Project Progress Report: Classification of Faculty directory pages and homepages

Author: Sameen Shaukat

Net ID: shaukat2

Overview:

ExpertSearch System is a search engine that is developed by students of University of Illinois to facilitate faculty search based on their research interests. The goal for this project is to enhance or aid the ExpertSearch System by writing a classification utility to identify good faculty directory and faculty webpage URLs. This utility/functionality can be added to the existing search system later on.

Goals:

Goals for this project are

- Identifying faculty directory URLs
- Identifying faculty webpage URLs

Proposed Solution:

Following methodology will be used as a starting point to achieve above stated goals.

1. Use data provided in Assignment MP2.1 and MP2.3 labeled as good examples and collect data for bad URLs and classify these as bad examples.
2. Based on this data, scrape these URLs to pre-process data.
3. Automatic or manual feature extraction based on data collected in step 2. This will be decided on the basis of results achieved.
4. Develop a utility in Python to identify good URLs from bad ones. This will involve thorough analysis in order to choose the best classification technique.
5. Integrate this utility in ExpertSearch System. (This is optional and time-dependent)

Progress Status:

Tasks Done or In Progress:

Data Collection: This part is completed to cater for both goals of this project. Bad URL examples were collected and added to the data set already available in MP2.1 and MP2.3. Labels were added to this dataset to help in classification.

Data Pre-Processing: This part is done almost 70 percent. I am still working on improving data extraction. The steps that I have followed so far are.

1. Scrape each URL to get data.
2. Remove tags to get text.
3. Remove Whitespaces & unwanted characters.
4. Remove Stop Words, Apply Stemming and Lemmatizing using NLTK
5. Replace email addresses and Web addresses with text tags

Feature Extraction: This part is still in progress. I have tried manual feature extraction and still want to explore automatic feature extraction as it is a particularly new area for me. Following features were extracted initially.

- No. of times professor, assistant, lecturer appears in text
- No of times faculty, staff and people appear in text
- No of times university, institute, school, department appear in text
- No of times email addresses appear in text.
- No of times Journal, Research, Publications appear in text.

This step is still in development. There is a chance I find more usable features to strengthen classification.

Tasks to be Done:

I intend to work on automatic feature extraction using simple bag of words model and enhancing it using TF-IDF. This is subject to change as per progress in project development and implementation as this is particularly a new area for me.

Classification model needs to be chosen. Currently, I am planning on using Naïve-Bayes, Linear Support Vector Machines, Logistic Regression, CNN (Time-Dependent) etc.

Average Accuracy Scores, F1-Scores and Execution Time on my current machine will be catered to provide performance comparison across all classification algorithms being used.

An integration utility needs to be written for integration with ExpertSearch System. (This is time-dependent)

Challenges:

Currently, I am having challenges with feature extraction as it is a comparatively new area for me as previously, I have only worked with full data set that had features already. Although, it is a challenge; it's also very exciting as I have learned a lot of techniques and tips during my research on this topic. Overall, I am content with the progress made so far.