MA331-Coursework

Text analytics of the TED talks by Tim Berners-Lee and Brian Cox

2110866-Khwanchanok-Chumkhun

Introduction

Text analysis is the methodology to investigate the insight emotions and opinions expressed in texts.
This report will utilize text analysis methodology to examine word frequencies and sentiment analyses in order to understand the meaning of 5 TED talk transcripts of 2 TED speakers, Brian Cox and Tim Berners-Lee.
In March 2008, Physicist Brian Cox introduced his work on the biggest scientific experiment, the Large Hadron Collider (LHC) at CERN. After that, in February 2009, he gave a short talk to update the progress of CERN LHC project and the future for the largest science experiment.
The 20th anniversary of the World Wide Web in February 2009, Tim Berners-Lee, the World Wide Web inventor, gave a talk at TED for his new project, “Linked Data” and requested for the raw data now. A year later, in February 2010 at TED University, he presented some interesting result of his Linked Data project from the raw data he requested in his first talk.

Methods

Since the Tidy data approach is one of the effective approach to manipulate text data, then it will be used to word with text to facilitate or to prepare the data before processing word frequencies and sentiment analyses. First, the tokenization is applied to separate the talk transcript into a token which is a single word for this study. After getting the data arranged in the tidy text format, to ensure that the study will focus only to the meaningful data, the list of irrelevant data (stop-words) will be remove from the data set. Then others ‘tidy’ tools will be used to manipulate the data set to analyze the insight meaning of it:
I) The words which are used the most frequent by each speaker will be clarified, and subsequently it will be transformed into wide format and the scatter plot will be used to visualize the comparison of the word frequencies for the 2 speakers (see Figure1 in Results section).
II) The Sentiment Analysis will be done using a lexicon based method. The implication or sentiment will be assigned to each individual word in tidy format using existing sentiment lexicons, then summarize the sentiment orientation of each word to be the representative sentiment of the whole text.
(1) First, this report will present the comparison of overall sentiment of the 2 speakers to show their emotions (sadness, surprise, positive, negative,…) compare to each other by using the lexicon called ‘nrc’. The result will be displayed in term of bar graph of log-odds ratio as shown in Figure2 in Results section, in order to present which speaker is less or more associate in which sentiment.
(2) Then the study will investigate a bit more detail of each talk, the ‘bing’ lexicon will be used to clarify the positive or negative sentiment and the result will be presented as a table of percentage of positive and negative sentiment in overall of each talk (Table1).
(3) Furthermore, again with ‘bing’ lexicon, the more detail analysis will be conducted. Instead of examine the sentiment of whole talk, each talk will be devised into fractions and studied sentiment of each fraction to understand how the speaker emotion changes toward more positive or negative sentiment over the trajectory of each talk. This study output will be demonstrated as histogram chart with the trajectory of talk in x axis and sentiment in y axis (See Results portion, Figure3).

Results

I. Word frequencies

This segment presents the comparison of the word frequencies for the 2 speakers which the result shows as following.

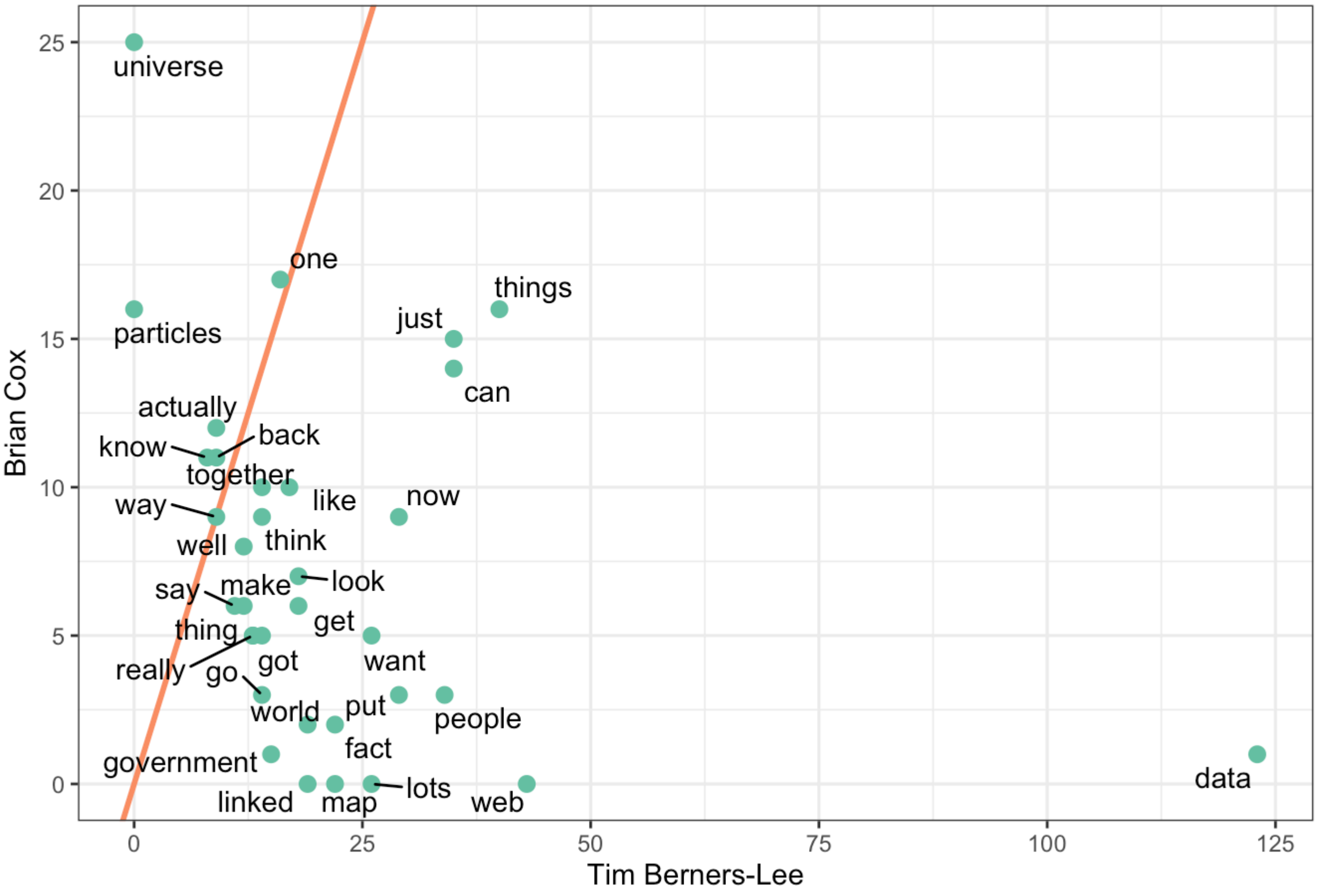

Figure1: Comparing the frequency of words used by Brian Cox and Tim Berners-Lee

The result,in Figure1, shows that there are words which are extremely frequent used by Tim Berners-Lee but not by Brian Cox. These words can lead the understanding of overall meaning of the talks. As in this report, we can know that Tim Berners-Lee's speeches must be something related to data, while Brian Cox might talk about the universe. Focusing at the plot, it is asymmetric, there are words from one side are more far away from the reference line than another side, because the frequencies of some word using by Tim Berners-Lee are immensely high comparing to the high frequent words of Brian Cox. There are 2 main reasons for this, the talks of Tim are longer than Brian. And the main purpose of Tim's talks is to request data for the web reforming so this is the reason why he keep repeating the words 'data' and 'web' many times. But there are some words which are near by the reference line, for instance, 'way', 'one', 'back',… are used about equal frequencies by Brian Cox and Tim Berners-Lee. These are the words that are common and general use and cannot lead to story of the talks.

II. Sentiment analysis

(1) For sentiment analysis portion, the study starts by comparing of overall sentiment of the 2 speakers to show their emotions compare to each other and the result is as below.

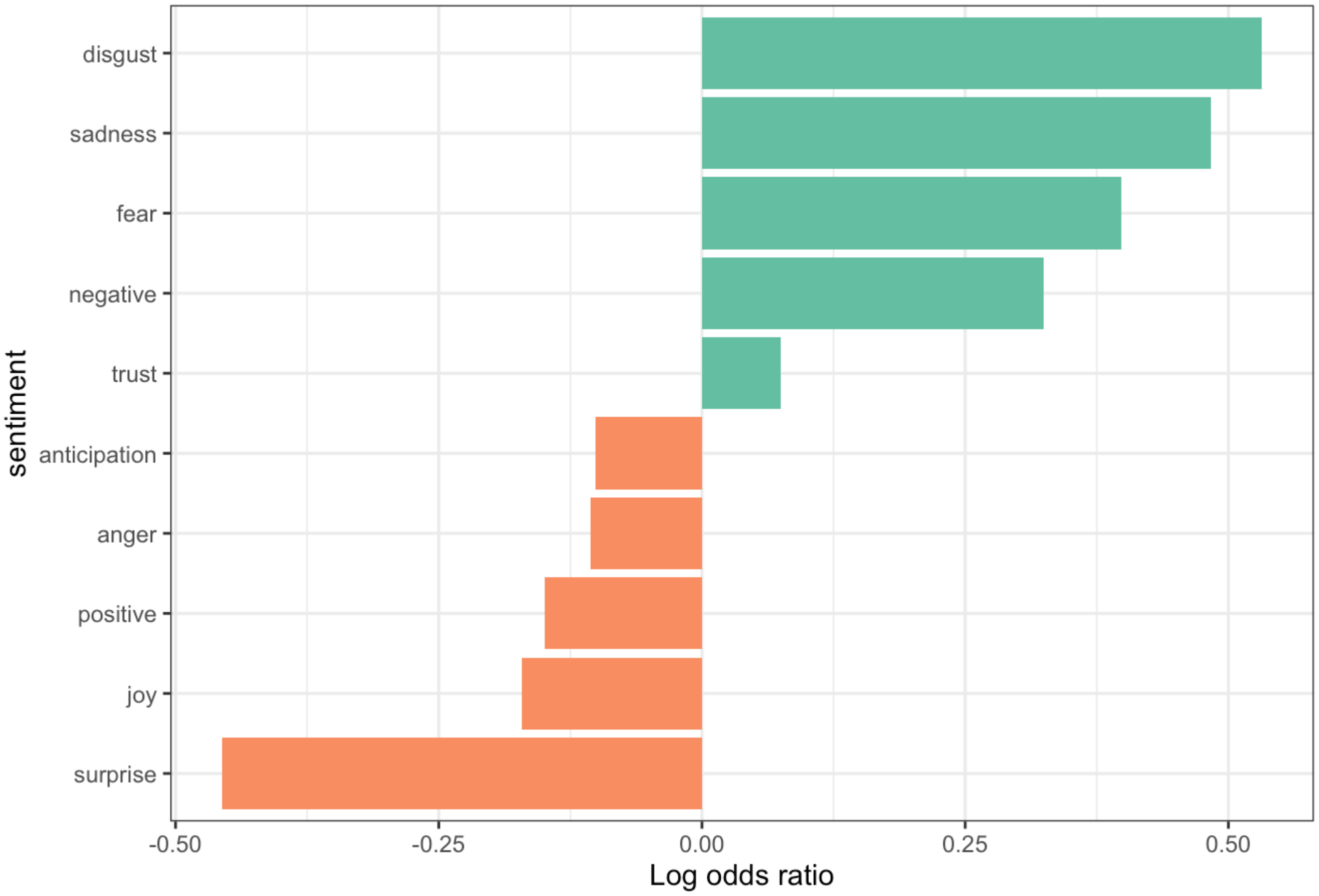

Figure2: The association between sentiments of Tim Berners-Lee's talks and Brian Cox's

The aforementioned Figure2 shows the correlation between sentiments of Tim Berners-Lee's talks and Brian Cox's talks. There are 10 different sentiments ('nrc' lexicon) using to analyze each speaker sentiment as shown in the figure. The positive value of log-odds ratio demonstrates that talk of Tim's talks are more associate in that sentiment more than Brian's talks and the higher the value shows that the sentiment is more strong. In this case, talks of Tim compound more sentiments of extreme disgust, strong sadness, fear, negative and few trust than Brian's. Whereas talks of Brian are more anticipating, angry, positive, very joyful and extremely surprising.

(2) Now more detail investigation is done by analyzing positive and negative sentiment in overall of each talk and show in percentage of it as shown in table below.

Table1: The comparison of negative and positive sentiment of each talk

speaker	headline	sentiment	n	Total	Percent
Tim Berners-Lee	A Magna Carta for the web	negative	15	49	30.61
Tim Berners-Lee	A Magna Carta for the web	positive	34	49	69.39
Tim Berners-Lee	The next web	negative	25	104	24.04
Tim Berners-Lee	The next web	positive	79	104	75.96
Tim Berners-Lee	The year open data went worldwide	negative	6	29	20.69
Tim Berners-Lee	The year open data went worldwide	positive	23	29	79.31
Brian Cox	CERN's supercollider	negative	27	117	23.08
Brian Cox	CERN's supercollider	positive	90	117	76.92
Brian Cox	What went wrong at the LHC	negative	10	29	34.48

speaker	headline	sentiment	n	Total	Percent
Brian Cox	What went wrong at the LHC	positive	19	29	65.52

Table1 shows the number, together with the percentage of negative and positive words using in each talk (headline), and show them ordered by speaker then talks. From the result, it shows that all talks contain more positive words than negative. The most positive talk is ‘The year open data went worldwide’ by Tim Berners-Lee, the talk about his new project, “Linked Data” and he used this occasion to request the cooperate to realize his project (raw data now). So it’s understandable why this talk is more positive comparing with others. On the other hand, the talk ‘What went wrong at the LHC’, which Brian Cox updates the issue that LHC project was facing, so it is also not surprising that this talk is the most negative.

(3) Now the study goes more deep, the sentiment (negative or positive) is showed over the trajectory of the story of the talks. And the results are presented in following Figure3.

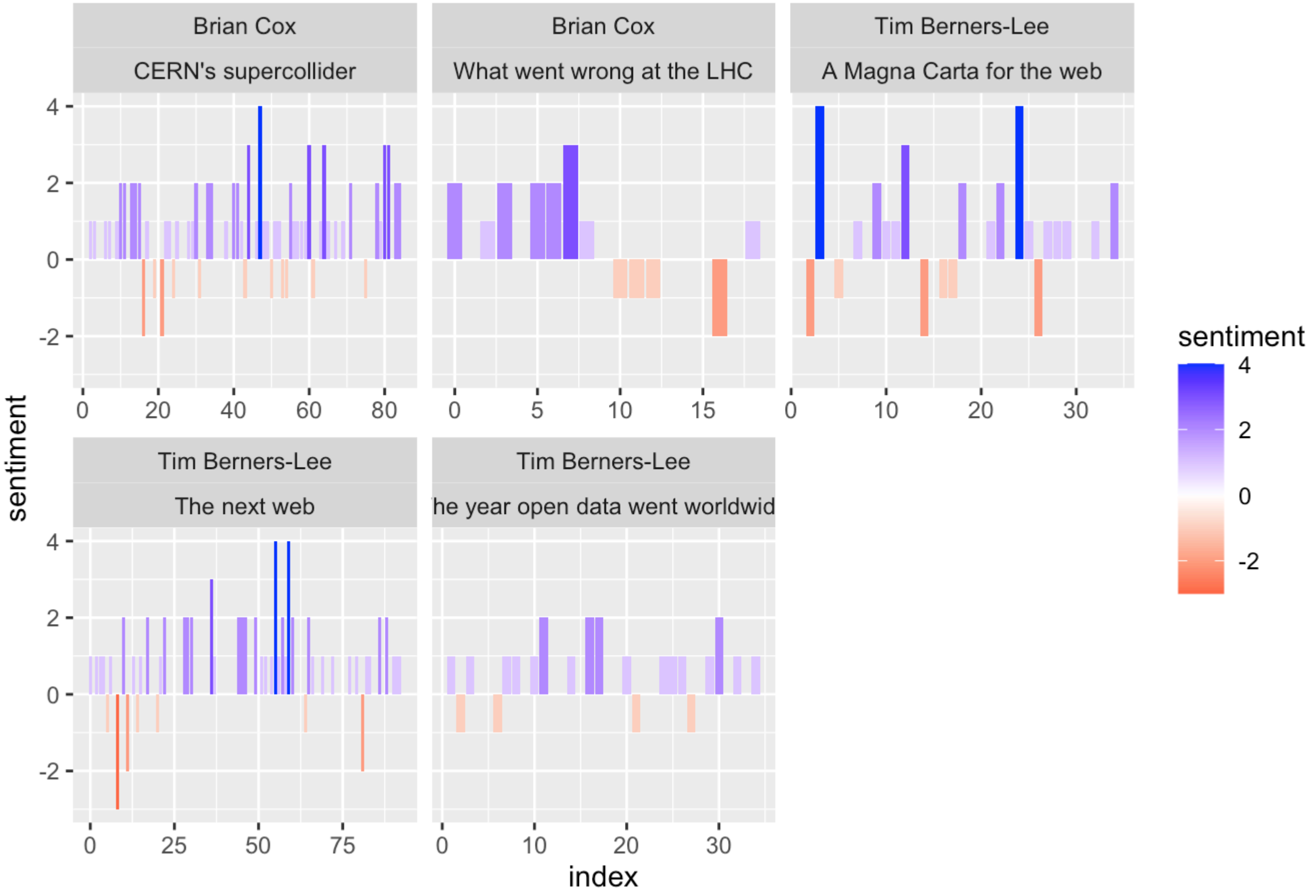

Figure3: Sentiment through the narratives of each talk

The result illustrates how speaker’s sentiment (positive or negative) change over period of time of each talk, it also shows how strong is that sentiment. The histogram graph, which is over 0, represents the positive sentiment of that narrative of the talk and the height of the histogram explains the intensity of the sentiment, the more graph is high, the more intense is the sentiment.
Overall result is the same as the one show in previous table. All talks are significantly more positive than negative. The distribution of positive and negative sentiment of 1st talk of Brian Cox, ‘CERN’s supercollider’ disperse through the talks, while the negative sentiment of his 2nd talk is concentrated in a period of time, it could be the timing that he presented the problem of his work. For ‘A Magna Carta for the web’ of Tim Berners-Lee, this talk is to warn and persuade to fight for the freedom to access data. This talk shows the 2nd negative sentiment in the previous analysis. Then it is normal that there are several strong negative histogram distributed over the trajectory of the talk. The remain topics of Tim shows the normal spread of negative sentiment through the timing of story.

Discussion

As conclusion from this study, Tidy data approach is an effective method to work or to manipulate text data, while lexicon based method is also the one of efficient way for the sentiment analysis.
However, there are some limitations of working with lexicon based method which are :
(1) The limitation of vocabularies in current existing lexicon which is not cover 100% of words. And some words can direct to several sentiments or emotions. These affect to the accuracy of the analysis.
(2) Current existing lexicons are not cover all languages, so this method could not be apply in all situations.
(3) Small section or short text might not be able to get properly sentiment analysis. Especially when the words contains in the text is not match with the word in lexicon.

The main challenges of this work are :
(1) For the person who does not have programming background, it is very challenge to work with the long code and to face and solve when any issue occurs.
(2) There are several solutions that can be used to visualize the data, the reasonable selection and interpretation of visualization is one of challenge point. The presented results should not be too few to unable to cover all and it should not be too many to make the presentation duplicated.
(3) One of speaker’s talks are very difficult and inapplicable, It not easy to understand whether the analysis result is reasonable or not, since the insight meaning of the talks are not clearly understood.

For further investigation, since now the lexicon in R language is not support my native language which is ‘Thai’. I would like to explore more on how to create lexicon for my native language and to be able to use it analyzing text in Thai.