Advances in Aeronautical Science and Engineering
ISSN: 1674-8190

Developing Novel NLP-Based Text Mining Techniques for Enhanced Information Extraction

Thandar Min
MS Student in Data Science, New York University

Dr. Sarah Patel
Professor of Data Science, New York University

ABSTRACT

The use of text mining in understanding the documents, and the novels can provide insights in supporting or making decisions. The purpose of this paper is to analyze an epic high fantasy novel "The Lord of the Rings", one of the famous novels written by the writer J.R.R. Tolkien in search for how the novel is delivered to the audiences such as the way the author expressed and portrayed the characters, tone, style, and the choice of words in artistic ways that it has reached up 100 million copies of books being sold, and impact to them. This paper will explain the different analyses performed on the Lord of the Rings novel using text mining and explain the results of it. Frequent words analyses show the words most commonly used by the authors in delivering the novel. In other word, this shows the tone, type, and the choices of words. The sentiments analyses describe how each book in the novel gives different sentiments and what are the overall sentiments the author delivered. With different types of sentiment analyses and different lexicons, sentiments other than dualism (positive or negative) such as trust, anticipation, and disgust can be discovered in analyzing the novel. Therefore, sentiment analyses result not only the sentiments the audience will perceive but also depict the tone and style used in the novel. The bigram analysis presents word relationships, and the use of the tf-idf method finds the important words by weighting words. Overall, this paper explains text mining with NLP techniques on the Lord of the Rings novel, and the results of it.

INTRODUCTION: THE SIGNIFICANCE OF NOVELS

Novels can deliver messages in terms of moral aspect. They teach society rules, give emotional feeling towards characters, and have effects on the large population. In addition, they organize us culturally and emotionally as well as enrich us intellectually by giving more wisdom, meaningful and manageable life, and expands our moral capacity which in turn benefit in how we view the world, life, and also help us in making decision. This can lead to improving how we communicate with others and how we impact and contribute to the society. While communication can bring peace and even prevent from wars, the contribution to the society can bring the development of the society as a whole. Because of these facts above, how the novels are being delivered, analyzing how the texts are being structured, expressed and delivered in the context of the novels which have reached to the large population in that way that has resulted them enjoyment or sadness toward characters along with shaping the way we view the life or the world is crucial.

LITERATURE REVIEWS ON NLP TECHNIQUES

Why frequent words are the way they are?

Frequents words are usually not the most meaningful words. The top five most frequent words are ‘the’, ‘of’, ‘and’, ‘a’ and ‘to’. Moreover, frequent words follow the Zipf’s Law and the principle of Least effort. According to the Zipf’s Law, the distribution of words can be explained through a formula that reveals that the frequency of a word through the statistical analysis of a collection of writing – corpus. It states that the frequency that a word appears is inversely proportional to its rank with the equation f * r = k. We can find the Zipf’s Law in many areas such as wealth of individuals, popularity of book, web-pages, and consumer products. The

principle of least efforts states that people, and even well-designed machines will naturally choose the path of least resistance or “effort”. Because of that two principles, the top most common five words are function words, and therefore when applying NLP techniques to find out frequent words, we take out those words as stop-words to find the common frequent meaningful words, and apply techniques such as tf-idf to weight the words in finding important words.

The evolution of the sentiment analysis

According to the research article “The evolution of Sentiment Analysis – a review of research topics, venues and top cited papers”, the roots of the sentiment analysis begins around 1900s in the text subjectivity analysis and public opinion analysis at the beginning of 2000s. Starting during WWII, public opinions are being wondered more by the academia even through the general desire was for political purposes. As the numbers of text on the web got increased, the numbers of analysis done on the sentiment analysis increased as well. (Mäntylä, et al.)

The application domains of sentiment analysis

Currently, sentiment analysis is used for variety of areas as mentioned in the article which aim to explain evolution of sentiment analysis. The author classified the field that sentiment analysis is focused into two classes based on the type of goals. The two classes are application domain, and Human and behavior domain. Application domain, which can also be called business domain is divided into six classes. They are society, security, travel, finance, corporate, medical and entertainment. Human and Behavior oriented is the several application domains with the focus on the research. The six domains are Expertise and Influence, Interaction, Globe, Truth, Language, Behavior and Emotions.

The two main approaches of sentiment analysis

Lexicon-Based Methods for Sentiment Analysis article presents the idea that there are two main approaches of sentiment analysis. The first is lexicon-based approach, which involves calculating orientation for a document from the semantic orientation of words and phrases in the document. The second approach is building classifiers from labeled instances of texts or sentences also called statistical and machine learning approach. (Taboada, et al.)

RESEARCH METHODOLOGY

In analyzing the text, store the text data, perform text mining, create visualization and user interface, R devtools, R markdown, plotly, Rshiny were used. Data cleaning methods such as stop-words removal, tokenizing, lowering the capital letters to small ones and so on are also used in order to get meaningful and correct results from this text mining. Many packages available in R-CRAN assisted in completing this project. Fews are deplyr, stringr, tibble, tm, RColorBrewer, and so on. Using R markdown and plotly, frequent words, sentiment analysis, word relationship, and long short-term memory for text generation is performed on this project. In the former three, frequent words can find which words helped the author pieced the six novels, sentiment analysis to see what emotional words are used to deliver the novels, in which specific part of the novel the author delivered and bigram analysis are performed to see the relationship between the significant words. The latter one can help in improving text generation entertainment purpose such as creating poet with J.R.R. Tolkien style.

In performing word frequency analysis, word frequency on each chapter, on each book, on each book with tf-idf, on each book with overall novel comparisons, and word frequency on combined all books are performed. Similar to word frequency analysis, sentiment analysis on

each book, sentiment analysis in the course of narrative, sentiment analysis using different lexicons, and negative and positive sentiments are figured out as well. To understand word relationship in the novel, bigram analysis and bigram analysis with tf-idf technique are used.

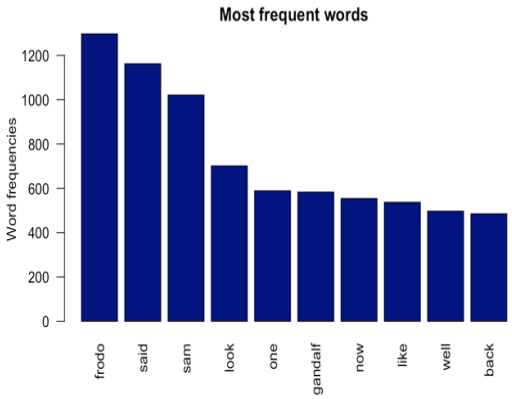
DATA COLLECTION AND STORAGE/WAREHOUSING

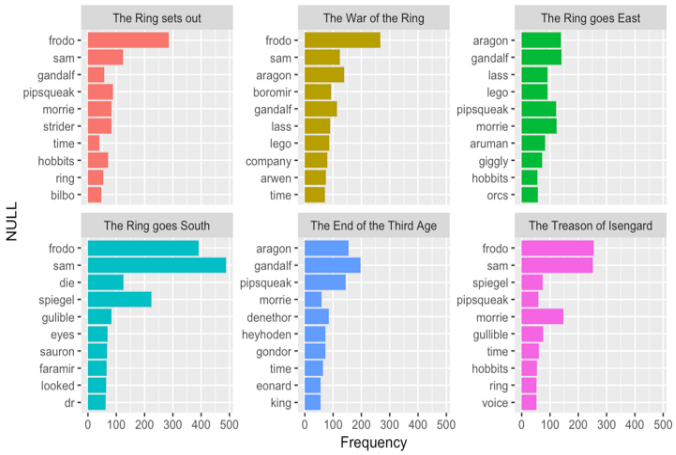
R data package is developed for easier reuse and as a contribution to analytic community as it will be able to utilize to practice and understand text mining using the Lord of the Rings dataset to find useful knowledge from the dataset. The six books of the Lord of the Rings are web scraped from an online website and packaged into the R CRAN (Inspired from Janeaustenr and Bradley Boehmke). A package is a convention for organizing files into directories. There are seven common parts of an R package. They are description file, which describes the work and sets up how the package will work with other packages and applies a copyright. The test file stores test that will alert you if your code breaks, man file contains the documentation for functions and the help pages in your package, vignettes teach how to use the tools to solve the real world problems(the package that I created doesn't include this), data to add data and namespace to organize the file . These components make the package self-contained and won't interfere with other packages. It is created as a local source package to share currently.

TEXT CLEANING/ PREPARATION AND PROCESSING/ DATA ANALYSIS

In search for finding which sentiments the author portrayed in the novel, the most common word frequency is first explored. To find the most common words the author used in the text and in each book, vector corpus is created, then the text is transformed into lower. Pronouns and articles such as "and", "a", and "the" are removed as most documents contain

words that can result the analysis's results got biased. Due to this concern, removing stop-words, misspelling, punctuation, whitespace, numbers and slangs is a necessary step in order to improve the performance of analysis. Then, the DocumentTermMatrix is created, and the top ten frequent words are sorted in the decreasing order. The two graphs below describe the common words that are present in all books and each book.





For sentiment analysis, the words that are cleaned are split and one of the lexicons based on all three unigrams lexicons are used to figure out sentiment analysis. When applying the lexicons, the lexicons are right joined, and the sentiments are grouped and summarized. The results are the spitted counts of sentiments into ten categories. Then, the six books are further analyzed to see on which part of each book, the most negative words or positive words are used using the index. The index can describe the amount of positive and negative words in the course of narrative.

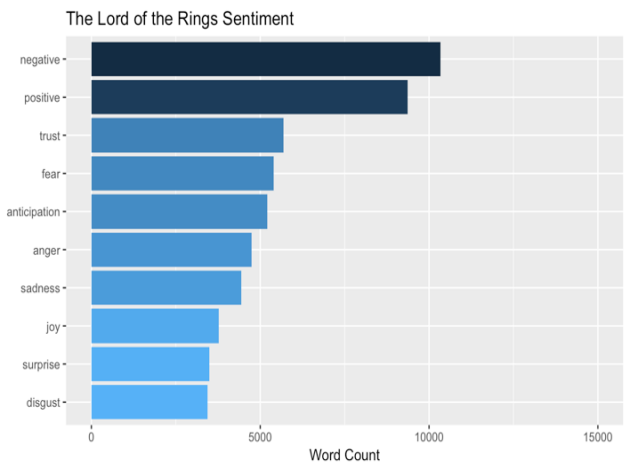


Figure 3

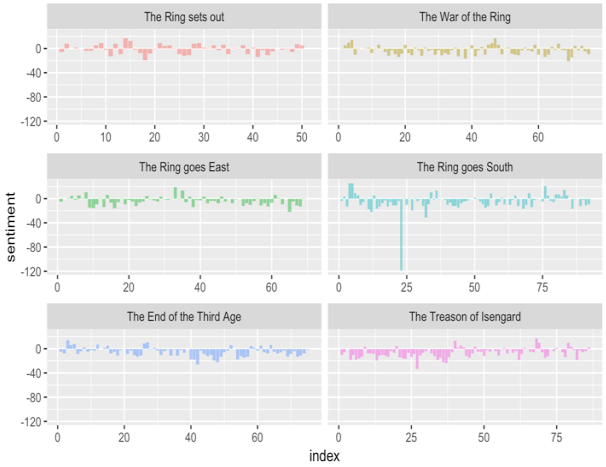


Figure 4

Moreover, to understand the degree of positive and negative words used in all six books, 3 different lexicon indexes are used and to understand how the scores are different based on the lexicon dictionaries. The three different lexicon dictionaries are AFFIN, bing, and nrc. As mentioned in the book "Text Mining with R," AFFIN from Finn Arup Nielsen assigns words into score between -5 to 5, bing lexicons categorize into positive and negative and nrc categorizes into yes or no.

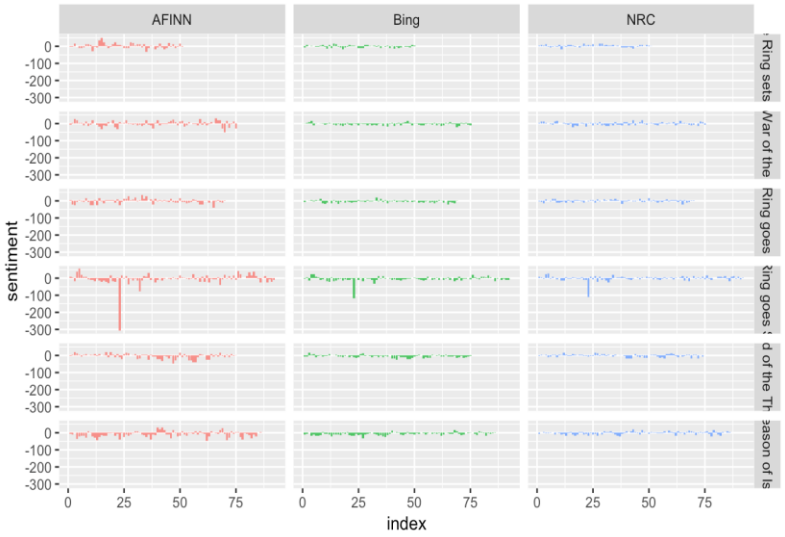


Figure 5

Using the bing lexicons, the top 10 positive and negative words used by J.R.R. Tolkien are also explored. The two bar-graphs below presents the findings.

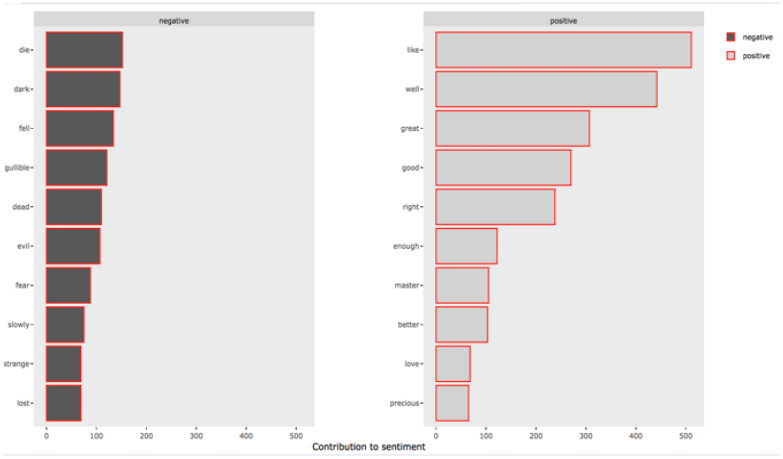


Figure 6

Moreover, the similarity of frequent words on all books versus on each book are compared. In this graph below, words that are located near the middle lines have similar frequencies than those that are located far apart (Silge, and Robinson). The deviation of words on each book to all book can be observed from this graph together with their correlation scores.

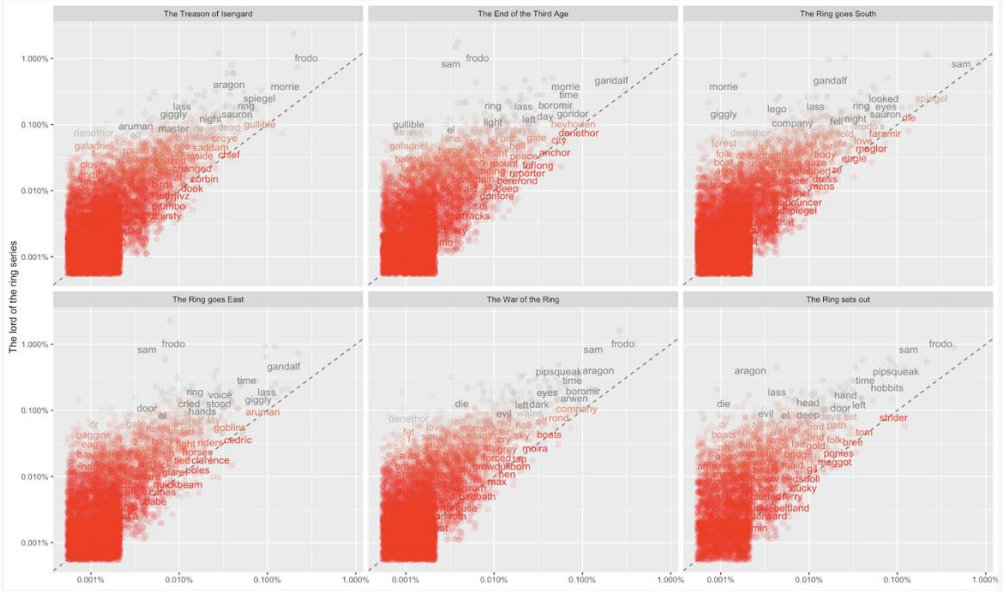


Figure 7

book <fctr>	correlation <dbl>
The Treason of Isengard	0.9043463
The End of the Third Age	0.6189687
The Ring goes South	0.8288405
The Ring goes East	0.6460124
The War of the Ring	0.8946792
The Ring sets out	0.8550044

6 rows

Figure 8

Using a term frequency (tf), a measure of how important a word is to the document is also used to determine the frequency a word occurs in a document combined with another measure inverse document frequency (idf), which increases the weight based on decreased frequency of usage and decreases the weight for increased frequency of usage. Tf-idf can measure how important a word is to a document in a collection of documents. This method selects out the important words which are not very common using the bind_tf_idf function, and the high tf-df words of each book are as seen in the picture below.

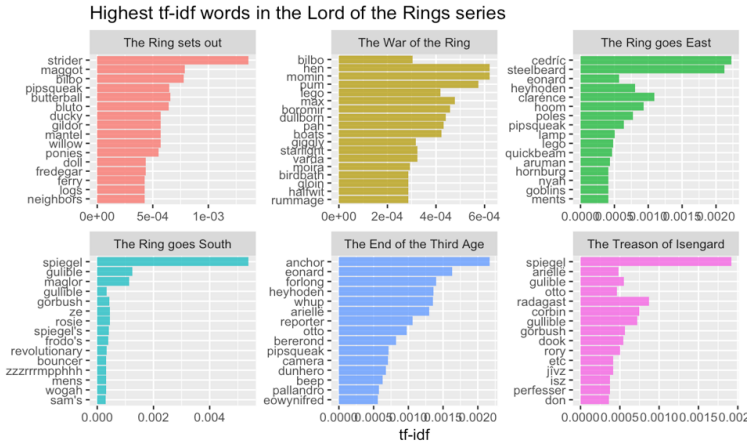


Figure 9

The ngram analysis is also performed in order to figure out the relationships between words that J.R.R. Tolkien used. The unnest_tokens function is used to figure out the ngrams of the words by counting and filtering the n-grams. However, the common words are “of the” and “to be”, which are stop-words and therefore they are filtered out using the filter and separating

the two words. After the filter function is performed, the words are united back to find the most common bigrams that doesn't contain stop-words.

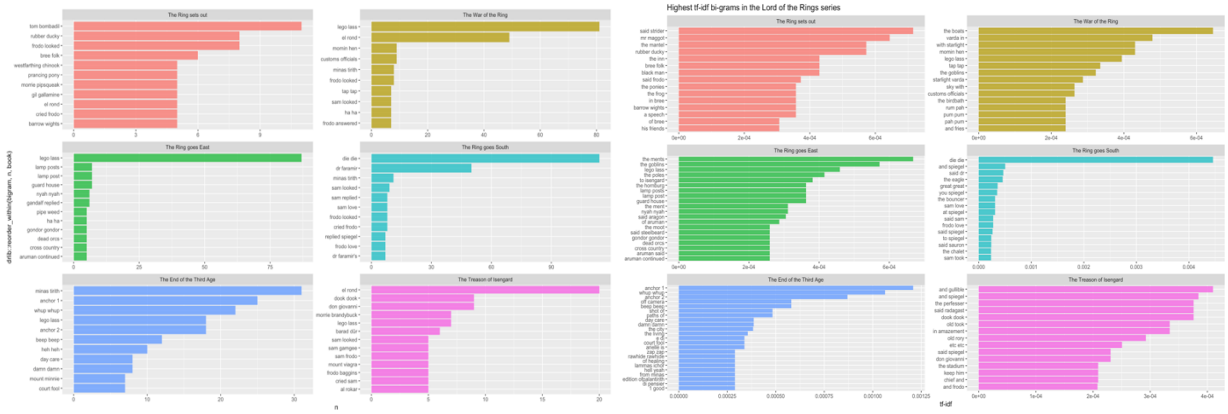


Figure 10

Figure 11

The most common sentiment-associated word that follow "not" to see which positive words are used as a negative in the all books is also analyzed.

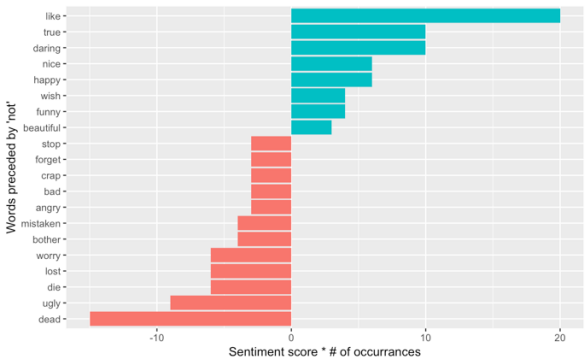


Figure 12

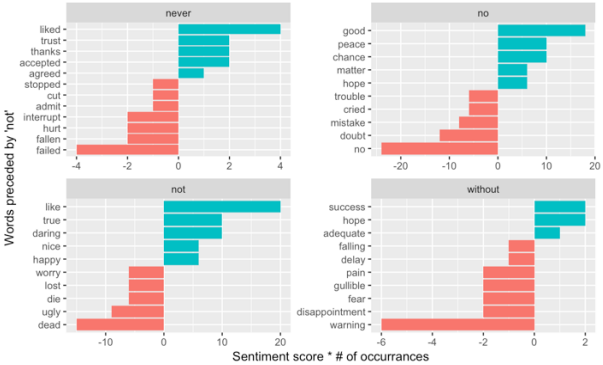


Figure 13

Ggprah is used to visualize bigrams with networks using from, to and weight. From - the node an edge is coming from, to - the node an edge is going towards, and weight - a numeric value associated with each edge are the variables used to construct them for the bigrams.

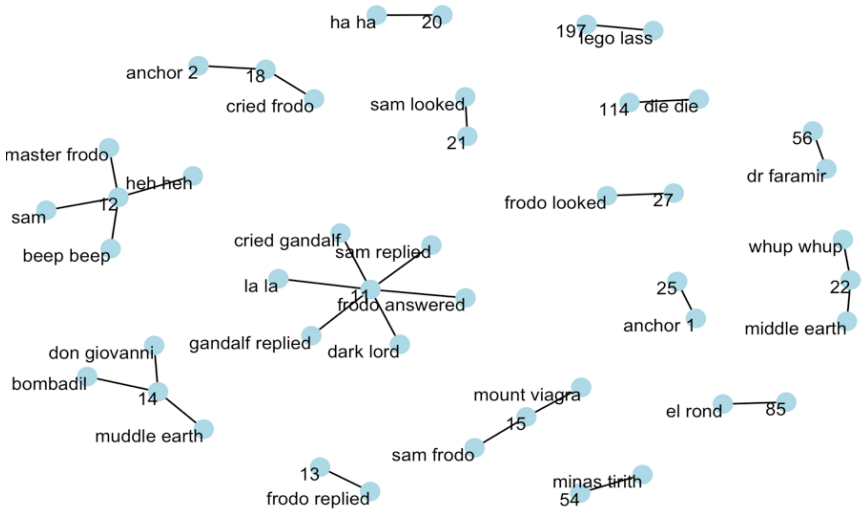


Figure 14

Long short-term memory neural network is also applied on the first book text dataset of for new text generation. After 5 epochs and with the batch-size of 128, the results had improved to a better level. Kera model with rmsprop optimizer is used in implementing this. This implementation can be improved to generate a new text similar to J.R.R. Tolkien's style of writing. Please see below for detail improvements after 5 epochs (Chollet, and Allaire).

First Epoch

```
Epoch 1/5
49408/49452 [============================>.] - ETA: 0s - loss: 2.4229epoch: 00 -------
--------

diversity: 0.200000 ---------------
 and bart and and a the sant of the sant of the the sand a stant and the santer and
the sant the sand the sand and and and the banted and and stant and the bere and the
sand the want and hant and a stare and the want stant and the sare the berter and the
barted and the banted and the mant the bant and and hant the sant a stant and and the
banter and the ore the bant the brout a the the want and th
```

Fifth Epoch

```
Epoch 5/5
49408/49452 [============================>.] - ETA: 0s - loss: 1.6282epoch: 04 -------
--------

diversity: 1.200000 ---------------

49452/49452 [==============================] - 206s 4ms/step - loss: 1.6286 frodo came
withfeek ol in any wawt swinced. "they-blowejulld, ondily set ackhen pleciored. "ohere
was stopped hoblick! or,elles,- wish a dallimetheh whila, juntow off-ey. urry seead to
```

```
care on at hispark.        that," he bree--your al s lucking enfeerchibuck, i's dooks
belherllan whosh. it all it.      "houried triecrforsonert
```

KEY FINDINGS

From the word frequency analysis, the top ten common words are found and the analysis result of six books describe the common words find in each book and from these common words, Frodo and Sam include the most part throughout the six books. In other word, these characters are the main characters of the novel. (Figure 1 and 2)

Sentiment analysis results show that the author use the words related to “trust”, and the second most used sentiment words are related to “fear”. The words that can represents the sentiment “disgust” is the least used throughout the novel. Moreover, the negative words overall are used over the course of all the novels than positive words. The top three most common negative words were die, dark, fell, and positive words were like, well, great. The results of sentiment analysis with the index shows that negative words and positive words usage throughout the narrative of each book. Negative words usage occurs most with the number of negative words count of 119 at the index 23, which is in the first quarter of the book “The Ring goes South.” The different lexicon usage of the six books show different number negative words usage, and findings suggest that the AFFIN lexicon can pick out the negative words and positive word counts more than Bing and NRC lexicons. (Figure 3, 4, 5, 6)

After searching for the most common words and sentiments pertains in each book, all six books’ word frequencies are compared with the word frequency of each book in search for frequent words similarity. The words that line close to the diagonal line in these plots have similar frequencies for both set of the text. The common words find in both all text and “The End of the third Age” would be "grandalf" and "gondor". The book: “The Treason of Isengard”, has the highest overall similarity compared to the Lord of the Ring series. (Figure 7, 8)

The highest tf-idf words in the Lord of the Rings series is “strider” in Book I, the Ring sets out. However, the most common word is a different word without tf-idf. (Figure 9)

Analysis results of bigrams show that there is a difference in bigrams without tf-idf approach to with high tf-idf in the Lord of the Rings series. The most common bigrams found in the first book of the Lord of the Rings is “tom bombadil” while the most common bigram with high tf-idf is “said strider.” (Figure 10, 11)

The analysis results that most common word preceded by the word "not" with the highest negative sentiment scores is “dead”, words with the neutral sentiment score is “stop” and “beautiful”, and word with the highest positive sentiment score is “like”. The most common words that follow a particular negation words are also text mined and the results are visualized as below. (Figure 12, 13)

Network analysis performed using bigram_graph with the occurrence of bigrams being more than 15 time shows that words such as Sam, looked, Frodo and cried are also highly related words that are seen most often together throughout the novel. (Figure 14)

RECOMMENDATION

Further analysis of text mining can be done on this dataset. For example- Name entity recognition would be able to extract the name, location and so on from the dataset. Text similarly such as cosine similarity can also be performed in order to figure out which documents are similar the most. A chatbot that reply with the tone and writing style of J.R.R. Tolkien can be created for entertainment purpose as well. This project’ result of sentiments can be compared with the reviewers’ comments toward the novel, and also see how each sentiment within the

index in the course of the narrative help in determining the overall sentiments toward the novel and the reviewers' comment.

CONCLUSION

Overall, NLP techniques such as finding word frequency, sentiments and word choices are used analyze how the author delivered the content of the novel is analyzed in this paper. From these findings, we can conclude that J.R.R. Tolkien used negative words more than positive words throughout in portraying the novel about the Lord of the Rings. Frodo and Sam are the main characters and their parts of sadness are more than happiness in the journey of destroying the ring to end the war because of the hardship and fear that they face while attaining the goal of destroying this ring. The sentiment "trust" is used as they all together try to achieve the common goal of destroying the ring. Using the frequent words, sentiments and word relationship explained above, the author portrayed the characters and delivered to the readers in effective ways that the readers can understand how the war is undesirable and the desire that people want to destroy the darkness from the world.

ISSN: 1674-8190

BIOGRAPHY

Kyi Win, M.S. is a graduate student in Data Science Program at George Washington University. She enjoys researching, travelling, driving and taking photos. She also has a huge interest in earning a lot of money quickly, eating and sleeping.

Nima Zahadat, Ph.D. is a professor of Data Science and Information Systems Security. He has also held positions as Chief Security Officer, Chief Information Officer, Director of security, Director of Training Solutions, Dean of Computer Science, Program Chair of Information Systems, and Director of Operations. Dr. Zahadat has worked extensively with public and private sectors throughout the years. Dr. Zahadat has taught at the George Washington University and George Mason University in the fields of data science, information systems, web development, and security. He has developed and taught over 100 different curricula throughout his career. He has an undergraduate degree in Mathematics from George Mason University, a graduate degree in information systems and a Ph.D. in Systems Engineering and Engineering Management from the George Washington University. Dr. Zahadat's research interests are data science, data mining, information visualization, mobile security, information security, digital forensic, and risk management.

All the files and paper are available in https://github.com/kyiwin/TheLordoftheRings
Package summary https://kyiwin.github.io/TheLordoftheRings
And Zenodo link is https://zenodo.org/record/2654983#.XMi2dy-ZNQI.

REFERENCES

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019, April 23). Text Classification Algorithms: A Survey. Retrieved April 30, 2019, from https://www.mdpi.com/2078-2489/10/4/150

Wang, et al. "Learning Natural Language Inference with LSTM." *SAO/NASA ADS: ADS Home Page*, 1 Dec. 2015, adsabs.harvard.edu/abs/2015arXiv151208849W.

Taboada, et al. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics*, www.academia.edu/14508040/Lexicon-Based_Methods_for_Sentiment_Analysis.

Silge, J., & Robinson, D. (2019, March 23). Text Mining with R. Retrieved April 30, 2019, from https://www.tidytextmining.com

Chollet, François, and J. J. Allaire. *Deep Learning with R*. Manning Publications Co., 2018.

Bradleyboehmke. (2018, March 16). Bradleyboehmke/R-Training-Package-Development. Retrieved April 30, 2019, from https://github.com/bradleyboehmke/R-Training-Package-Development

Bradleyboehmke. (2016, December 30). Bradleyboehmke/harrypotter. Retrieved April 30, 2019, from https://github.com/bradleyboehmke/harrypotterInsert in-text citation Edit Delete

Juliasilge. (2018, May 29). Juliasilge/janeaustenr. Retrieved April 30, 2019, from https://github.com/juliasilge/janeaustenr

George Zipf: Human Behavior and the Principle of Least Effort. (n.d.). Retrieved July 24, 2015, from http://csiss.org/classics/content/99

IGF 2006 – 2011: The most frequently used words. (n.d.). Retrieved July 24, 2015,

from http://www.diplomacy.edu/IGFLanguage/most_frequent_words_background

Piantadosi, S. (2015, June 2). Zipf's word frequency law in natural language: A critical review and future directions. Retrieved July 27, 2015, from http://colala.bcs.rochester.edu/papers/piantadosi2014zipfs.pdf