**CS021E Final Project**

**Jordan Bourdeau**

**12/7/21**

**Summary:**

For the final project I designed a web scraper which takes a number of pages to scrape, a .txt file name, and a starting Wikipedia link as inputs. The program takes all the text from the paragraph elements on the Wikipedia page, makes them into a string, strips every sentence from within the submission file of whitespace and periods, exclamation points, and question marks then checks them against the string to see if there are any matches. Every time there is a match, the program will increment the variable containing the number of matches and add the matching sentence to a list. After performing this on the Wikipedia page, the program will then follow as many links from the references section of the page as given earlier from the user input and repeat the process. The program is designed to take a Wikipedia link because of how the page is structured, and would not be able to reliably follow links from other pages unless they followed the same structure so it will only follow links from the main. Every iteration of its run, the program will output the number of matches from a given page as well as a list of matched sentences which it will then add to a corresponding dictionary variable with keys “Matches” and “Sentences”. After the web scraper has finished running, the program will then determine how plagiarized the submission file is by comparing the number of matched sentences to the total number of sentences in the file. Depending on this value, the program will output different messages to the user along with a list of sentences which were flagged by the program to investigate.

**Necessary Installs:**

This program uses the scrapy and os libraries. os is a default python module so you will not need to install it if python is installed, however you will need to install scrapy so it runs on your computer.

Python install (if not already done) through python website or windows store (also must be added to the computer’s path!):

Scrapy install command (run in command prompt): pip install scrapy

**Scripts:**

For this project the main script to look at is wikipedia\_spiders.py located in the spiders subfolder. To run this script all you need to do is run wikipedia\_spiders.py through an IDE or terminal and follow the instructions for inputs using the testing files.

**Testing:**

To test this program, you will need to enter the number of pages for the scraper to scrape, the file name, and the link to use. For these sample papers the link is the python Wikipedia page.

Copy Link from Here: <https://en.wikipedia.org/wiki/Python_(programming_language)>

Here are some sample .txt documents prepared to test this program and they are given below:

* sample\_plagiarized\_paper.txt
  + Plagiarizes 4/20 sentences from Wikipedia page so gets a 10/10 score (20% plagiarized = 10). To test this one, you only need to enter 1 as input for the number of pages to scrape but can enter more. It does not flag sentences from other sites.
* sample\_clean\_paper.txt
  + Plagiarizes 0 sentences from anywhere and does not flag any sites regardless of how big you set the number of pages to scrape (the number of pages to scrape can only be as big as the length of the Wikipedia references section.
* sample\_external\_plagiarized\_paper.txt
  + 100% plagiarized from <https://julialang.org/blog/2012/02/why-we-created-julia/>
  + Set number of pages to around 50 to access this link with the scraper

References

“How to Check If a File Exists in Python.” *Python Tutorial - Master Python Programming For Beginners from Scratch*, Python Tutorial, 23 Oct. 2020, https://www.pythontutorial.net/python-basics/python-check-if-file-exists/.

“How to Run SCRAPY Spider FROM within a PYTHON SCRIPT.” *Youtube*, Code Monkey King, 16 Feb. 2020, https://www.youtube.com/watch?v=Z9tFRznPAS8&ab\_channel=CodeMonkeyKing. Accessed 4 Dec. 2021.

“Python Web Scraping and Crawling Using Scrapy.” *Youtube*, Buildwithpython, 17 Jan. 2019, https://www.youtube.com/playlist?list=PLhTjy8cBISEqkN-5Ku\_kXG4QW33sxQo0t. Accessed 27 Nov. 2021.

“Scrapy Tutorial¶.” *Scrapy Tutorial - Scrapy 2.5.1 Documentation*, Scrapy, 6 Oct. 2021, https://docs.scrapy.org/en/latest/intro/tutorial.html.