An analysis of Uber related tweets

Assignment 1: COSC2671 - Social Media and Network Analytics
By Adelbert Choi

1 Introduction

1.1 Background

Founded in 2008, *Uber Technologies, Inc.* (Uber) is a company that develops, markets, and operates a ride-sharing smartphone application aiming to connect drivers of vehicles for hire with riders (people who needs a ride). Overall, this application allows the effective and efficient scheduling of transportation services between its users (bloomberg.com 2018). In the recent years, Uber was also able to successfully enable delivery servicing (e.g., Uber eats) related capabilities through a different application platform.

On March 18, 2018, Uber's self-driving car test in Arizona resulted in the death of a pedestrian. This unfortunate incident is expected to negatively impact the public's perception towards the company. This assignment mainly aims to collect and utilise Uber related Twitter tweets data to explore this belief. In addition, this assignment also aims to determine a set of topics that can summarise the current Twitter content associated with Uber.

2 Methodology

2.1 About the Data

This assignment utilised Twitter tweets data to achieve the aforementioned objectives. In particular, tweets exhibiting the keyword uber (and #uber) were collected, pre-processed and analysed. In addition, this assignment focused on analysing Uber related tweets from the United States (US). Only analysing tweets from the US significantly decreased the number of tweets collected, making it more manageable to process.

Uber related tweets data were gathered using Twitter's REST Search API. Using Twitter's REST Search API, 8,010 tweets from March 23 2018 to April 2 2018 in the US were obtained. It is worth noting that not all Twitter tweets have a reported location; thus, the tweets obtained for this assignment is not an exhaustive list of all Uber related tweets in the US. This assignment will assume that the tweets data obtained will be able to reflect the overall situation in the US.

2.2 Data Preprocessing

A number of text pre-processing procedures were implemented on the collected tweets data. These include:

1. **Transformation** of words. All words in each tweet were transformed to be in lower case. This transformation was implemented to allow easier identification of identical words.
2. **Tokenisation** of each obtained tweet. This was performed to ensure that each word across all retrieved tweets can be pre-processed accordingly in a similar systematic manner.
3. **Removal** of irrelevant words across each tweet. These include english stopwords, urls, frequently occurring words, punctuations, and other words and characters determined to be meaningless for this assignment were all removed from the analysis. Based on an initial observation of the data, showcased below are some of the words/characters that were determined to be irrelevant and were removed from the analysis.

   ['rt', 'via', '...', '…', ".", "’", "…", "!", "?", "...", ":", "“", "”", "$", "'", "/", "-", "..", "u"]

4. **Lemmatisation** of words. Lemmatisation is a process that aims to group together variant forms of the same word (thefreedictionary.com 2018). In other words, it attempts to transform various grammatical variations of a certain word into a single standard. Unlike, Stemming, it ensures that a valid word is provided (a word’s lemma). For this assignment, a part of speech tagger function was created, attempting to obtain the most appropriate lemma, based on the context, for each of the words in the tweets obtained.

2.3 Analysis

After the data is pre-processed, this assignment applied topic modelling and sentiment analysis to the Uber related tweets data.

Topic Modelling was first applied to the pre-processed tweets. Specifically, topic modelling was implemented using a method called **Latent Dirichlet Allocation (LDA)**, which is a probabilistic modelling approach that aims to collect or group discrete variables together (e.g., text corpora) (wikipedia.org 2018). For this assignment, LDA was utilised to assist in identifying topics associated with Uber. Using the topics identified, tweets can be classified accordingly and the main idea of the tweets can also be easily understood at a high level.

Sentiment analysis was also applied to the pre-processed tweets. Sentiment Analysis was done to obtain insights regarding the public’s perception towards Uber. Particularly, **Vader (Valence Aware Dictionary and Sentiment Reasoner) Sentiment Analysis** was used to obtain a score measuring the sentiment of tweets. The sentiment score is obtained by summing the positively classified words versus the negatively classified words in each tweet. The resulting score produced will have a value between -1 (negative sentiment) to 1 (positive sentiment). Descriptive summaries of the acquired sentiment scores were then used to provide insight into the public’s view with regards to Uber.

3 Results and Discussion

3.1 Descriptive Summaries

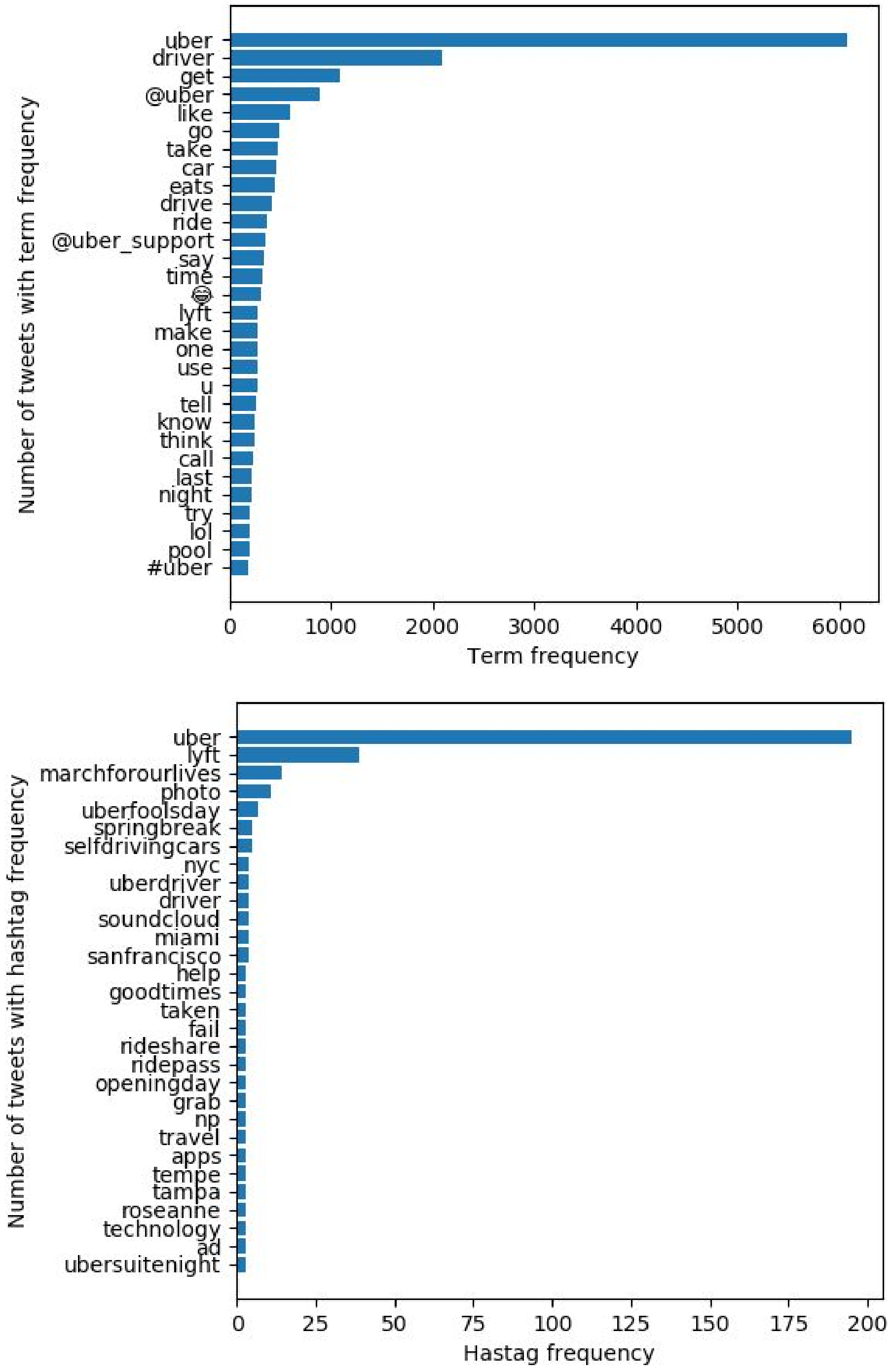


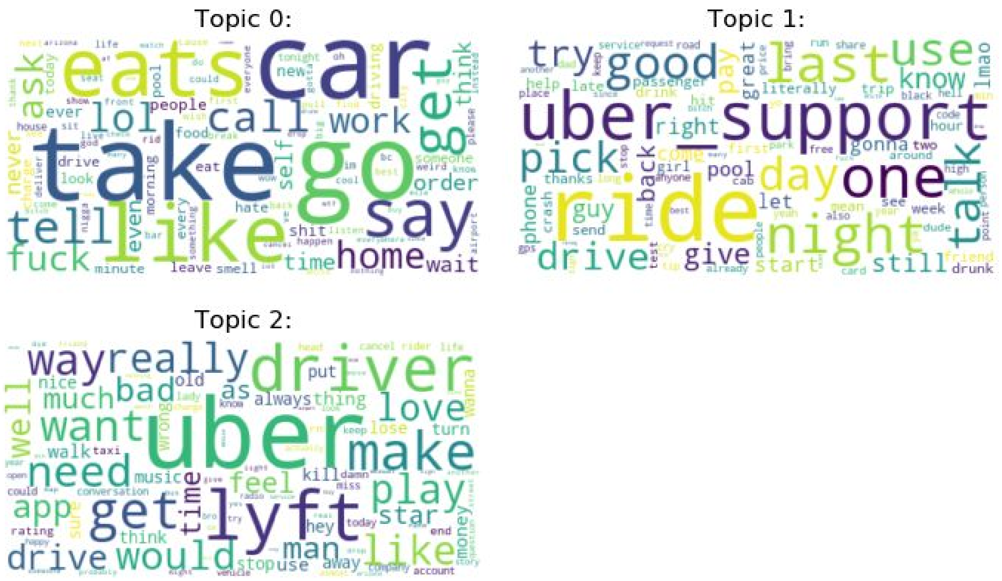
Figure 1. Top 30 most frequent terms across tweets (top) and Top 30 most frequent hashtags across tweets (bottom)

The top panel of Figure 1 showcases the top 30 most frequent words observed across all collected Uber related tweets. It is not surprising to see that the word ***'uber'*** is the most frequent term observed across the tweets. Specifically, the word was observed in 6,081 of the collected tweets. This is followed by the word ***'driver'***

(1,096), then ‘***get***’ (2,093). It is good to note that these three terms were removed from subsequent analysis. These terms were removed due to their high volumes. To be specific, when these words are included in topic modelling, these words were observed to be consistently present across constructed topics, making it difficult to interpret results generated. This might be due to their relatively high volumes. Compared to other words on the dataset, these three words were seen to consistently have higher frequencies than other words even when divided across different topics.

On the other hand, the bottom panel of figure 2 illustrates the top 30 hashtags observed across the collected tweets. Once again, it is not surprising to observe ***‘#uber’*** (195) being the most common hashtag across the collected tweets. This is followed by ***‘#lyft’*** (39), one of Uber’s competitors in the US, and ***‘#marchforourlives’*** (14). ***‘#marchforourlives’*** is a trending hashtag used by students advocating for gun-safety in the US. It can be seen from this frequency distribution that there are numerous unique hashtags across the collected tweets. This could indicate the potential of observing various topics being discussed in association with Uber. However, the low volumes counts of these hashtags may not contribute as much when constructing topics. For these tags to have a significant impact, more tweets data containing these keywords must be collected.

3.2 Topic Modelling



Topic 0:

Topic 1:

Topic 2:

Figure 2. Word clouds of topics generated

Continuously exploring the number of possible underlying topics making up Uber related tweets, it was decided that three topics was the most meaningful. These three topics identified can be considered to be overall topics, where each generated Uber related tweet can be classified to one or more of the topics identified. Figure 2 displays the word clouds of each of the three topics identified, the words making up each topic’s word cloud are the keywords found to be commonly co-occurring and can be used to describe the topic.

Focusing on topic 0, it can be seen that the keywords making up the topic include ***eats***, ***car***, ***take***, ***like***, ***go***, ***home*** and ***work***. These keywords lean towards indicating tweets talking about the services Uber provides to its users. Generally, these include Uber food delivery and car rides home and to work.

On the other hand, some keywords contained in topic 1 includes ***uber_support***, ***ride***, ***good***, ***talk***, ***use***, and ***drive***. The keywords in this topic could possibly indicate tweets relating to customers' interactions with the Uber support team (uber_support). Based on the keywords, it can be said that some tweets to uber_support are about a customer's car ride experience with a Uber certified car.

Lastly, topic 2's keywords may suggest tweets relating to Uber drivers in general. Numerous words in this topic contain adjectives that can be used to describe Uber drivers. These include ***well***, ***love***, ***play***, ***like*** and ***bad***. Also, it can be seen that the keyword ***lyft*** is commonly occurring in this topic. This could possibly indicate tweets comparing Uber drivers to Lyft drivers.

Table 1. Overall topic distribution

Rank	**Topic Number**	**Number of Tweets (Relative Proportion)**
1	0	3680 (0.4594)
3	1	2042 (0.2550)
2	2	2288 (0.2856)

Using the identified three topics, each tweet in the dataset was assigned a topic. Assignment of a topic to each tweet was based on the highest probability of a tweet to be a certain topic, which is based on a tweet's words. Table 1 presents the topic distribution between the three identified topics. It can be seen that almost half (45.94%) of the gathered tweets are classified to be topic 0. Provided this observation, it can be said that Uber related tweets on Twitter are mostly about the services Uber provides.

3.3 Sentiment Analysis

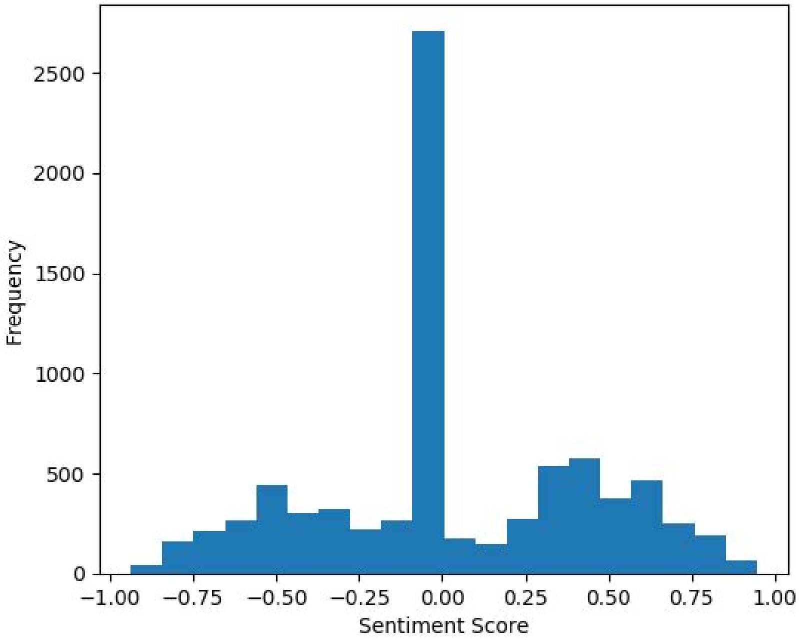


Figure 3. Histogram of sentiments scores (sentiment scores obtained from vader sentiment analysis)

Figure 3 shows a histogram of the sentiment scores found for each of the obtained Uber related tweets. It can be observed that most of the tweets have a score of 0. This indicates that most of the Uber related tweets collected do not express a positive or negative opinion (or perception). On the other hand, it can be observed that there is also a notable amount of positive and negative tweets.

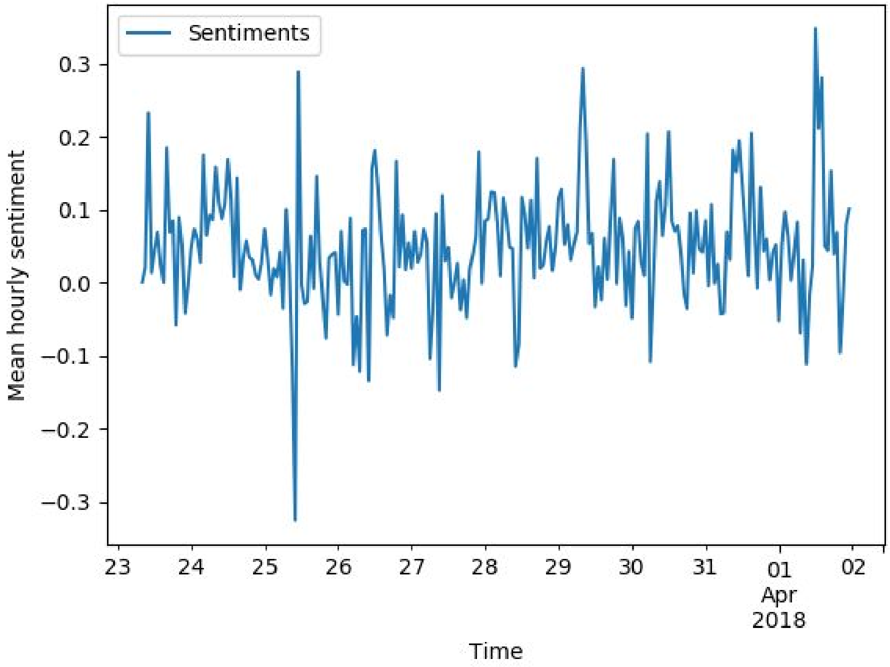


Figure 4. Time series plot of mean hourly sentiments scores (sentiment scores obtained from vader sentiment analysis)

To observe how the sentiments towards Uber changes through time, a time series plot of the sentiment scores was constructed. Figure 4, displays a time series plot showing how the average sentiment score per hour changes from Mar 23 2018 to April 2 2018. No increasing or decreasing trend in the time series can be observed. In an overall sense, this indicates that the public's perception towards Uber is constantly neutral (or slightly above neutral or slightly positive). Thus, it can be said that the involvement of Uber in the recent self-driving car accident did not negatively impact the public's perception towards the company.

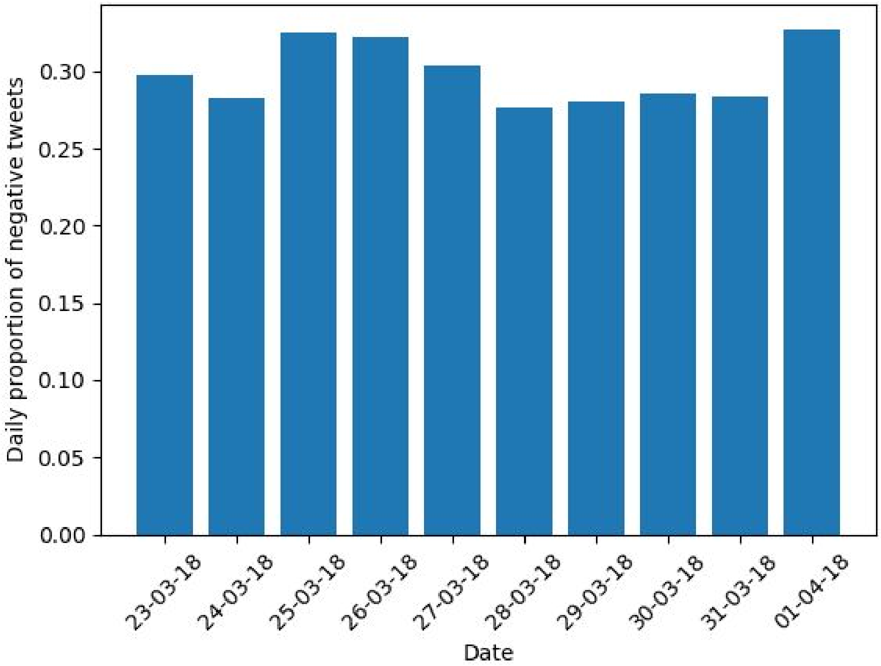


Figure 5. Bar plot of daily proportions of negative sentiments (all sentiment scores that is less than 0)

It was found that 30.14% of the gathered tweets exhibited a negative sentiment value (sentiment score<0). To see if the proportion of negative tweets changes across the days, a bar chart was created. Figure 5 illustrates this, and it can be seen that the daily proportion of negative tweets is generally constant. Specifically, around 30% of tweets everyday have a negative sentiment score. Similarly, this observation could indicate that Uber did not experience any notable negative impact, public perception-wise.

The constant proportion (about 30% daily) of negative tweets seen may be the result of a certain cause. Hence, for future analysis collecting Uber related tweets data before March 18, 2018 may provide a better view if the public's perception towards Uber did change or not.

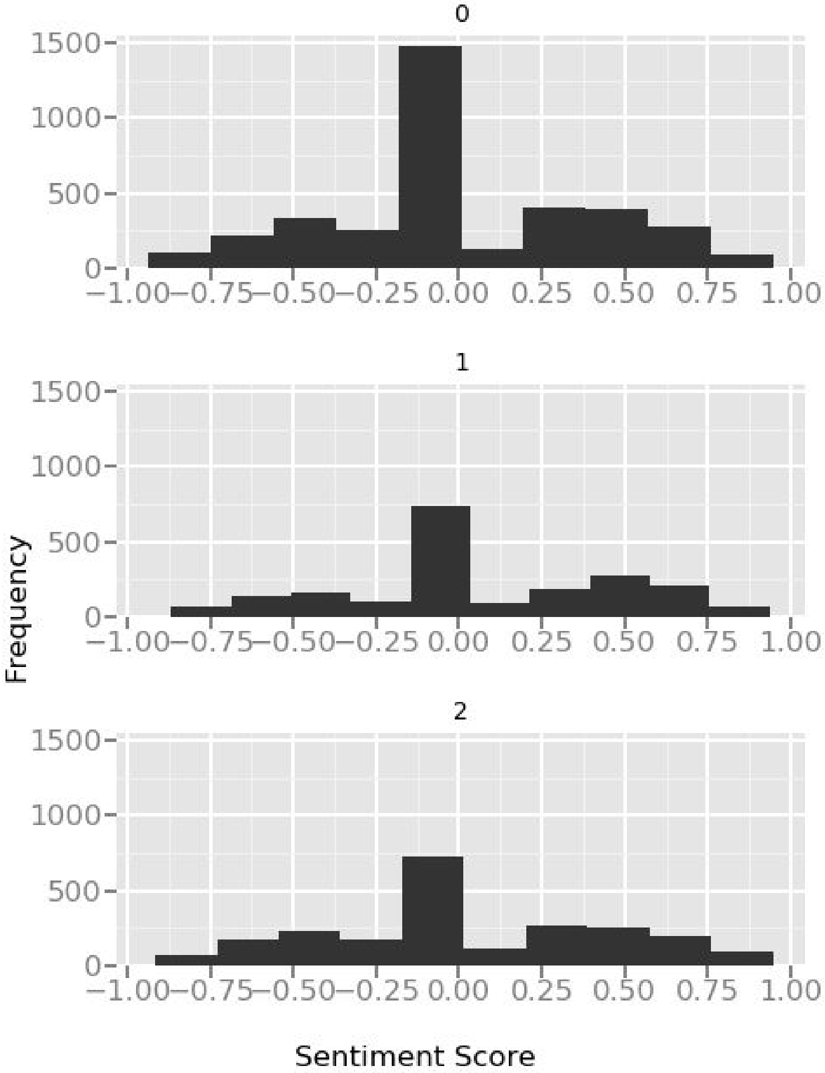


Figure 6. Distribution of sentiments scores by generated topics

Figure 6 allows a quick exploration to observe if the distributions of sentiments differ between the identified topics. In general, it can be seen that all histograms of sentiment scores show a similar distribution to the overall sentiment distribution (see figure 3). From this visualisation, it can be said that no particular topic has more volumes of positive or more negative sentiments.

4 Conclusion

Based on the conducted sentiment analysis, resulting sentiment scores show that a constant neutral public perception towards Uber. Having been involved in an unfortunate self-driving car accident, this event did not negatively impact the public's perception towards Uber. Overall, sentiment analysis results show that everything is quite consistent throughout time.

On the other hand, topic modelling of the Uber related tweets resulted in the identification of three meaningful topics. In summary, these topics relate to Uber services, Uber support, and Uber Drivers. New incoming tweets can be classified to one or more of these topics to obtain a good idea of the tweet’s context.

For future studies, it is recommended that more data is collected. This will result in the discovery of more topics associated with Uber. Additionally, it is also recommended that Uber related tweets from all over the world be gathered and analysed. This will provide a better view of the world’s perception towards Uber. Even though, negative impact was not observed in the US, other countries in the world having an Uber market may negatively perceive Uber as a result of the self-driving car accident.

5 References

[1] *Company Overview of Uber Technologies, Inc*. (2018). Retrieve date April 7, 2018 from https://www.bloomberg.com/research/stocks/private/snapshot.asp?priv capId =144524848

[2] Lemmatisation - The Free Dictionary (2018). Retrieve date April 7, 2018 from https://www.thefreedictionary.com/Lemmatisation

[3] Latent Dirichlet Allocation - Wikipedia (2018). Retrieve date April 7, 2018 from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation