Assignment 03 - Web Scraping and Text Mining

Due Date

March 17, 2022 by 11:59pm.

The learning objectives are to conduct data scraping and perform text mining.

APIs

- Using the NCBI API, look for papers that show up under the term "sars-cov-2 vaccine." Look for the data in the pubmed database, and then retrieve the details of the paper as shown in the lab. How many papers were you able to find?
- Using the list of pubmed ids you retrieved, download the details of each paper using the query parameter rettype = abstract. If you get more than 250 ids, just keep the first 250.
- As we did in the lab. Create a dataset containing the following:
    1. Pubmed ID number,
    2. Title of the paper,
    3. Name of the journal where it was published,
    4. Publication date, and
    5. Abstract of the paper (if any).

Text Mining

The pubmed.csv dataset https://github.com/JSC370/jsc370-2022/blob/main/data/text/pubmed.csv contains abstracts from articles across 5 search terms. Your job is to analyze these abstracts to find interesting insights.

1. Tokenize the abstracts and count the number of each token. Do you see anything interesting? Does removing stop words change what tokens appear as the most frequent? What are the 5 most common tokens for each search term after removing stopwords?
2. Tokenize the abstracts into bigrams. Find the 10 most common bigram and visualize them with ggplot2.
3. Calculate the TF-IDF value for each word-search term combination. (here you want the search term to be the "document") What are the 5 tokens from each search term with the highest TF-IDF value? How are the results different from the answers you got in question 1?