Natural Language Processing – CS 4420

Assignment 1

Shrey Chhabra

1.b. My scraper uses selenium, beautifulsoup4, and pandas. I opted to scrape Google, Bing, and MSN news articles to get a total of 150 articles (even though the requirement is 100, a lot of articles are just *stubs* which are more difficult to clean, which is I scraped several extra articles as well). The reason to use these three news aggregators is that since selenium allows me to use a headless webdriver (basically running chrome without the browser window actually opening), scraping the news articles is akin to browsing these on google news. This ensures that the articles that are scraped successively are the most recent possible articles, and I would not have to spend time setting up logic for the program to filter out older articles. After this, I only needed to specify the search queries (eg. “Ozempic”, “Ozempic FDA approval”) to direct the webdriver to scrape the articles with these matches. Finally, I used bs4 to parse the HTMLs and pandas to store the articles in a CSV.

2.d. For entity resolution, I opted to use fuzzy matching. It works best for this specific use case since different articles refer to the same entities in different ways. There are several examples that support this, for instance; “Novo Nordisk” (a Danish pharma company) is also often referred to in different ways, i.e., “Novo-Nordisk Ltd.”, “Novo Pharma”, etc. A lot of the scraped articles are written by entirely different people, from different countries and different writing styles. Fuzzy matching thus becomes the best method for entity resolution (well, specifically the Levenshtein Distance method of fuzzy matching), measuring string similarity, and is flexible, albeit slow for large datasets. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. If the similarity score is above a certain threshold (90% in my case), then the entities are considered to be a match.