CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computing@CIT

DISTRIBUTED DATA MANAGEMENT

Assignment 1 – Part 1: MongoDB

Background.

Rachel Allen is one of the most famous Cork-based Irish chefs, well known for her work on television and as a writer. Let’s suppose that, after years earning some savings, Rachel is decided to follow one of her dreams: Open her own new restaurant in New York City!

She has been deep in thought about how to make this project successful:
What is New Yorkers taste like? How often do they go to restaurants? How much do they pay? What do they prefer and valuate the most? Do her famous recipes fit into the new market, or should they be adapted at somehow?

The research of Rachel has achieved a milestone: She has found out a database with information of all the restaurants of New York. This database, with tenths of thousands of restaurants, is on the form of a MongoDB collection. Rachel regrets of having no knowledge of MongoDB in particular (and of data analytics in general), as the data might contain some insights answering the following questions:

1. What kind of cuisine do New Yorkers prefer?
2. Which area represents the biggest market opportunity for opening a new restaurant of this kind of cuisine?
3. Who are the biggest competitors in this area?

As you can imagine, your goal for the assignment is to help Rachel on answering these questions by applying data analytics on the aforementioned MongoDB collection.

The MongoDB Collection.

The MongoDB collection of restaurants used as dataset for the assignment can be found at the file **restaurants_dataset.json**. Each line of the file represents 1 document (restaurant), and provides the following information:

- The name of the restaurant (together with a unique identifier).
- The borough of the city in which the restaurant is placed, together with its complete address (including its zipcode, street, building and coordinates).
- The kind of cuisine offered in the restaurant.
- A list of reviews from customers, including the date, grade and score of each review.

The following document / JSON object (used as an example) is one of the documents of the collection:

```
{
        "address": {
                "building": "1007",
                "coord": [-73.856077, 40.848447],
                "street": "Morris Park Ave",
                "zipcode": "10462"
        },
        "borough": "Bronx",
        "cuisine": "Bakery",
        "grades": [
                { "date": { "$date": 1393804800000 },  "grade": "A",  "score": 2 },
                { "date": { "$date": 1378857600000 },  "grade": "A",  "score": 6 },
                { "date": { "$date": 1358985600000 },  "grade": "A",  "score": 10 },
                { "date": { "$date": 1322006400000 },  "grade": "A",  "score": 9 },
                { "date": { "$date": 1299715200000 },  "grade": "B",  "score": 14 }
                 ],
        "name": "Morris Park Bake Shop",
        "restaurant_id": "30075445"
}
```

Goal of the Assignment.

As mentioned before, Rachel would like to apply some data analytics on the restaurants collection, as the data might contain some insights answering the following questions:

1. *What kind of cuisine do New Yorkers prefer?*
2. *Which area represents the biggest market opportunity for opening a new restaurant of this kind of cuisine?*
3. *Who are the biggest competitors in this area?*

She will make her final decisions based on the following approach:

i. Instead of opening an Irish-based restaurant in which she will directly apply her famous recipes, she will adapt them to whatever style New Yorkers like the most. Thus, she wants to know the kind of cuisine with higher number of restaurants in the city. Whatever this cuisine-style is, she will open a restaurant of this kind.

ii. Once the kind of cuisine is fixed, she wants to decide the borough in which she will open the new restaurant. For this reason, she wants to know the ratio (percentage) of restaurants of this kind of cuisine per borough. Her approach would be to pick the borough with smaller ratio (i.e., the one with less competence).

Of course this approach might not be ideal. For example, if Manhattan has a 50% of restaurants of type A and Brooklyn has only a 20%, perhaps is because people in Brooklyn are not that much interested in restaurants of type A. In other words, there is a risk. But, come on, there is going to be a risk whatever the approach being followed. So that's why we are doing data analytics, to get some insights helping she *to maximise the chances* of making good decisions.

iii. Once the kind of cuisine and the borough are fixed, she wants to follow the same approach for the zipcode of the borough in which she will open the restaurant. She will only consider the 5 best zipcodes (i.e., the 5 zipcodes of the borough with higher number of restaurants in total). For these 5 zipcodes, she wants to know the ratio (percentage) of restaurants of this kind of cuisine. Her approach would be to pick the zipcode with smaller ratio (i.e., the one with less competence).

iv. Finally, once the kind of cuisine, borough and zipcode for the new restaurant are fixed, she wants to know which ones are the 3 biggest competitors of the zipcode. That is, which are the best 3 restaurants of this very same kind of cuisine, placed in this very same borough and zipcode. Her approach would be to visit them all to see how to differentiate her new restaurant from them. To select the three restaurants to visit, she wants to follow the reviews available in the collection. She wants to consider only restaurants with, at least, 4 customer reviews, and then select the 3 with best (higher) average review scores.

Exercise.

The Python file **data_analysis.py** is to be filled to perform the data analytics answering Rachel's questions. It uses the pymongo library to connect to a **mongod** server and query the restaurants collection.

Complete the functions:

i. **most_popular_cuisine**:
   - It receives the test database of the cluster in which the restaurants collection is.
   - It returns the name of the kind of cuisine with higher number of restaurants in New York and its ratio (percentage).

ii. **ratio_per_borough_and_cuisine**:
   - It receives the test database and the kind of cuisine we are interested in.
   - It returns the name of the borough with smaller percentage of restaurants of this kind of cuisine. It also returns the proper percentage.

iii. **ratio_per_zipcode**:
   - It receives the test database, the kind of cuisine and the borough we are interested in.
   - It returns the name of the zipcode with smaller percentage of restaurants of this kind of cuisine. It also returns the proper percentage.

iv. **best_restaurants**:
   - It receives the test database, the kind of cuisine, borough and zipcode we are interested in.
   - It returns the names of the 3 best restaurants (the ones with, at least, 4 reviews and higher average scores). It also returns the proper average scores.

Additional Requirement.

When coding the functions, minimise the amount of data being transferred from the cluster to the Python program.

That is, one possible approach would be to gather all the documents of the collection and bring them to the Python program (so as to process them within it). We do not want to follow this approach, as the amount of data to be transferred represents an unaffordable bottleneck!

Instead, use the MongoDB aggregation framework. That is, within the Python program create the pipeline of steps to process the query. Then, connect to the cluster to trigger the entire pipeline of commands there. In this context, the cluster will only transfer back to the Python program the final list of documents achieved as a result of executing the entire pipeline.

Marking Scheme and Submission Date.

- Total marks: 30 (7.5 marks per function).
- Submission: Upload to Canvas the filled file **data_analysis.py**.