BE0M33BDT – Homework

Deadline: January 10th, 2020

Send to jan.hucin@profinit.eu both source code of commands and results that you get. Make a separate text file for each section.

A. Hive

Source data

We will work with monthly ratings of chess players from December 2017 to December 2019. The file with data is on the local filesystem:
`/home/pascepet/fel_bigdata/data/chess-ratings.zip`

Final table

The goal is to make a Hive managed table in your database from this data. The table should have these properties:

- managed (internal) table;
- containing only fields *name, fed, sex, rat, gms, bday, year, mon*
- format ORC with ZLIB compression;
- containing only records with non-empty name and non-zero rating and with sex “F” or “M”.

Analytic queries

How many records are there in the final table?

What is the difference of average ratings of Germany men (fed = GER) between December 2017 and December 2019?

For every player, we consider the maximum of all his/her ratings. Among women find five of them that have the highest maximum ratings.

B. Spark RDD

Source data

We will use the file of songs and lyrics.

- path (on HDFS): `/user/pascepet/data/lyrics/lyrics.csv`
- separator: `','` (comma)
- header: no
- fields: id, name, year, interpret, genre, text

Read the file into a RDD.

Text-mining

Keep only songs with the interpret ‘eminem’ and with a non-empty text. How many songs are there?

In next work, use words converted to lowercase. Find 20 most frequent words in texts of songs. If a word is multiple times in one song, consider it multiple times.

Do the same, but consider only words with at least 3 characters.

C. Spark SQL

Source data

We will use the same file of songs and lyrics as in the B section:

- path (on HDFS): `/user/pascepet/data/lyrics/lyrics.csv`
- separator: `','` (comma)
- header: no
- fields: id, name, year, interpret, genre, text

Read the file into a DataFrame (with automatic schema inferring).

Rename columns to have proper fields names (see above).

Cache the DataFrame into memory.

Exploratory analysis

How many songs are there? How many songs are from the year 2000?

How many interprets do have 500 songs or more? Who are they?

Advanced analysis

Create new column *word_cnt* containing number of words in each song’s text.

Consider only songs with non-empty text. Count average of *word_cnt* for each genre and display it, but only for genres with at least 20 000 songs.