Web Application for Text Mining Analysis Using R Shiny

26.06.2023

Name: Petar Todorovski

Study program: Computer Science  
Mentor: dr. Uroš Godnov

The goal of this thesis is to create a web application that performs an in-depth analysis of text documents uploaded by users with the intention of providing users with valuable insights into their documents. The application will be developed using the R Shiny framework and will be hosted on the Shiny Server.

The web application will have a homepage that briefly explains how it works and what it does, and it will be optimized for use on desktop devices. Users will be able to upload English-language documents ranging in length from a single page to a novel. The analysis of the document performed by the application will include word frequencies and text classification analysis such as sentiment analysis and topic modeling.

Users will also have the ability to customize the analysis by selecting options such as the number of words to include in the word cloud or the type of graph to use for displaying summaries.

The programming will incorporate stemming and lemmatization as a default option it will automatically remove noise such as punctuation and stop words. To clean and preprocess the text data, the pre-built libraries in R such as `tm` or `stringr` will be used. Alternatively, machine learning algorithms may be used to accomplish these tasks.

The text classification analysis will be performed to determine the overall mood or tone of the text. This will be achieved using a pre-built library in R such as `tidytext` or `sentimentr`, which will identify positive, negative, and neutral sentiments in the text. Moreover, if implemented, the programming will also allow for multiple text documents to be uploaded at once.

In the event that the uploaded file contains numerous misspellings and inconsistencies, the user will be informed to double-check the document and attempt uploading it again. The analysis will cover all words on the stop word list, but any rare words with low frequency will be included in the output as rare words.

**LIST OF CONTENTS**

1 INTRODUCTION

1.1 Text Mining Analysis

1.1.1 Components of Text Mining Analysis

1.2 R programming   
1.2.1 R packages   
1.2.2 R Shiny

2 METHODS

2.1 Word frequency

2.2 Text Classification  
2.2.1 Sentiment analysis  
2.2.2 Topic Modeling

3 RESULTS

4 CONCLUSION

5 REFERENCES

**REFERENCES**

[1] *What is Text Mining* <https://monkeylearn.com/text-mining/>

[2] *Text Mining with R*. <https://www.tidytextmining.com/index.html>

[3] shi*ny: Web Application Framework for R*.   
<https://cran.r-project.org/web/packages/shiny/index.html>