Homework 4 – MongoDB & Vector Databases
100 Points

Due: Monday, May 5, 11:59PM EST

All datasets are in Jupyuterhub shared drive

Q1. Write MongoDB queries (2 points each) 40 Points:

Datasets: restaurants.json

1. Count the number of documents in the collection.
2. Display all the documents in the collection.
3. Display: restaurant_id, name, borough and cuisine for all the documents
4. Display: restaurant_id, name, borough and cuisine, but exclude field **_id,** for all the documents in the collection
5. Display: restaurant_id, name, borough and zip code, exclude the field **_id** for all the documents in the collection.
6. Display all the restaurants in the Bronx.
7. Display the first 5 restaurants in the Bronx
8. Display the second 5 restaurants (skipping the first 5) in the Bronx.
9. Find the restaurants that achieved a score, more than 80 but less than 100.
10. Find the restaurants that do not prepare any cuisine of 'American' and their grade
    score more than 70 and latitude less than -65.754168.
11. Find the restaurants which do not prepare any cuisine of 'American ' and achieved a grade point 'A' and not in the borough of Brooklyn, sorted by cuisine in descending order.
12. Find the restaurant Id, name, borough and cuisine for those restaurants which contain 'Wil' as first three letters for its name.
13. Find the restaurant Id, name, borough and cuisine for those restaurants which contain 'ces' as last three letters for its name.
14. Find the restaurant Id, name, borough and cuisine for those restaurants which contain 'Reg' as three letters somewhere in its name.
15. Find the restaurant Id, name, borough and cuisine for those restaurants which belong to the boroughs of Staten Island or Queens or Bronx or Brooklyn.
16. Find the restaurant Id, name, borough and cuisine for those restaurants which are not belonging to the borough Staten Island or Queens or Bronx or Brooklyn
17. Find the restaurant Id, name, borough and cuisine for those restaurants which achieved a score below 10.

18. Find the restaurant Id, name, borough and cuisine for those restaurants which prepared dish except 'American' and 'Chinese' or restaurant's name begins with letter 'Wil'.
19. Find the restaurant Id, name, and grades for those restaurants which achieved a grade of "A" and scored 11 on an ISODate "2014-08-11T00:00:00Z" among many of survey dates.
20. Find the restaurant Id, name and grades for those restaurants where the 2nd element of grades array contains a grade of "A" and score 9 on an ISODate "2014-08-11T00:00:00Z".

Q2. Restaurant foreclosures in North Carolina – 60 Points

Some Background: geospatial logic is possible in MongoDB using the geopastial library/facilities. https://docs.mongodb.com/manual/geospatial-queries/

Datasets:
Restaurants_in_Durham_County_NC.csv
durham-nc-foreclosure-2006-2016.json

2.1. Find the **center point (geolocation) or centroid** for the region that includes all Rpt_Area_Desc=”restaurants” and Seats>=40.

2.2. Find the number of foreclosures within 2 miles of the Point in Q2.1 above.

Q3. Extra Credit – 40 points
Book Sales Analysis - Store and Query Book Sales Data

You are given a dataset of book sales from an online bookstore. Your task is to:
- Download the dataset from the provided URL.
- Import the dataset into a MongoDB collection named 'book_sales'.
- Perform queries to answer specific questions about the data using Python.

Dataset
Book Sales Dataset (CSV)
https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/books.csv

Schema: book_id, title, authors, average_rating, language_code, ratings_count, and publication_year.
Instructions

4. Write queries to:
  - a) Find the top 5 highest-rated books.
  - b) Find the number of books published after 2010.

- c) Find the average rating of books with more than 100,000 ratings.
- d) Find the distinct languages the books are published in.

Expected Results

Task	Example Output
Top 5 rated books	List of titles and ratings
Books after 2010	245
Avg rating of highly rated books	4.34
Languages	['eng', 'spa', 'en-US', ...]