BDMM - 1st Project

April 22, 2021

1 Big Data Modeling and Management Assigment

1.1 The Beer project

As it was shown in classes, graph databases are a natural way of navegating distinct types of data. For this first project we will be taking a graph database to analyse beer and breweries!

For reference the dataset used for this project has been extracted from kaggle, released by Evan Hallmark. Even though the author does not present metada on the origin of the data it is probably a collection of open data from places like beeradvocate

Problem description Explore the database via python neo4j connector and/or the graphical tool in the NEO4J webpage. Answer the questions. Submit the results by following the instructions

Connection details to the neo4j database

```
Host: rhea.isegi.unl.pt:7474
Username: neo4j
Password: F3cfcrnvBev57KZ8mcMk78L9wHgJVZuJ
Connect URL : bolt://rhea.isegi.unl.pt:7687
```

Questions

0. **Example Question** *How many beers does the database contain?*
1. How many different countries exist in the database?
2. Most reviews:
    1. Which `Beer` has the most reviews?

    2. Which `Brewery` has the most reviews for its beers?
    3. Which `Country` has the most reviews for its beers?
3. Find the user/users that have the most shared reviews (reviews of the same beers) with the user CTJman?
4. Which Portuguese brewery has the most beers?
5. From those beers (the ones returned from the previous question), which has the most reviews?
6. On average how many different beer styles does each brewery produce?
7. Which brewery produces the strongest beers according to ABV?
8. If I typically enjoy a beer due to its aroma and appearance, which beer style should I try?
9. Using Graph Algorithms answer the questions:
    1. Which Countries are most similiar when it comes to the most produced Beer styles
    2. Which beer has the most similar reviews as the beer `Super Bock Stout`

10. If you had to pick 3 beers to recommend using only this database, which would you pick and why?

Questions 8 to 10 are somewhat open, which means we'll also be evaluating the reasoning behind your answer. So there aren't necessarily bad results there are only wrong criteria, explanations or execution.

Groups Groups should have 4 to 5 people
You should register your group on moodle. An email will be going out to everyone with the credentials for the database to use when storing the results.

Submission Submission of the query results to be done to the group's redis database (explained on the first class, credentials sent via email).
The following format is expected:

```
>>> redis.set("0", "358873")
```

This result should be the anwser of the group to question 0

The code used to produce the results and respective explations should be uploaded to moodle. They should have a clear reference to the group, either on the file name or on the document itself. Preferably one Jupyter notebook per group.

Delivery date: Until the **midnight of May 2nd**

Evaluation This will be 20% of the final grade.
Each solution will be evaluated on 2 components: correctness of results and simplicity of the solution.
All code will go through plagiarism automated checks. Groups with the same code will undergo investigation.

Note: Remember the Neo4j is a shared database and when creating in-memory graphs please use your group's prefix.
Ex. Instead of `my-graph` as the name of your graph please use `group0-my-graph`.

[ ]: