Jamel Chouarfia

Dr. Biplav

CSCE 240

5 February 2023

Project 1 Report

For the very first part of making the chatbot, I had to create a program that downloads and filters data from the disease website. The downloading part was something that I worried about, but it turned out to be easy. The filtering part, however, was the complete opposite. I’ve filtered strings before, but the strings extracted from the file were messy. Both pages’ code files had random lines that would just go on for over 10000 columns. Within these huge lines were multiple sentences that had to be separated from dozens of tags. This part took most of the time, as I had to deal with conflicting issues with both pages. Removing a restriction that deleted useful information from the CDC file resulted in the WebMD file gaining extra junk data. At the end of it, I had to make a compromise between lost data and junk data. I plan on fixing this bug before the actual chatbot is delivered.

As for how the code works, it uses two classes. The class MainExtractor is where the main method is located. It first creates SiteHandler, which is the second class, and attempts to use it to download html code. If successful, it sends a confirmation message. The first site downloaded is the CDC page about Avian Flu. The second site is the WebMD page about the same disease. After that, it uses SiteHandler to filter the downloaded data and place it in a new file. SiteHandler does this by using a helper method called lineTrim. lineTrim takes in a string from the unfiltered file and cleans it up into a more readable format. It does this by deleting everything in a string that is not in between a paragraph tag. Once all the data is filtered, it stores it into a new file. This is done for both the CDC and WebMD html files.