PhD course on Big Data – Project description

We provide two examples of a publicly available datasets, and some queries that can be done of such datasets. The student should choose one of the two datasets and prepare a notebook with:

- Some exploratory analysis (see examples below).
- The answer to one/two queries among the ones proposed for that dataset.

Alternatively, the student may propose a different dataset and the corresponding queries, if such dataset is related to the student's research activities.

Yelp dataset

Consider a dataset collected by a website for maintaining the information related to users' reviews about businesses (restaurants, pubs, etc.). The complete description of the dataset can be found at https://www.yelp.com/dataset/documentation/main

Here we report some relevant tables and fields:

- *Business*: information about the business, such as its unique ID, its name, its location;
- *User*: information about the registered user, such as the unique ID, the name, the number of reviews and her/his friend IDs;
- *Review*: each entry contains a review made by a user (identified by her/his ID) for a business (identified by its ID).

A sample with 10k entries of the relevant tables described above can be found at:
https://univr-my.sharepoint.com/:u:/g/personal/damiano_carra_univr_it/EajzSVc_wllEgAmw3t7uUggBoOE8_rZ3saGgiTBaMoV0zg?e=ljBd5C
Note that, since the dataset has been sampled, the result of some queries may be empty. Clearly, we will evaluate the correctness of the query, rather than the specific result.

Exploratory analysis (can be done on the "Review" table):

- Number of reviews for each business (and its distribution).
- Number of reviews for each user (and its distribution).
- Average score received by each business.
- Average score given by each user.
- Top K businesses (e.g., K = 10) with at least R reviews (e.g., R>20), i.e., the K businesses with the highest average rate that have at least R reviews.

Queries:

- For a given user, provide the average difference between the vote she/he gave to a business, and the average vote that business has. For instance, if the user visited two business whose average score is 4.5 and 4.8, and she/he gave 3 to both of them, then the average difference would be: ((4.5-3) + (4.8-3))/2;
- Answer to the above query with different time granularities, e.g., in the last month, year, two years.
- For a given business, provide the average number of reviews of the users that wrote a review for that business. For instance, if user A and B wrote a review for a business, and user A has written 50 reviews in total, while user B has written 30 reviews in total, then the average number of reviews of the users that wrote a review for that business is (50+30)/2;
- Answer to the above query with different time granularities, e.g., in the last month, year, two years.
- For a given city (e.g. San Francisco), the average rating given for businesses in that city in the last month or year;
- Answer to the above query showing the evolution over time, e.g., the average rating per year in the last 10 years.

Movielens dataset

Consider a dataset that contains the ratings that users assigned to movies:
https://grouplens.org/datasets/movielens/

(for the project, the student may consider small older datasets, such as the 1M dataset)

The dataset is composed by multiple files (tables), such as:

- *Movies*: information about the movie, such as its movieId, title and genres;
- *Ratings*: each entry contains a rating made by a user (identified by her/his ID) for a movie (identified by its ID) and a timestamp.

Exploratory analysis (can be done on the "Ratings" table):

- Number of ratings for each movie (and its distribution).
- Number of ratings for each user (and its distribution).
- Average score received by each movie.
- Average score given by each user.
- Top K movies (e.g., K = 10) with at least R ratings (e.g., R>20), i.e., the K movies with the highest average rating that have at least R reviews.

Queries:

- Find if there is a correlation between the standard deviation of the ratings a movie has received, and the number of ratings.
- Find the evolution over time (with a granularity of N months) of the number of ratings and the average rating: do high rated movies maintain their ratings? Are low rated movies "abandoned" after a while?
- Find how the average rating of each movie changes as we progressively remove the ratings from users that rated more and more movies. For instance, we can identify different groups of users (who rated less than 10 movies, who rated between 11 and 30 movies, …) and we can compute the average rating considering all the groups, then only the groups of users with at least 11 ratings, and so forth.
- [Hard] Is it possible to identify groups of similar movies based on the ratings they received from the users? For instance, if movies m_1 and m_2 have both obtained 5 stars from users u_1 and u_2, they may be considered similar.