ExpertSearch Improvements

Progress Report

Topic Modeler

Assignee: Karthik Rajagopal (kr22@illinois.edu)

Updates:

Task	Progress
Implemented topic modeling using spaCy python library and it accurately predicts topics for individual bio pages	**Completed**
Integrating topic modeler in ExpertSearch system to display top 5 topics for each search result	**In Progress**

Challenges:

Running topic modeler for all bios is computationally extensive. We are incrementally optimizing the algorithm to achieve a better performance but this has proven to be challenging.

Bio Page Classifier

Assignee: Bashir Partovi (partovi2@illinois.edu)

Updates:

Task	Progress
Implemented a URL crawler using Scrapy to crawl Carnegie Mellon University and University of Maryland in order to generate	**Completed**

negative labels for classifier	
Using Keras, implemented a deep learning layer and trained it with the compiled bios from class project and the URLs that were crawled by the Scrapy spider, achieving 99% accuracy on the test data	**Completed**
Write a wrapper to load the model and use it with the URL crawler in order to identify bio pages	**In Progress**

Challenges:

It was difficult to understand how Keras deep neural network layers work, especially for someone who has never worked with the library before. In addition, transforming the text data into a feature vector in order to fit the model was very challenging.

Automatic URL Crawler

Assignee: Mohana Venkata Kalyan Cheerla (cheerla3@illinois.edu)

Updates:

Task	Progress
Code development to dynamically route from University home page to all its subsequent web pages and scraping their content has been completed. This scraping activity covers the extraction of the entire text information of those pages along with some important metadata that can feed additional information to "Faculty Bio Page Classifier" beyond what it requires today to help further improvements in future.	**Completed**
Adding additional filtering criteria to the crawler to eliminate crawling of uninteresting URLs. Configuration of Crawler to be more dynamic	**In Progress**

is also pending along with integrating it with the classifier	

Challenges:

Crawler runs for a long time and fetches over 40K+ web pages per university. Identifying the filter criteria to reduce the false positives is challenging.