Assignment 1.1: Text Preprocessing using NLP Techniques

Patricia Enrique

Department of Engineering, University of San Diego

AAI-511: Neural Networks and Learning

Prof. Mokhtari

September 11, 2023

Introduction

Preprocessing text data is an essential step in developing a natural language processing (NLP) model. The dataset used for this analysis is compiled by Standford (Maas et al., 2011) and is comprised of 50000 movie reviews with a binary sentiment. The distribution of positive and negative sentiment reviews is equal with 25000 for each classification. The reviews are of varying lengths and can contain emojis, breaks, and spelling errors. For the purpose of illustrating tokenization, stemming, lemmatization, and stop words, five shorter reviews are selected and summarized in Table 1.

Table 1

Movie review examples

ID	Review	Sentiment
415	A rating of "1" does not begin to express how dull, depressing and relentlessly bad this movie is.	negative
9437	Hated it with all my being. Worst movie ever. Mentally- scarred. Help me. It was that bad.TRUST ME!!!	negative
1419	For pure gothic vampire cheese nothing can compare to the Subspecies films. I highly recommend each and every one of them.	positive
9746	A great film in its genre, the direction, acting, most especially the casting of the film makes it even more powerful. A must see.	positive
21044	Brilliant. Ranks along with Citizen Kane, The Matrix and Godfathers. Must see, at least for basset in her early days. Watch it.	positive

Process

The preprocessing techniques for the movie reviews are completed using Spacy and NLTK libraries. When using Spacy, the trained language pipeline for English is loaded containing the components and data to process text. Tokenization, lemmatization, and stop words are implemented using Spacy; however, this library does not contain a function for stemming, so this technique is implemented using NLTK.

The process for tokenization, lemmatization, and stop words are similar. A single list is created with the five reviews previously selected and is iterated through to process one review at a time. Each review is passed into the trained pipeline and the tokens are extracted from the text. For tokenization, the tokens are printed. For lemmatization, each token is reduced to its dictionary form before being printed. For stop words, the tokens that provide little contribution to the meaning of the text such as propositions, conjunctions, or pronouns are extracted and printed (Mokhtari, 2023).

To apply stemming to the text, a PorterStemmer() class is created using the NLTK library. The tokens extracted using the Spacy library are then iterated through and the porter stemmer class is used to find the root of the words.

Analysis

The importance of applying preprocessing techniques to text becomes apparent when viewing the most frequent words in the movie reviews. To illustrate this, Figure 1 summarizes the top 20 words in the first 1000 movie reviews from the original dataset. It can be seen that the list is majorly comprised of stop words and punctuation. For most applications, this result is not beneficial and provides no valuable insight on the text being analyzed. In comparison, Figure 2 summarizes the results after removing stop words and punctuation. From this plot more valuable information is extracted such as there are more references of the word "good" and "like" than "bad" implying there may be more positive movie reviews than negative. It is important to note that these results can be further improved by additionally applying lemmatization. From Figure 2 it can be seen that "movie" and "movies" are listed separately as are "characters" and "character" which both represent the same word. Similarly, "good" and "great" express the same sentiment which is captured by applying lemmatization.

Figure 1

Top 20 words in the review without removing stop words

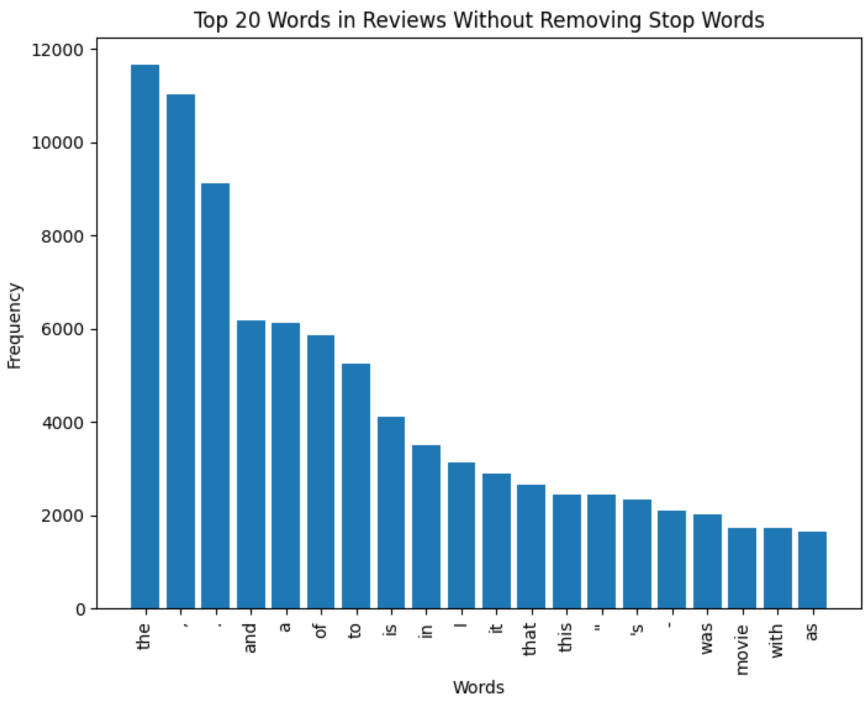
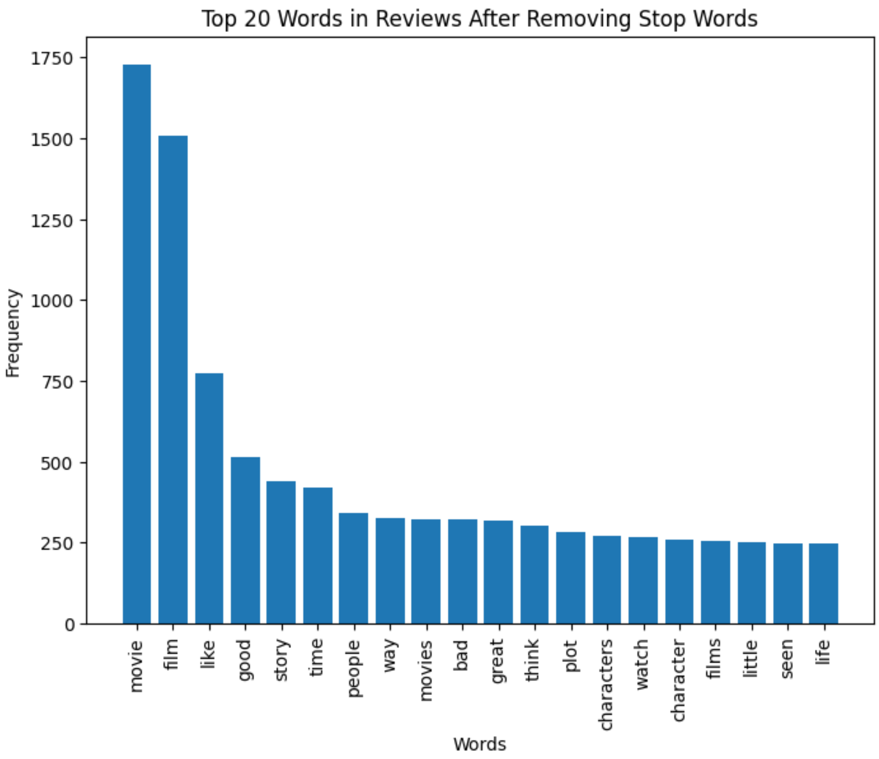


Figure 2

Top 20 words in the review after removing stop words



Conclusion

Text preprocessing improves the performance of NLP models by making the raw data a more manageable size and standardizing the text across multiple sources. The process removes unwanted or unimportant information which may affect the accuracy of the model results and makes them more difficult to understand.

A few challenges that appear when handling text data are spelling errors and shorthand typing or slang. These can be prevalent issues in search engines, chatbots, and social media which can hinder the language processing and understanding of the text. These common problems do not have a simple solution; however, Microsoft released a REST API for potential spell checking in Python which can be helpful (Vajjala et al., 2020). The processing of slang and shorthand in NLP models can benefit from the model's ability to understand context such as in a language translation application.

From this analysis it can be seen that text preprocessing techniques are beneficial in providing a sentiment analysis of movie reviews. Applying stop word removal and lemmatization provides a clearer summary of the words appearing in the reviews and thus if more positive or negative words are used. Another application is document retrieval based on user input. By removing stop words and applying lemmatization to the documents and the user input, a better match between the two can be made.

References

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Large Movie Review Dataset* (Version 1.0) [Data set]. Stanford University. http://ai.stanford.edu/~amaas/data/sentiment/

Mokhtari, K. (2023, September). *Natural Language Processing Basics* [Lecture notes]. University of San Diego. https://sandiego.instructure.com/courses/829/pages/presentation-1-dot-1-natural-language-processing-basics?module_item_id=65362

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Practical natural language processing: A comprehensive guide to building real-world NLP systems. O'Reilly Media.