CSE 4510/5310 – Mgmt & Proc Big Data

Fall 2022

Term Project Guidelines

Instructions: Please use the following guidelines when planning your term project. After selecting a project, please post your topic and group members list on Piazza so the class can know what ideas are already taken.

The term project is a project with a big data component. You will be required to process at least 5GB of data using a big data architecture. You will be given access to several AWS platforms via the AWS Academy. Please use your credits sparingly as you will not have access to the platform when your credit expires. The project is divided into the following components:

Milestones

- Group Selection (must have 8-9 members per group)
- Project proposal (5 points)
- Project proposal presentation (10 points)
- Literature survey (10 points)
- Report outline (5 points)
- Final report (10 points)
- Final presentation (10 points)

Potential Topics/Problem Description

1. Using a recipe dataset provided by the instructor, build a web application that allows people to filter out foods based on allergies (Florida Tech's Panther Dining Hall menu manager (https://app.mymenumanager.net/fit/) lists allergies based on ingredients). Use a machine learning model to filter instead of checking for hard-coded strings like wheat and soy.
2. Download FBI crime data from the FBI API (https://api.data.gov/docs/fbi/) and visualize hot spots by cities on a map. Provide the user with insights and trends of crime over time.
3. Download housing data from the Zillow API (https://www.zillowgroup.com/developers/api/rentals/public-api-spec/) or similar API and do an analysis to see how the prices of homes or rent has changed throughout the COVID-19 pandemic.
4. Use the cost of goods and services over time (https://data.world/datasets/price) to evaluate the impact of the war in Ukraine on inflation across the world versus other wars in history
5. Use data from UNCHR (https://www.unhcr.org/refugee-statistics/) to visualize on a map the destinations that refugees impacted by the war in Ukraine end up. Determine the impact of these migrations on the source country and the destination country.
6. Build a web interface that allows users to search in big data (e.g., Health records (https://healthdata.gov/browse), census data (https://www.census.gov/data/developers/data-sets.html, ... etc.) for insights.
7. Learn Neo4j for handling large graphs. Then consult the instructor about a health project he is working on to build a graph and analyze the dataset.
8. Collect tweets (https://developer.twitter.com/en/docs/twitter-api) and use them to run some correlation analysis or sentiment analysis, etc., on how people in different states perceive brands (car brands, food brands, ... etc.).

9. Collect tweets (https://developer.twitter.com/en/docs/twitter-api) regarding prevalent diseases such as Monkeypox or COVID-19 and use them to run sentiment analysis, e.g., how people in different states/regions perceive the disease.

10. Download satellite data (https://developers.google.com/maps, https://docs.aws.amazon.com/location/latest/developerguide/location-actions.html, https://docs.microsoft.com/en-us/bingmaps/rest-services/imagery/) and find the correlation between temperature, vegetation, precipitation, and fires. For example, you can compute the average per day/week/month/season/year and show how the averages change over time.

11. Collect census data (https://www.census.gov/data/developers/data-sets.html), places of interest (POI - [https://developers.google.com/maps/documentation/places/web-service/overview])) data, lakes, parks (https://www.nps.gov/subjects/science/science-data.htm), ..., etc. and try to rank cities in the US by their quality of life.

12. Build a 3D road network and visualize it on Google Earth. Collect the road network from OpenStreetMap (OSM - [https://wiki.openstreetmap.org/wiki/Databases_and_data_access_APIs]) and adjust it with a Digital Elevation Model (DEM) to assign an altitude to it.

13. Extract a clean dataset from OpenStreetMap (OSM - [https://wiki.openstreetmap.org/wiki/Databases_and_data_access_APIs]), i.e., a dataset that can be directly used in applications.

14. Build an efficient sampler for text files in Hadoop. One that is faster than the existing one. For example, it could iteratively read bigger samples until some statistical measure is met.

15. Build a map of images that shows an image for each region where the regions change as we zoom in/out.

16. Build an interactive web visualizer that visualizes the functional map of the world (FMOW) which consists of satellite images and annotated regions (e.g., buildings or parks). Read more here: https://github.com/fMoW/dataset

17. Build an interactive 3-D visualization for a given geospatial dataset using existing 3-D visualization packages like Cesium, I3S or three.js. Feel free to choose any other 3-D visualizer if you wish to. The input dataset can be obtained from us or from open source environment like https://www.data.gov/

18. In all existing geospatial visualizations like Google Maps, Bing Maps or HadoopViz, the spatial datasets are preprocessed into small tiles and generated into maps. All of these systems work with static datasets (datasets that are constant and not updated periodically). If some new points are added to the input dataset, in order to visualize it, we need to rebuild the tiles from scratch. Build a similar system with a dynamic dataset that can help us view geospatial points on maps for a data stream like Twitter data. The system should be able to incorporate the new datapoints that were added to the input dataset without having to reconstruct the whole thing.