RELEVANCE OF HASHTAGS IN KEYWORD EXTRACTION OF TWEETS

BY
ANJANA KUMMARI - 1782634
HOANG NGUYEN - 1838540

TABLE OF CONTENTS

GLOSSARY

PART A :

1. NLP : Natural Language Processing.
2. Keyword : Subject/Topic for tweets.
3. TF : Term Frequency
4. IDF : Inverse Document Frequency.
5. TF*IDF/ tf-idf : Term Frequency - Inverse Document Frequency Algorithm.
6. Max count : Maximum Count of the words Algorithm.
7. Weight : Impact factor of a word calculated using tf-idf.
8. Wordcloud : Data visualization technique to represent most dominant words based on their frequency/weight.
9. Hashtags : Words/Phrases used in tweets that describe the topic discussed in tweets.
10. Dominant subject : most repeated topic among the tweets.
11. Stopwords : Most common words such as *a, the, an, of, but,* etc.,
12. Preprocessing : Data Cleaning

PART B :

1. Sections 1, 2, 3 are included in the Survey Part.
2. Sections 4 and 5 is the development of the idea, implementation and the analysis.
3. Section 6 and 7 discuss the conclusion and references.

PART C :

This project comes under : ***You could implement a program (in any language you want) that tries to solve a particular problem of your interest (e.g., search technique, planner, logic, etc.).*** The topic discussed in this project are :

1. Text Analysis.
2. NLP/Machine Learning.
3. TF-IDF
4. Topic Discovery

PART D :

In case the results are not clear in the report, the images of histograms and the wordcloud have been included along with the code files.

1. INTRODUCTION :

Text Analysis can simply be defined as a process that allows the machines to make sense of a bunch of data. The most common source of data is Twitter API, Facebook API, etc., With the growing technologies, researches have come to a point where they can use textual data to interpret the damages caused and/or predict the damage caused by be it hurricanes, earthquakes, etc., For example, one can analyze the damage caused by Hurricane Harvey in Houston by pulling up the tweets from that time. Few of the most common applications text analysis would be Social Media Marketing, Product Reviews, Business Intelligence, Spam Filtering, etc., With the applications being in wide range and only becoming more and more, it has become important that the analysis is done more efficiently and effectively. One of the major problems that would be discussed in this project would be keyword extraction.

Keywords can be defined as the most dominant topics being discussed in the text data. With the help of keywords, we would be able to get a high level understanding of the data. The applications of keyword extraction are Information Retrieval, SEO, etc., Traditional methods such as manual assessment would take hours or days to get through all of the data. Even though it is the most accurate interpretation but it is a long and tedious. As it turns out, keyword extraction is still one of the most difficult tasks in NLP. Ideally, keywords extracted should give the gist of the data we have. But most of the times we do not have the resources or the right approach. In this project, one of the problems faced during keyword extraction is discussed which is the inclusion of hashtags.

2. LITERATURE REVIEW:

For the application of tf-idf on twitter data, we have looked at the paper [1]. In this paper, it is explained how we can implement tf-idf on text data. The exact methodology is followed as explained in the paper but on tweets. The preprocessing steps and the calculation of tf-idf are implemented using the tidyText package. The representation of the results however is not from the paper[1]. Mathematical understanding of the algorithm is done based on the reference from paper[2]. The authors discuss the implementation of tf-idf algorithm on UN database from 1988 on about 1400 documents.

Based on the results from paper[2] which imply that tf-idf is an efficient algorithm for extracting keywords, we have decided to work on this particular algorithm. To get more information on the application of tf-idf algorithm [3] was referred to. Information on text analysis and preprocessing was gathered from [5] and [6]. These two sites explained the implementation of text analysis. R package tidyText[4] was used all through for all the algorithms. Twitter API was used from [7].

3. ALGORITHMS :

3.1 PROBLEM STATEMENT:

The main idea of this project is to do a comparison analysis of keyword extraction algorithms which give importance to hashtags and the ones which don't. Ideal approach for keyword extraction is the main goal. The algorithms are discussed in the following subsections. For this project, only tweets which have hashtags have been extracted. Hashtags are separated into a new column which would be used in the hashtags approach. But the tweets even though they do contain hashtags in them, during the preprocessing step, the punctuation marks are removed and they are considered as any other word and no importance is given to them in the tf-idf approach or the max count approach. By the end of the experiments, the relevance of hashtags for this field is going to be determined.

3.2 KEYWORD EXTRACTION :

As it's discussed what exactly keywords are in the previous sections, this section is mainly going to cover how it is done. The input data is tweets, the main ideology is to break the data down into simple words and calculate the impact factor of each word and determine the most dominant subject i.e., the word(s) that discuss the main/major topics of the data. For example, if we have tweets from 25th of December, it is expected that the most dominant subject would be Christmas, Winter, Break, etc., This can be used in a lot of fields marketing especially. For this project, three approaches have been taken under consideration, which are Max Count,TF*IDF(Term Frequency * Inverse Document Frequency) and Hashtags Approach.

3.3 MAX COUNT :

This approach is as simple as it sounds. The tweets are broken down into a set of words and each word has a counter. Stopwords are removed. Punctuation marks are removed. The words that have the maximum count are returned which means that the words with the maximum count are the keywords and the most dominant subjects. In the case where the tweets are the input, the first step is to pre process the text and then divide them into words. A counter is initialized which keeps a count of each word and for this project, 25 most common words have been plotted in the histogram and 150 most common words have been used to plot the word cloud.

EXAMPLE :

1. Winter is a beautiful season.
2. Iced Tea is better than Iced coffee.

RESULT :

	Word	Counter
1.	Winter	1
2.	Beautiful	1
3.	Season	1
4.	Iced	2
5.	Tea	1
6.	Coffee	1

These results are calculated under the assumption that stopwords are efficiently removed. Based on the counter, the words with maximum count are considered as the most common topics i.e., the keywords.

3.4 TF*IDF :

TF*IDF is just another way to judge the topic of an article or a document based on the words it contains. In this algorithm, each word would be given a TF*IDF score which would be the weight of the word and not the frequency. In the TF part, the frequency is calculated i.e., the number of times the word has appeared in the document in this case in one tweet. However, the longer the document higher the chances are for the word to be repeated. To avoid this problem, the score is normalised. Once the tf is calculated, which is the word frequency, the next step is to calculate the idf. IDF can be explained as a process that would determine the importance of each word and throughout the project the importance is going to be referred to as weight of the word. While calculating TF, the consideration is that all terms are equally important. IDF is used to determine the significance of each word in the entire dataset. This is calculated by looking at the total no. of documents and in this case total no. of tweets and the number of tweets that have the term in it. The mathematical formula is given below on how to calculate both TF and IDF.

For a term ‘t’ in a document ‘d’, the weight ‘Wt’ of term ‘t’ in document d is given by

$$Wt = TFt * IDFt$$

Where :

- TFt is the number of occurrences of t in document d.
- IDFt is the significance measure of the word.

Term Frequency can be defined as the number of times a word has appeared in a document and it is ideal to normalize the value. Hence TFt can be given by

$$TFt = \#words/length(document)$$

Inverse Document Frequency on the other hand can be used to get the significance of the word in the whole corpus. We use log based 10 on IDF to dampen the effect. For example, if the words have a count which are in 1000s or 10000s, the easier way to represent would be using the logs. IDFt can be given by

$$IDFt = log(N/DFt)$$

Where :

- DFt is the number of documents containing the term t.
- N is the total number of documents in the corpus.

In simpler words, the higher the score is, the rarer the word appears and vice versa.

3.5 RELEVANCE OF HASHTAGS :

Hashtag is basically a word or a phrase that is included in a tweet and it basically summarises a tweet. Given that’s exactly what keyword extraction is used for, hashtags can be considered as one of the most relevant subjects for this process. For example, let’s look at the following tweets.

a. This product is extremely overrated. I regret buying it more than anything.

b. This product is extremely overrated. I regret buying it more than anything #kyliecosmetics.

The examples above are the same tweet with and without the hashtag. Taking a look at the first example, it is obvious that it is a negative tweet but it is unknown what the user is talking about which makes the tweet completely irrelevant in the case of product reviews or any study for that matter since the subject is undefined. However upon looking at the second example, it is clear what the user is reviewing. Hashtags can be very relevant if the motive is to find out the most dominant subjects on a given day or on a given keyword.

In this project, the comparison analysis algorithms is done. The first approach is the max count method, the second approach is the TF*IDF algorithm, and the third approach is the inclusion of hashtags using the max count method. In the following sections, the implementation and the results of the algorithms is mentioned.

4. IMPLEMENTATION :

This section discusses the code implementation of the algorithms discussed in the previous sections. For this project, only tweets with hashtags are taken under consideration for better results. All the tweets are in English.

4.1 PLATFORM AND PACKAGES :

All the experiments are done in R studio. Twitter API was used to gather the tweets. All the data subjected in this project is extracted by the author alone. The R packages used are “rtweet” and “dplyr” for the extraction of tweets, “dplyr” and “tidytext” for preprocessing the tweets and for the extraction of keywords for both max count and tfidf. “Ggplot2” and “wordcloud” are used for the visualization of the data.

4.2 INPUT DATA :

The input data for this project is a set of tweets which are around 520 in number. They have been extracted on the day of 20th July, 2019 for a certain keyword. A keyword has been used to extract these tweets since randomly extracted tweets are mostly neutral and contain no subject generally. To avoid that, the tweets which have the keyword ‘sad’ in them were extracted which means the tweets which have the word sad in them as a part of the tweet or a hashtag. Although extracted tweets contain a lot of information, for this project only the following were taken under consideration *user_id, created_at, text, location, hashtags.* In short, this project would be predicting why people were sad on 20th July. At the end of this experiment, there would be a bunch of words which would be the subjects that are the reasons for people's sadness for that particular date.

4.3 CODE FILES :

There are totally three code files and one csv file. "tweets_2.R" is the file that extracts the tweets from the twitter API. Twitter Developer account is needed to get access to the live tweets and with the help of that tweets with the keyword 'sad' are extracted on the day, 20th July, 2019. The user_id gives the unique id for each tweet. Created_at is the time at which these tweets were posted(the data is converted as characters hence the format in the csv file). The text is the tweet by each user. Location gives the city and state or country from where the user tweeted. Hashtags are extracted as a different column.

"tidyText.R" is used to determine the keywords for the dataset using the max count approach. Firstly, the data is preprocessed and punctuation is removed. The tweets are then divided into a bunch of words. All the steps are carried out using the tidyText package. The section text is used for this algorithm. The next step is to remove all the stopwords. Although the predefined stopwords remove most of the unwanted words but since the derived tweets have a common keyword, "sad", sad and a few other irrelevant words were removed as well. A histogram is plotted next for the top 25 words that have the maximum count. Following which a wordcloud is made for the top 150 words. The images of the plot and the wordcloud shall be put in the following sections. In the same file, the same process is repeated for the hashtags except for preprocessing. Max count is used to determine the most dominant subject among the hashtags since they are purely words and the ideal way to determine them would be to check the words that have been repeated the most. A histogram and a wordcoud has been plotted for the hashtags as well.

"tttfidf.R" is used to determine the keywords based on the TF*IDF algorithm. The initial steps are the same as the above method. tf*idf is calculated for each word. Created_at has been used here since user_id may be repeated and that would cause a discrepancy in calculating the term frequency if the document is not uniquely identified. Once the weight of the word is calculated, the top 25 words are plotted. As discussed already, if the tfidf is close to 0, it means that the word is more significant. Top 150 words are used for the wordcloud.

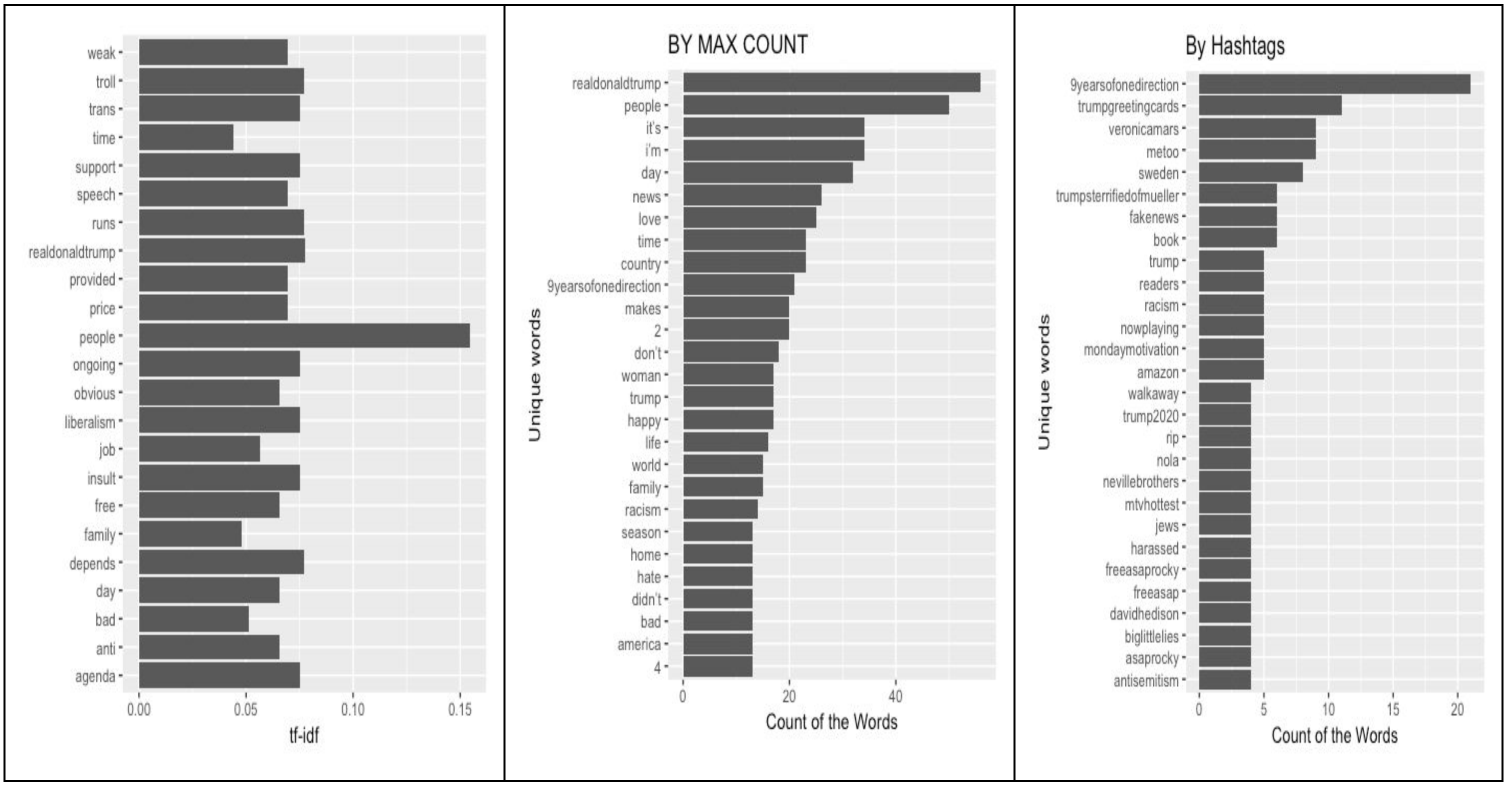
4.4 OUTPUT FORMAT AND FILES :

The output for the tweet extraction code is a csv file which is named as tweets_h.csv. This has the tweets for the keyword sad, which would be used as the input file for the rest of the two files. The csv is included along with the report and the code files. The output of the rest of the two files is histograms and word clouds. These are going to be included in the results section for comparison purposes. Altogether, there is 1 csv file and 6 output graphs/images

5. EXPERIMENTAL RESULTS :

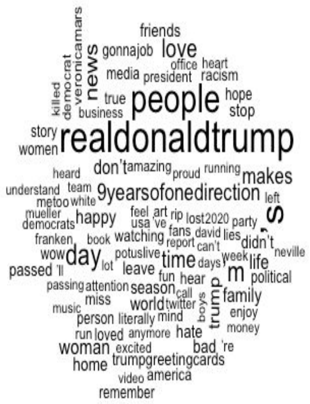
This section is divided into two parts. The first part will include the histograms from three algorithms and the second part would contain the wordcloud from the three algorithms.

5.1 HISTOGRAMS

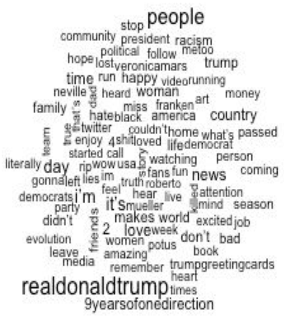


By looking at the histograms, the most dominant subjects can be seen. As it is clear, both tf-idf and the max count method for the tweets are not very clear on what exactly is the topic . Even though they give a list of words that have been used the most and the words that have the most weight, it is not enough to pinpoint the topic from those bunch of words. However by looking at the histogram from the hashtags, it has a bunch of topics that are prominent. For example, 9yearsofonedirection is the most common topic in the 500 tweets that are used. Even though it has been identified in the other histograms as well, there are other words which have dominated that topic such as ‘people’, ‘woman’, which are common nouns which might occur in a lot of tweets but do not specifically give out a topic.

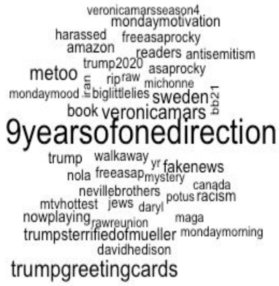
5.2 WORDCLOUD



Tf-idf



max count



hashtags

By looking at the wordcloud above, we can tell that they are a better representation when compared to the histograms since they appear in different sizes and are more easy to read and understand. Tf-idf’s wordcloud looks more relevant with the one of the hashtags. Again, hashtags wordcloud seems more relevant and accurate against the manual comparison.

6. CONCLUSION

The motive of this project was to understand how relevant hashtags are when keywords are being extracted to determine the most dominant subjects. The experimental results have been compared to manual assessment of the data. Upon that comparison, it is observed that the results with hashtags are more accurate when compared with the results of the entire text data. Even though the hashtag would be present in the tweets when tf-idf or max count was performed, the reasons why only hashtags results is more accurate is because of the presence of more number of words. People tend to describe a topic without even mentioning the topic in their tweets and just the topic as a hashtag. In such cases, it is difficult to determine the topics as the words like people, women, men, books etc., which could easily appear in many tweets would be considered more relevant.

However, when only hashtags are taken into consideration, we already have the topics on which the tweets are made. So the only other thing that needs to be determined is the most discussed topic which can be done easily by using the max count algorithm. However in the tf-idf method and the max count method of the tweet text, during the preprocessing step where the punctuation marks are eliminated, the hashtag is removed and the word is considered just like any other word in the text data. Not giving importance to the hashtag would result in the irrelevant dominant subjects.

Upon working on three different methods and looking at the results from all the three methods, it is understood that hashtags carry a lot of importance in extraction of dominant subjects and weightage given to them during the analysis should be more. At the end of the day, the results that would make sense are the topics and not any random words and based on the results from these methods, sensible outputs have been shown by the hashtags approach. However, tf-idf and max count of the tweets also have some of the topics highlighted but the importance given to them is low even when they are the most common subjects(from manual assessment).

7. REFERENCES :

[1] DOMAIN KEYWORD EXTRACTION TECHNIQUE: A NEW WEIGHTING METHOD BASED ON FREQUENCY ANALYSIS by Rakhi Chakraborty, Department of Computer Science & Engineering, Global Institute Of Management and Technology, Nadia, India.

[2] Using TF-IDF to Determine Word Relevance in Document Queries by Juan Ramos, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855

[3] TF-IDF https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

[4] tidyText https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html

[5] Text Analysis Algorithms https://algorithmia.com/tags/text%20analysis

[6]Text preprocessing https://hackernoon.com/text-processing-and-sentiment-analysis-of-twitter-data-22ff5e51e14c

[7] Twitter API https://developer.twitter.com/en/apps