**Homework 6: Text Mining**

**CPLN 6710/MUSA 5000**

For HW 6, you are asked to do the following:

1. Obtain a text data set (some potential data sources are below) and convert it into a corpus. The corpus you analyze should have at least 50 entries.
2. \* For **extra credit**, summarize the entries using ChatGPT using the ChatGPT R Markdown.
3. Clean the corpus using the procedures outlined in the R Markdown (removing certain characters, lemmatization, etc.)
4. Create a Word Cloud of the terms that appear most frequently and discuss it.
5. Conduct a sentiment analysis using one or more lexicons of your choice.

Structure your report as your previous homework assignments:

1. In the Introduction (1-2 paragraphs), describe what you’re trying to look at and explain the relevance of the problem.
2. In the Methods, describe what you will be doing in this report. Be sure to describe any method you use in your own words. Every term that’s part of the text mining lingo should be explained as well (e.g., stemming, lemmatization, stop words, word clouds, sentiment analysis, lexicon, etc.)
3. In the Results, describe your results. For example, what are the findings of the data cleaning, word cloud, sentiment analysis?
4. In the Discussion, briefly (1-2 paragraphs) describe why these findings might be interesting and any future analyses you may want to conduct to gain further insight about your data.

Potential data sources are in the table below. Feel free to use ChatGPT to get suggestions for additional data sources.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Summary** | **Link** | **R Library Support** |
| Wikipedia Corpus | Full dump of Wikipedia articles, suitable for a wide range of NLP tasks. | [Wikimedia Downloads](https://dumps.wikimedia.org/) | No direct library, but data can be processed with **tm** or **text** packages. |
| Project Gutenberg | Large collection of books, primarily classics and historical texts. | [Project Gutenberg](https://www.gutenberg.org/) | **gutenbergr** for downloading and processing texts. |
| Reuters News Dataset | Collection of news stories, used in text classification and sentiment analysis. | [NLTK Library](https://www.nltk.org/nltk_data/) | Accessible via **tm** package's **getReuters21578** function. |
| IMDb Reviews | Dataset of movie reviews, used for sentiment analysis and text classification. | [Kaggle](https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews) | No specific package, but can be imported and processed. |
| Common Crawl | Massive collection of web pages, useful for various applications including web text analysis. | [Common Crawl](https://commoncrawl.org/) | No direct library, requires manual processing. |
| Amazon Product Reviews | Product reviews, often used for sentiment analysis and recommendation systems. | [Kaggle](https://www.kaggle.com/snap/amazon-fine-food-reviews) | No specific package, but can be imported and processed. |
| Stack Overflow Data Dump | User-generated content from Stack Overflow, useful for technical text analysis. | [Archive.org](https://archive.org/details/stackexchange) | No direct R library, but data can be processed using text mining packages. |
| Enron Email Dataset | Emails from the Enron corporation, used in various NLP tasks. | [Kaggle](https://www.kaggle.com/wcukierski/enron-email-dataset) | No specific package, but can be imported and processed. |
| Google Books Ngrams | Collection of word sequences from books scanned by Google, useful for linguistic research. | [Google Ngrams](https://books.google.com/ngrams) | No direct library, but **text** and **tm** packages can be used for analysis. |
| UCI ML Repository Text Datasets | Various datasets for text data mining and machine learning research. | [UCI Archive](https://archive.ics.uci.edu/ml/index.php) | No direct library, but data can be imported into R. |
| Blog Authorship Corpus | Large number of blog posts, suitable for authorship identification and analysis. | [UMass School of CS](https://www.cs.umass.edu/~mccallum/data.html) | No specific R package, requires manual processing. |
| Yelp Reviews Dataset | Business review dataset, used for sentiment analysis and recommendation systems. | [Yelp Dataset](https://www.yelp.com/dataset) | No specific package, but can be imported and processed in R. |
| Hansard Corpus | Edited transcripts of debates in the British Parliament, useful for political and historical analysis. | [UK Parliament Hansard](https://hansard.parliament.uk/) | No direct R library, but data can be processed using text analysis packages. |
| Brown Corpus | One of the earliest corpora in computational linguistics, containing diverse text sources. | [NLTK Data Download](https://www.nltk.org/nltk_data/) | Available in **corpora** package in R. |