MongoDB Assignment

Task Description

In this coursework work you will write spark code to load two datasets and import them into MongoDB. Then, you will write queries in MongoDB query language.

Data Description

This task consists of two datasets.

1) **Zip codes**: This dataset contains zip codes of US cities, along with their location and state.
   Download: http://www.doc.ic.ac.uk/~theinis/zips.json
   Docker location: `/zips.json`

2) **Prescriptions**: This dataset contains prescription records by various doctors. More information about the dataset can be found here and here.
   Direct link: http://www.doc.ic.ac.uk/~theinis/prescriptions.jsonl.zip
   Docker location: `/prescriptions.jsonl`
   Notes:

- NPI (National Provider Identifier) is a unique Identifier for healthcare providers (simply assume they are all doctors). Each JSONL record belongs to a unique doctor.
- Assume that the value of each drug is the "number of times it is prescribed".

Both datasets are in JSON-lines format, a special case of JSON format where each line contains a separate JSON object. The JSON-lines format makes it possible to load JSON records in Spark.

Write the following queries in MongoDB and run against the collection that you created in part A.

Queries for "Zip codes"

1) Count the total number of cities in Washington state (code: "WA").
2) Find the total population of each state (i.e., sort states by their population in the ascending order).
3) Find the 10 closest cities to WEST BROOKLYN, IL. You might want to use the "$near" operator.
4) Considering the `region` of each US state, according to this source, find the total population in each of the four regions (West, South, Midwest, and Northeast).
5) Find the 3 most populated cities for each state (in one query).

Queries for "Prescription-based prediction" dataset

6) Find the specialty of all doctors who have prescribed "HALOPERIDOL".
7) Find the total number of doctors, separately for each region (in one query).
8) Find the total amount of prescribed "ATORVASTATIN CALCIUM"

9) Find the drug that is prescribed by the most of doctors working in "non-urban" areas. (in terms of number of doctors who prescribed it, not the total amount of prescriptions).

10) Considering the region of US states (Query #4) and the region of each doctor, find the average number of doctors per capita in each of the four regions in US.

Notes

- Please upload both the .ipynb, .html, and the PDF of your notebook.
- Do not print intermediate data in the notebook.
- Do not use Python as a part of query. Python can only be used to prepare and print the output.
- If installing MongoDB manually, Deploy the latest version of MongoDB (3.4.4 or above), as you might need some operators introduced from v.3.4.4.
- The output should be printed nicely. Do NOT simply print an array or JSON output. Instead, print the results such that somebody can easily read it.
- Briefly explain your code and query. If you make any assumptions, please highlight them.
- Take a look at the schema and try to understand the data. Make sure that you handle missing or incomplete values (e.g. NULL values). Also, use column(s) that are most accurate and meaningful for each task.
- The efficiency of your code is important.