:20:15|train|
|    010|20080405160011|2008-04-05|2008-04-05 16:00:25|34.532132|106.243835|8589934623|2008-04-05 16:00:00|2008-04-06 01:20:15|train|
|    010|20080405160011|2008-04-05|2008-04-05 16:00:40|34.531473| 106.24638|8589934623|2008-04-05 16:00:00|2008-04-06 01:20:15|train|
|    010|20080405160011|2008-04-05|2008-04-05 16:00:55|34.530827|106.248903|8589934623|2008-04-05 16:00:00|2008-04-06 01:20:15|train|
|    010|20080405160011|2008-04-05|2008-04-05 16:01:10|34.530167|106.251432|8589934623|2008-04-05 16:00:00|2008-04-06 01:20:15|train|
+-------+--------------+----------+-------------------+---------+----------+----------+-------------------+-------------------+-----+
```

Drugi korak podrazumeva računanje razdaljine između svake dve sukcesivne tačke u jednom pređenom putu. Važno je da se razdaljina računa samo između tačaka koje pripadaju jednom kontinualnom očitavanju koordinata (nema smisla računati razdaljinu za dve tačke između kojih nedostaju podaci).

Parcijalne razdaljine se računaju korišćenjem spark window funkcija. Jedna particija prozora odgovara jednoj očitanoj labeli iz fajla labels.txt.

```
partition = [df.user_id, df.label, df.label_id]
window_spec = Window.partitionBy(partition).orderBy(df.datetime)
df = df.select(df.user_id, df.label, df.label_id, df.latitude, df.longitude, df.datetime, #and other...
            lead(df.latitude).over(window_spec).alias("next_latitude"),
            lead(df.longitude).over(window_spec).alias("next_longitude"),
            lead(df.datetime).over(window_spec).alias("next_datetime"))
```

Rezultat je red koji sadrži trenutnu koordinatu i vreme, kao i narednu koordinatu i vreme. Svaki poslednji red u particiji će imati null vrednosti za narednu koordinatu. Potrebno je isfiltrirati poslednje redove iz skupa pre prelaženja na sledeći korak. Takođe se mogu ukloniti redovi u kojima je trenutna i naredna koordinata identična. Za svaki red se sada može izračunati parcijalna distanca, računanjem distance između trenutne i naredne koordinate.

```
Calculating partial sums
+-------+-----+----------+---------+----------+----------+-------------------+-------------+-------------+-------------------+--------------------+--------------------+
|user_id|label|  label_id| latitude| longitude|      date|           datetime|next_latitude|next_longitude|      next_datetime|        dist_part_km|         time_part_h|
+-------+-----+----------+---------+----------+----------+-------------------+-------------+-------------+-------------------+--------------------+--------------------+
|    010|  bus|8589934627|39.895505|116.316188|2008-04-06|2008-04-06 09:30:20|    39.895585|    116.317153|2008-04-06 09:30:35| 0.08280382731118512|0.004166666666666667|
|    010|  bus|8589934627|39.895585|116.317153|2008-04-06|2008-04-06 09:30:35|     39.89559|    116.318235|2008-04-06 09:30:50| 0.09230748001453351|0.004166666666666667|
|    010|  bus|8589934627| 39.89559|116.318235|2008-04-06|2008-04-06 09:30:50|    39.895597|    116.319602|2008-04-06 09:31:04| 0.11662182153602813|0.003888888888888...|
|    010|  bus|8589934627|39.895597|116.319602|2008-04-06|2008-04-06 09:31:04|    39.895572|     116.32116|2008-04-06 09:31:19| 0.13294260645972888|0.004166666666666667|
|    010|  bus|8589934627|39.895572| 116.32116|2008-04-06|2008-04-06 09:31:19|    39.895545|    116.321745|2008-04-06 09:31:34|0.049996933563819575|0.004166666666666667|
+-------+-----+----------+---------+----------+----------+-------------------+-------------+-------------+-------------------+--------------------+--------------------+
only showing top 5 rows
```

Korišćenjem spark groupby i sum funkcije, mogu se izračunati razdaljina i potrošeno vreme za svaki pojedinačni put.

```
Calculating sums per trip
+-------+-----+----------+------------------+--------------------+
|user_id|label|  label_id| total_distance_km|   total_duration_h|
+-------+-----+----------+------------------+--------------------+
|    010|  bus|8589934627|15.808112079995405| 0.7408333333333326|
|    010| taxi|8589934599| 4.997940890116999|0.14805555555555558|
|    010| taxi|8589934601| 6.031302670990729|0.18083333333333337|
|    010| taxi|8589934603| 5.266303014263252|0.13138888888888892|
|    010| taxi|8589934605| 4.282445785187889|0.09777777777777777|
+-------+-----+----------+------------------+--------------------+
only showing top 5 rows
```

Konačno, korišćenjem group by i funkcija min, max, avg, count može se izračunati statistika korišćenja prevoznog sredstva korisnika.

```
Calculating statistics per transportation mode
+-------+-----+---------------+-------------------+-------------------+-------------------+-------------------+--------------------+-------------------+
|user_id|label|number_of_trips|       min_trip_km|       max_trip_km|       avg_trip_km|max_trip_duration_h| min_trip_duration_h|avg_trip_duration_h|
+-------+-----+---------------+-------------------+-------------------+-------------------+-------------------+--------------------+-------------------+
|    010|  bus|              1| 15.808112079995405| 15.808112079995405|15.808112079995405| 0.7408333333333326|  0.7408333333333326| 0.7408333333333326|
|    010| taxi|              4|  4.282445785187889|  6.031302670990729| 5.144498090139717|0.18083333333333337| 0.09777777777777777| 0.1395138888888889|
|    010|train|              6|   244.909496040808|  782.5191195460434| 466.8920486845687|   9.21638888888875|  1.8688888888888842|   5.01175925925924|
|    010| walk|              3|0.13761899378345765|0.34808975175238804|0.2531838390784024|0.11444444444444446|0.032777777777777774|0.07666666666666667|
|    068| bike|             11|0.19818101632308704|   4.48607296391748|1.6283949277247012| 0.7155555555555533|0.027777777777777766|  0.169873737373737|
+-------+-----+---------------+-------------------+-------------------+-------------------+-------------------+--------------------+-------------------+
only showing top 5 rows
```

Dnevno korišćenje prevoznih sredstava

Analiza se može naći u jobs -> daily_statistics

Analiza podrazumeva korišćenje prevoznih sredstava na dnevnom nivou. Analiza sadrži sledeće statistike:

- Najkraća i najduža razdaljina koju je korisnik dnevno prešao određenim prevozom
- Najduže i najkraće dnevno korišćenje prevoznih sredstava
- Prosečno dnevno korišćenje određenog prevoznog sredstva - vreme i dužina

Prva dva koraka ove analize se rade na isti način kao u prethodnoj sekciji:

1. Učitati skup podataka, ujediniti puteve i labele
2. Naći parcijalne distance između sukcesivnih tačaka (iako se analiza radi na dnevnom nivou, i dalje je bitno da se distanca računa na smislen način)

Razlika između ove i prethodne analize je način sumiranja parcijalnih distanci. U ovoj analizi, sumiramo sve parcijalne distance za isti dan. Na taj način dobijamo dnevnu pređenu distancu za to prevozno sredstvo i dnevno utrošeno vreme.

```
Calculating daily sums
+-------+-----+----------+-------------------+--------------------+
|user_id|label|      date|  daily_distance_km|    daily_duration_h|
+-------+-----+----------+-------------------+--------------------+
|    010| walk|2008-04-06|0.13761899378345765|0.032777777777777774|
|    068| bike|2008-09-18|   9.34960330812416|  1.1316666666666626|
|    010|train|2008-04-01| 490.77153416305066|   6.241388888888861|
|    010| taxi|2008-04-01| 15.580051470441873| 0.41000000000000014|
|    068|  bus|2008-09-20|  15.67700590195688|  1.0052777777777517|
+-------+-----+----------+-------------------+--------------------+
only showing top 5 rows
```

Na osnovu dnevnih vrednosti, možemo dalje raditi statistiku dnevnog korišćenja prevoznih sredstava:

- Najveća i maksimalna pređena razdaljina
- Prosečno dnevno korišćenje

```
Calculating daily statistics
+-------+-----+--------------------+--------------------+--------------------+-------------------+-------------------+-------------------+
|user_id|label|max_daily_distance_km|min_daily_distance_km|avg_daily_distance_km|max_daily_duration_h|min_daily_duration_h|avg_daily_duration_h|
+-------+-----+--------------------+--------------------+--------------------+-------------------+-------------------+-------------------+
|    010|  bus|   15.808112079995405|   15.808112079995405|   15.808112079995405|  0.7408333333333326|  0.7408333333333326|  0.7408333333333326|
|    010| taxi|   15.580051470441873|    4.997940890116999|   10.288996180279437| 0.41000000000000014| 0.14805555555555558|  0.2790277777777779|
|    010|train|   1031.4713652710404|   490.77153416305066|    700.3380730268525|    12.3494444444443|   5.432777777777754|   7.517638888888831|
|    010| walk|   0.6219325234517497|  0.13761899378345765|  0.37977575861760365| 0.19722222222222227|0.032777777777777774| 0.11500000000000002|
|    068| bike|     9.34960330812416|    8.562740896847556|    8.956172102485858|  1.1316666666666626|   0.736944444444442|  0.9343055555555523|
+-------+-----+--------------------+--------------------+--------------------+-------------------+-------------------+-------------------+
only showing top 5 rows
```

Reference

https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide

[1] Yu Zheng, Lizhu Zhang, Xing Xie, Wei-Ying Ma. Mining interesting locations and travel sequences from GPS trajectories. In

Proceedings of International conference on World Wild Web (WWW 2009), Madrid Spain. ACM Press: 791-800.[2] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, Wei-Ying Ma. Understanding Mobility Based on GPS Data. In Proceedings of

ACM conference on Ubiquitous Computing (UbiComp 2008), Seoul, Korea. ACM Press: 312-321.

[3] Yu Zheng, Xing Xie, Wei-Ying Ma, GeoLife: A Collaborative Social Networking Service among User, location and trajectory.

Invited paper, in IEEE Data Engineering Bulletin. 33, 2, 2010, pp. 32-40.