Project on Big Data Processing
CIS 490/590
Sunnie S Chung

Group Project with Presentation of a Project : 25%

You can choose one of the suggested projects below or you can create your own. You can change some details of the project that you choose from the list as you wish.

For some of those projects, you will need to get an account approval from the public social media sites like LinkedIn, Twitter, Yelp, Facebook sites to register your Project (App) as a developer and download data from the sites. Check their Developer/App/Tool sites for this process. Or you can choose any machine (server) generated log files for data to process for your project.

Submit 1-2 page proposal on a project your group choose to specify major tasks and tools to use and plan a time line by the deadline of Task 1

Each group (two person group) will give a 15 min presentation on a project and the related topics, tasks and APIs (tools) used for this project during the last week of the class in the semester. The group project presentation will be scheduled after the proposal submission.

Final Group Project will be done the following 4 Tasks. Check the deadlines of these 4 Task Submissions in the Project Section of the Class Website.

- **Group Project Proposal**
- **Group Project Status Report**
- **Group Project Presentation**
- **Final Group Project Report**

Project List
CIS 490/590
Sunnie S Chung

Big Data Project is to build an Intelligent System that collects, processes, and analyzes Big data in your Big Data Processing Infrastructure with NoSQL systems. (and/or RDBMS as needed)

This project consists of two phases of work as follow:

1. Phase 1: Big Data Processing Infrastructure commonly requires three tasks as follow:

   1-1. Data Collection:
   You can use any available APIs or tools (for example, Flume, Tweetpy) to collect/download a stream of Social Media Data to your local files. Or optionally you may have to write a simple script/program with REST API to collect data from the social media sites, which is a common method to collect data set in real life applications from web applications in those social media sites into your file system.

   1-2. Transform any streams of unstructured/semi structured data found in public social media sites or any system generated log files (web server log files or system log files) into the structured/semi-structured files

   1-3. Create permanent stores as databases/collections in a NoSQL system (MongoDB) and RDBMS that you learned in class to retrieve (query) data for further processing in Phase 2 as in real life applications for Data Analytics.

   Common structured/semi-structured file formats could be:

   - Table format in any RDBMS
   - CSV(Comma Separated Value), TSV (Tab Separated Value) or any delimiter separated text file format
   - Key Value Stores such as JSON file format as Document Collections in MongoDB
   - XML format generated by XML tools (XML Builder/Parser)

2. Phase 2: Big Data Analytics commonly requires three tasks as follow:

   2.1 Find Analytic Information by Writing a Complex Aggregation Pipelining queries from the Knowledge Bases that are Indexes in a Permanent Store (Databases/Collections) in RDBMS or NoSQL System that you created Or

Write a script that retrieves (in queries) some selective data you need from in RDBMS or NoSQL System to create an input as a Training set/Test set for any Data Analytics/Machine Learning Algorithms.

2.2 Build a Simple Search Engine for a Set of a Webpage Collection by building Inverted Index to build TF-IDF vector matrix of documents to calculate Similarity of each document to a user given query.
OR
Perform the Classification process using Machine Learning Algorithms to Get Analytic Results

2.3. Visualize your Data Analytic results using any existing APIs (tools) like Tableau

Suggested Data Sources

The suggested public social media sites or known data collection sites for data analytics are listed below with related industry research papers.

1. LinkedIn

2. Any well-known Newspaper or Magazine sites on Facebook:

3. Twitter Message data transformation:

5. Yelp Data Challenge: Business Data set

https://www.yelp.com/dataset_challenge

6. Transform log files in any system into either one of the platforms:

https://www.kaggle.com/datasets

Related papers to read: will be given

7. Webpage Processing for Text Analysis

Download all the webpages in one domain sites in any well known public news sites of your choice:

www.cnn.com
https://en.wikipedia.org/wiki/Main_Page
https://dumps.wikimedia.org : Wiki Page Data set
Or a collection of webpages in any website domain like
www.csuohio.edu

- You can build your Document Frequency and Inverted Index described in the Lecture Notes on Information Retrieval to build Any IR related Metrics in an algorithm
  For example, The Lecture notes show how cosine similarity is adopted as vector space scoring for document ranking. One example is building weight matrix by calculating weighted score based on Tf-Idf on page in 24 - 26 in the lecture note. Then you can calculated Cosine similarity between documents and the keyword using the weight score based on Tf-Idf you built. At the end of lecture notes, there are variations of the scoring matrix to optimize. Cosine normalization is one of them as well.
  OR
  Create your own Classification Process with Machine Learning Algorithms.
  Suggested Classification Tasks:
  - Categorizing each website into Sports, Entertainment, Politics, Social, Personal Webpage.
  - Sentiment Analysis of Twitter Messages or
  - Topic/Opinion Analysis of Twitter Collected on a Selected Topic
    For example, for a Collection of Twits on the Topic Corona Virus/COVID-19, find out which facts people are most interested in or which people are most worried about
  - Sentiment Analysis of Product/Business Reviews

8. Transform any electronic books or online documents for text processing analysis

   Any Electronic book on line
   See item 7 Webpage Processing above for processing.

For those who have already taken CIS660, The Final Group Project should include fully analytic processing:

- For Text Analytics like Sentiment Analysis or Opinion Analysis:
  NLP Techniques - POS, NER Tagging, Bi-Gram Handling are required for Preprocessing.
- For Document Categorization: Inverted Index Building is required to construct TF-IDF Vectorization.
- Building Word2Vec Embeddings for a Collection of Documents/Webpages with Training Set Generation in Skip Gram Model
- For Other Types of Projects, the Proposal is required to be approved to meet the complexity of Final Project.

Other Data Sources:

https://archive.ics.uci.edu/ml/index.php
https://www.kaggle.com/datasets

You can download preprocessed Wikipedia texts (in XML) here.
https://dumps.wikimedia.org

Text data sets for Sentiment Analysis
https://nlp.stanford.edu/sentiment/
https://nlp.stanford.edu/software/

Review Data Sources for Sentiment Analysis

Amazon Product Review Data:
http://jmcauley.ucsd.edu/data/amazon/

Movie Review Data
http://ai.stanford.edu/~amaas/data/sentiment/

https://datasets.imdbws.com/
http://www.imdb.com/reviews/index.html

Yelp Data Set
https://www.yelp.com/dataset/challenge