20231218_Status

December 18th, 2023

Project Goals

- Use machine learning (supervised or unsupervised) to classify ASM papers for two variables
  - "containing new sequence data" or not
  - "having data available online" or not
- This is based on most common "grams" (words) from the paper, using a pre-determined set of test data hand coded for both variables
- Report statistics on number of citations/paper as a function of data availability

Finished Tasks

- Create “groundtruth.csv” file of N=480 papers coded for the two variables of interest as a training set
- Learn about R packages ‘rvest’ and ‘polite’ for querying servers to collect HTML text of websites

Current Status

- Working on text scraping for each paper and discussion of what to keep in text
- Book: Text Mining with R by Silge & Robinson
- Set up script for using Great Lakes cluster to scrape all N=480 texts

To Do Next

- Separate each word in each paper(?)/ come up with some way to make "grams"
- Calculate tf-idf statistics for each word in each paper
- Determine possible keywords for a supervised ML attempt
- Outcome of unsupervised ML attempt (mikropml?)