MongoDB Coursework

Task Description

In this coursework work you will write spark cods to load two datasets and import them into MongoDB. Then, you will write queries in MongoDB's 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 link]
2) Prescription-based prediction: This dataset contains prescriptions records by various doctors. More information about the dataset can be found here and here. [Download link]

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.

Tasks

The task consists of two parts, to be done on both datasets.

Part A

Load each dataset using Spark and persist it directly into MongoDB. Feel free to use Spark SQL facilities. You should share the code as well the schema of the dataset.

Part B

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 ascending order).
3) Find the 10 closes 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). Use linking to do it.

Queries for "Prescription-based prediction" dataset

5) Count the total number of records
6) Find the specialty/-ies of all doctors who have prescribed "HALOPERIDOL".
7) Find the total number of patients visited, separately for each region (in one query).
8) Find the total amount of prescribed "ATORVASTATIN CALCIUM"
9) Find the drug that is prescribed the most in "non-urban" areas. (in terms of number of prescriptions, not the total amount).

Queries on both datasets

10) * Considering the region of US states (Query #4) and the region where each prescription is recorded, find the average number of prescriptions per capita in each of the four regions in US. Again, use linking and integrate in the query code.