ALY 6040  
Module 5: Text mining

Professor: Justin Grosz

Chongfan Bao CPS in Informatics

Northeastern University

06/26/2021

**Abstract**

For this project, I will implement text mining to get an insight about text. During this procedure, I will omit unnecessary punctuation, number, and common English letter such as “the”, “at”, “I”. This can help me understand the content and get wordcloud image more precisely. Based on wordcloud and word frequency histogram, I will explain the meaning of the char and think about how I can associate this kind of technique with business, what information the enterprise can get.

*Keywords:* text mining, business.

**Introduction**

The first four weeks, we almost deal with the structured data. However, in the complicate world, there are many unstructured data included emails, social media posts, comments, reviews, subjective survey results, new articles and other human-written text we need to explore. This also help us get insight about content and corporate with our business questions, which is involved in mining text data. By this technique, we are able to extract relevant data through the use of text mining and achieve our business needs.

# Text overview and modeling

This text is about “dream” presentation retrieved from the Internet. The article is English and there are totally 46 paragraphs in the RStudio. Also, as a result of this information is unstructured and in text format, we need to a process that transforms unstructured information into structured data that can be analyzed in a traditional way. Therefore, I will leverage text mining to convert the unstructured data into structured data. In addition, I am about to extract the letters that are frequently used, which is apart from pronoun, preposition, number and so on. This is because these kinds of words often show in the articles. When we draw a wordcloud chart and words frequency histogram, it will cause huge bias for our outcomes. Finally, we can analye our chart more accurately and align with our business questions.

**Data Analysis**

From figure 1, we can observe that the top 10 words showed in the contents. The first place is “will” that totally presented 17 times. Then, the “freedom” is mentioned 13 times follow by the “ring” and “dream”. There are 6 words show more than 10 times. Based on these 10 words, we can guess that this paper is talking about dream regarding freedom if we don’t read the article firstly. In the figure 2, we display the 200 words showed mostly in the article. We can see the most obvious and frequent word located in the center of image. Furthermore, the size of word will shrink by their frequency in the article. The color contributes to our audiences to figure out the importance of word and latent meanings as well. We also rely on outside green words to get more information regarding the article. From figure 3, we can find words that occur at least four times in the article by findFreqTerms function. There are 15 words, as dream, day, nation, one and so forth. Moreover, we also can check which words are associated with freedom in the term document matrix we create. We notice that there are 8 words associated with freedom over 0.3 correlation circumstance. They are let, ring, Mississippi, stone, mountainside, state, every, and mountain. Finally, we can intuitively know in the figure 4 the most frequent words in the text. People can be more easily know what the article is.

**Business Insight**

After we discussion so much. We want to know what we can do or what the company can do with text mining. As far as I am concerned, text mining is a good support tool for organizations because it can understand information more deeply, understand and identify relevant business insights from the content, and highlight the connection between the information in one or more texts. If we use traditional tools or search engines cannot find this useful information.

First of all, we guess we have the same experience. When we search Google or Amazon, we just enter the server word, and the search engine will display relevant information or products. This is because the company can broadly match customers’ search criteria and then show your ads to users. This means that when customers search with the words they set, their products will be displayed. When users search for words related to keywords, advertisements will also be displayed, which may include synonyms, misspelled words, variant words, plural words, word order, etc.

Secondly, we also leverage the text mining to implement sentiment analysis. We collect the customers’ feedbacks, reviews and comments. Then, we use the sentiment score to evaluate if the text is positive, negative, happy, sad, or neutral. Given this, the business or agency can get an idea of how users feel about a particular product. Based on these sentiment analyses, the company can fast make a response to their customer and merchandises. For instance, If the text mining find that the customers aren’t satisfied with logistics. The enterprise needs to promote the speed of logistics or ensure the safety of product while in transit.

**Conclusion**

Text mining ensures that all available information is used, enabling us to make more informed decisions, automate information-intensive processes, gather key business insights, and reduce operational risk. If used properly, text mining can address the high value of many different application areas Knowledge discovery issues, including R&D using sentiment analysis and social media mining, competitive intelligence, patent analysis, and market research.

**References**

Expert.ai Team, (2016). The value of text mining for unstructured information. Retrieved from https://www.expert.ai/blog/value-text-mining-unstructured-information/

Text Mining and Wordcloud Fundamentals with R Code. Retrieved from https://northeastern.instructure.com/courses/68030/pages/lesson-4-3?module\_item\_id=5788070

Appendix A

**Figure1:** *Wescenting Words frequency*

A picture containing graphical user interface

Description automatically generated

**Figure2:** *Wordcloud*

*Text

Description automatically generated*

**Figure3:** *word frequency and relation*

*Graphical user interface, application

Description automatically generated*

**Figure4:** *Frequent words histogram*

*Chart, bar chart

Description automatically generated*

Appendix B

R code

A picture containing text

Description automatically generatedGraphical user interface, text, application

Description automatically generated