CS 410: Text Information System
Expert Search System
Project Progress Report

Name	**Net-Id**
Govindan Kutty Menon	gvmenon2
Harikrishna Bojja	hbojja2
Pushpit Saxena	pushpit2

Team Name: BayToBay
Team Captain: Pushpit Saxena (netid: pushpit2)

1) Which tasks have been completed?

- Design and architecture of the proposed modules of the project.
- Extract and Crawl Web Pages
    - Utilized scrapy framework with Python to crawl web pages and identify connected links.
    - Utilized BeautifulSoup toolkit with Python to extract text from web pages.
    - Removed special characters using regular expressions and extracted text data from web pages.
- Expert Bios page classification
    - PreProcessing Data
        - Using NLTK Python toolkit
            - Removed stop words
            - Stemming
    - XGBoost Text Classifier
        - Tf-Idf vectorizer

        - Assumes that tokens are stemmed and lowercase
        - Remove stopwords
        - Used the dataset from here as positive text and regular web pages as negative text.
    - Logistic Regression Classifier
        - Tf-Idf vectorizer
        - Assumes that tokens are stemmed and lowercase
        - Remove stopwords
        - Used the dataset from here as positive text and regular web pages as negative text.

2) Which tasks are pending?

- Extract and Crawl Web Pages
    - Increase scalability to crawl and extract data from web page with higher data content and connected links
    - Multi-thread extract and crawl scripts to achieve parallel processing
    - Utilize a delimiter to segregate page identifier and content information
    - Redesign process to accept input web pages from a file and make the process more configurable
    - Validation of script for different University web pages
- Expert Bios page classification
    - PreProcessing Data
        - Writing unit tests
    - Bi-LSTM text classification (using tensorflow/pytorch on Google Colab)
        - Use the data set from here
        - Compare accuracy against other classifiers
    - XGBoost Classifier
        - With word2Vec vectorizer (Glove)
        - Assumes that tokens are stemmed and lower case

        - Remove stopwords
        - Used the dataset from here as positive text and regular web pages as negative text.
- (Stretch goal) Topic modelling on expert bios.
- Integration
    - Integrate different components which are being developed individually
    - Integrate with existing functionality of the expert search system
- End to end script execution/ validation
    - Run scripts (in order) and validate conformance to the need/ original design.
- Submission/ Demo delivery
    - Finalize and create deliverables for project submission
    - Finalize and create deliverables for demo

3) Are you facing any challenges?

- Extracting and crawling web pages using lower powered CPUs and less memory on personal machines is posing a challenge from performance and scalability perspective.
- Extracting/ crawling web pages of different universities using a scrapy framework requires understanding of the page structure. Page structure could be different for different web pages and this is a challenge for crawling and scraping required contents.
- While we were able to collect positive training set for classifiers, collecting “quality” negative data set could be tricky
    - Collected positive training set from CS410-MP2
    - Trying to use general web crawled data for negative examples.