Assignment 03 - Web Scraping and Text Mining

September 10, 2021

Due Date

November 5, 2021 by midnight Pacific time.

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 trial vaccine." Look for the data in the pubmed database, and then retrieve the details of the paper as shown in lab 7. How many papers were you able to find?
- Using the list of pubmed ids you retrieved, download each papers' details using the query parameter rettype = abstract. If you get more than 250 ids, just keep the first 250.
- As we did in lab 7. 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

A new dataset has been added to the data science data repository https://github.com/USCbiostats/data-science-data/tree/master/03_pubmed. The dataset contains 3241 abstracts from articles across 5 search terms. Your job is to analyse 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?

PM566: Introduction to Health Data Science - PM 566 (Fall 2021)

University of Southern California

Department of Population and Public Health Sciences

George Vega Yon, Kim Siegmund, Abigail Horn

vegayon@usc.edu