**Final Project Report**

1. Function Overview

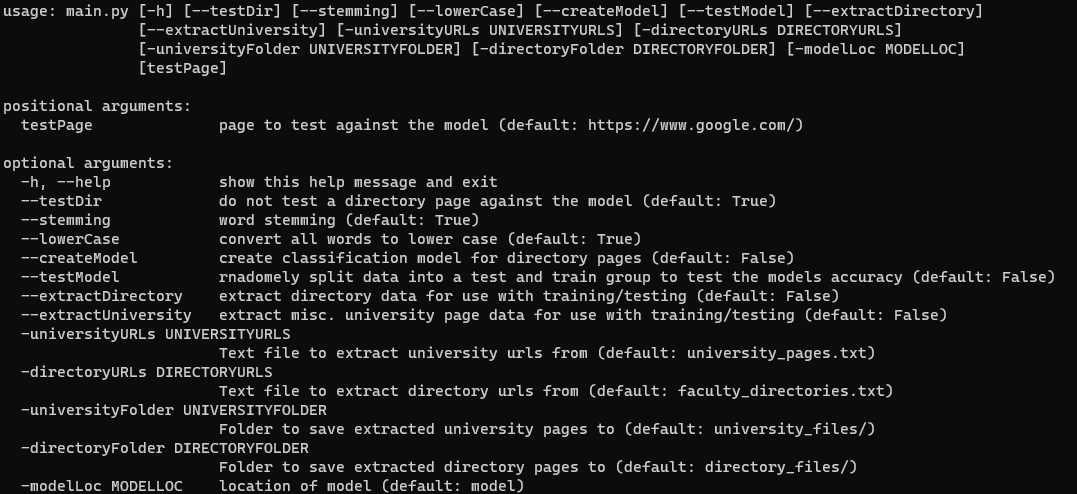
Directory page classifier with cli to control preparing the data for the model, creating the model, and testing against the model.

2. Implementation Detail

Built in Python using the bs4, requests, and nltk outside libraries. First to prepare data for the model, webpages are pulled from a list of directory and non-directory webpages. The html text content from these websites is then cleaned, scrapped, and saved in a directory for use later. The text content in these files is then tokenized and modified using the nltk library to be fit to create a model with. The modified terms are then passed through to create a bigram bag of words model, which is then saved to a file for use later classifying directory and non-directory pages. With the model created, a webpage can be entered into the cli, which will make a determination based on the model whether it thinks it is a directory or non-directory page.

3. Usage Documentation

All function of the program occurs through the cli. For the program to work python3 and its standard libraries are required alongside bs4, requests, and nltk. To run the program, the user should navigate to the repository and type `python3 main.py`. This will run the program at all of its default settings, utilizing the pre-built model and checking it against <http://www.google.com>, so the program will return ‘Classified as NOT a Directory Page’. If the program is run against a directory page such as `python3 main.py https://cs.illinois.edu/about/people/all-faculty` it should return ‘Classified as a Directory Page’. To view the other options the program has, the user can type in ` python3 main.py -h` which will return:

****

This shows all options of the program including those related to extracting the initial data and creating the model.  
  
4. Video Overview

https://www.youtube.com/watch?v=KabKW-5aXY8