Data Engineering — Intermediate Report 1

Daan Brugmans (s1080742)
Radboud University
Nijmegen, Netherlands
daan.brugmans@ru.nl

1 DATA & PROJECT PROPOSAL

I would like to work with data sources that I think are an interesting alternative to the types of data that are commonly used by students in academia. If my project proposal and/or data sources are considered inappropriate for the purposes of the Data Engineering course, please notify me. I will then search for data sources from a different domain.

For the Data Engineering project, I would like to build a pipeline that serves data to data scientists working on a beer recommendation system. The pipeline should aggregate from multiple sources that serve data(sets) of beer and their attributes. Examples of attributes are name, brewery, alcohol by volume (ABV), international bitterness unit (IBU), category/style, textual description, flavor profile, and personal ratings. The pipeline should serve two sets of data to the data scientists: a big dataset containing beer attributes from varying publicly available sources that can be updated regularly, and a small dataset of an end user of the beer recommendation system that includes personal ratings of beers they have had before. These datasets will be served as tabular data. The data scientists can then use the ratings of the end user and the data of both datasets to build a system that can recommend beers from the big dataset to the end user's non-specified preferences, that can be extracted from the data using machine learning. A realistic use case for such a recommendation system could exist for social media platforms revolving around beer and online retailers of beer that can use the system to increase revenue and beer sales.

For the small dataset of an end user, I will provide my own data of beer ratings that I have collected on Untappd. Untappd [2] is a social medium where users rate beers they try and share their ratings with friends by registering their rating on the medium ("check-in"). I participate in Untappd and have collected a dataset of these check-ins with ratings. I can access an export of my check-in data per GDPR request. Although I do not know yet which attributes are included for certain in this export, I expect they include name, category/style, brewery, check-in location, purchasing location, flavor profile experienced by the end user, rating, and possibly timestamp. The data will be provided in CSV format and contain almost 400 check-ins, slightly over 300 of which being unique.

For the big dataset containing beers that can be recommended to an end user, I have found multiple data sources of beer attributes. I want to aggregate the data from these sources into a single database of beers. Because my data sources serve data in varying formats, I expect that the aggregating of my data into a single collection will be a major challenge of the project. For example, I will have to make sure that all data uses the same formats and standards, that there are no duplicates due to beers being present in multiple sources, and that there will likely be missing data due to some data sources not storing certain attributes.

I want to make use of the following data sources:

- OpenBeerDB [1] is an archive of beer data. Although it currently serves an older, static dataset, OpenBeerDB is planning on updating this dataset in the future, possibly including regular updates to the data. The OpenBeerDB data is served as a set of `.sql` files that will create tables for beer, breweries, categories, and styles (finer granularity categories), and insert them with data. The resulting SQL database contains most attributes I would want to serve to the data scientists, with the exception of flavor profile.
-

REFERENCES

[1] OpenBeerDB. 2024. OpenBeerDB. https://openbeerdb.com/
[2] Untappd. 2024. Untappd. https://untappd.com/

DE'23-'24, Radboud University, Nijmegen, Netherlands
2024.