Joseph Prieto

CS 410 Text Information Systems

November 19, 2023

Project Progress Report

So far, I have begun my initial research on the project. I have searched for medical device websites with reviews to web scrape. I have also investigated existing web scrapping tools and have determined that I will have to write my own web scraping code using the beautifulsoup library in python if I want to have full control of the output format of the data scraped from the websites. I have begun writing the code and testing it on the websites chosen. After further research I will include the following attributes as part of my data set: star review out of 5, text review, medical device, source website, reviewer name, and date posted. The tasks that remain are finishing writing the code for the web scraper and adjusting it per website to retrieve the review data, cleaning and formatting the data, and then writing code to obtain statistics about the data set. I have run into a few challenges so far. The first one was not being able to find robust free web scraping tools, so I decided to write my own code. The code will be uploaded to the GitHub. The other challenge I faced was finding medical devices with bad reviews. This will lead to a skewed data set in favor of good reviews. Other than those challenges, the project is coming along nicely.