# Text Analysis of Ted Talks using RStudio and Orange

Christopher Dillard

**TED Talks "Ideas Worth Spreading": A Text Analysis To Measure Word-Usage and Its Implications on the Success of These Events**

---


*Photo from: https://tedxwinterpark.com/how-the-ted-conference-become-a-worldwide-community-of-passionate-individuals/*

## How did Ted Talks become what they are today?

The first TED Talk was held in 1984, as a private, invitation-only event that addressed the convergence of technology, entertainment, and design and catered to individuals working within these industries. The first Ted Talks celebrated exciting achievements and fostered creativity only for those with a formal invitation, something which of course limited the events' reach. That being said, these events proved to be a sensation and decades later, in 2006, the first TED talk was made available online which saw immediate success as over 1 million views were reached within that year. In 2007, the company made its online content free for its global audience and, by 2009, the view count reached over 100 million views, just 3 years before it reached its one billionth view in the fall of 2012 (https://www.ted.com/about/our-organization). The company's platform has helped disseminate ideas on a global scale and open discussion in several disciplines beyond the initial 3 from its first event, to include history, economics, social-justice, the meaning of life itself and a broad variety of topics of niche interests in between

which, according to its official website, “invite everyone to engage with ideas and activate them in your community”.

While the company’s growth is impressive and does not appear to be stopping anytime soon, a deeper analysis into a large body of transcripts could provide a better understanding into how their events reach such large audiences and whether the specific word-usage shared across the transcripts from these events might be a factor in their success. Considering how Ted Talks are limited to 18 minutes, word choice can play a decisive factor in the reception of speeches by captivating audiences in this limited time frame. Perhaps certain words are shared across seemingly unrelated Ted Talks for a reason?

## Problem Statements:

For this analysis, I will be using the dataset “TED Talk Transcripts (2006-2021), published on Kaggle by Ramshankar Yadhunath (https://www.kaggle.com/datasets/thedatabeast/ted-talk-transcripts-2006-2021?select=transcript_data.csv), to answer the following questions:

1. **Are there any common themes that can be identified among Ted Talks?**
2. **Does word-usage tend to be more more positive (ie. motivating/inspiring) in Ted Talks?**
3. **Can we use bigrams to explore whether topics discussed in Ted Talks are relevant to a global audience?**

## Importing and Cleaning the Data Set:

```
library(tidyverse)
library(tidytext)
library(readr)
transcript_data <- read_csv("transcript_data.csv")

summary(transcript_data)
```

```
 title            transcript
 Length:4442        Length:4442
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Prior to cleaning the dataset, there are 4442 observations, which should each represent a single Ted Talk. The two columns contain the “title” and “transcript” variable. I am interested in looking at the text contained in the “transcript” column to answer the three problems previously listed above.

### Dealing with empty rows:

```
colSums(is.na(transcript_data))
   title transcript
       0        144
```

There are 144 empty rows in the transcript column which is, of course, useless for the sake of this analysis. So, using the na.omit() function I created a more useful dataframe. Similarly, there are many stop words that do not add any meaningful value in the transcripts that need to be filtered out.

```
tidy_ted <- na.omit(transcript_data)

library(stopwords)
tidy_ted %>%
  filter(!("word" %in% stopwords(source = "snowball")))

tidy_ted_token <- tidy_ted %>%
  unnest_tokens("word",transcript)

tidy_ted_token %>%
  count(word,sort=TRUE)
```

Below, as to be expected, we can see that the transcripts are full of stop words that we don’t want to include in our results:

A tibble: 110,314 × 2

| word <chr> | n <int> |
|---|---|
| the | 350837 |
| and | 247634 |
| to | 222899 |
| of | 195201 |

| | |
|---|---|
| is | 191589 |
| a | 178304 |
| that | 158828 |
| i | 132386 |
| in | 131365 |
| it | 117437 |

Next

123456…100

Previous

1-10 of 110,314 rows

Using the following functions, we can now see a more useful list of the word counts after filtering out the unwanted stopwords:

```
tidy_ted_token <- tidy_ted_token %>%
  anti_join(stop_words)

tidy_ted_token %>%
  count(word,sort=TRUE)
```

A tibble: 109,654 × 2

| word <chr> | n <int> |
|---|---|
| people | 29036 |
| time | 15388 |
| world | 14308 |
| laughter | 14120 |
| life | 8720 |
| applause | 7680 |
| lot | 7163 |
| day | 6685 |
| called | 5988 |
| human | 5599 |

Next

123456…100

Previous

1-10 of 109,654 rows

Although we cannot yet make out any specific topics from the top 10 words above, these words do help us see how the Ted Talks' slogan and purpose play out in their events. "People" is the top word in the transcripts, which shares the list with other words like world, life, and human. We can infer that many of the topics present in the transcripts deal with treating humanity as a global community where ideas are shared at the benefit of everyone. This is further supported by the fact that the vast majority of the instances where "laughter" and "applause" appeared in the transcripts, they appeared in parenthesis to show that they were not spoken words and therefore irrelevant to our analysis. This word cloud produced using Orange Data Mining shows us just how "central" peopl(e) are to the Ted Talk events:


## Sentiment Analysis:

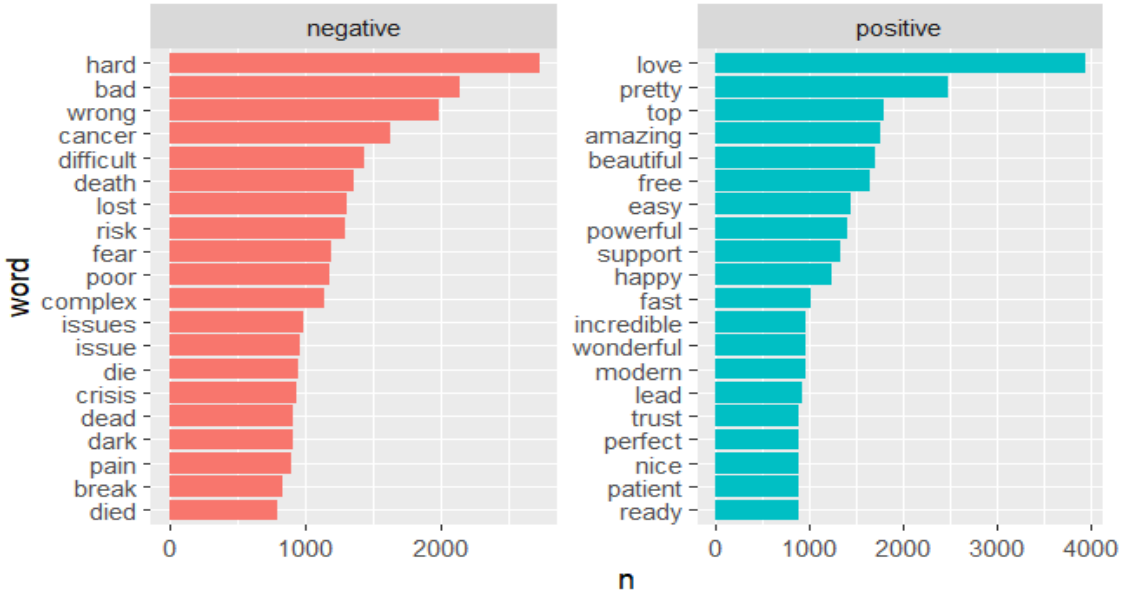
### Top 20 Words:

In this section, I will be running a sentiment analysis on the transcripts to gauge how "positive" or "negative" the word-choice in the Ted Talks are, along with a graph to visualize the frequency of words belong to each sentiment.

```
tidy_ted_sentiment <- tidy_ted_token %>%
  inner_join(get_sentiments("bing"))

tidy_ted_sentiment %>%
  count(word,sentiment) %>%
  group_by(sentiment) %>%
```

```
top_n(20) %>%
ungroup() %>%
mutate(word=reorder(word,n)) %>%
ggplot(aes(word,n,fill=sentiment))+
geom_col(show.legend=FALSE)+
facet_wrap(~sentiment,ncol=5,scales="free")+
coord_flip()
```




## Bigrams

I will be using bigrams to explore how words are used together, which can bypass the problem previously discussed about how context plays a big role in how homonyms can be distinguished and interpreted in text analysis. By grouping words in pairs, topics can began to emerge which provide insight into what themes are being discussed in the transcripts and which topics are discussed the most by measuring their frequency using the count() function. This will also require stop words to be filtered out, as shown in the code below:

```
bigrams <- tidy_ted %>%
unnest_tokens(bigram,transcript,token="ngrams",n=2)

bigrams %>%
  count(bigram,sort=TRUE)
```

A tibble:1,565,765 × 2

| bigram<br><chr> | n<br><int> |
|---|---|
| it is | 39392 |
| of the | 35507 |
| in the | 33036 |
| that is | 20132 |
| is a | 19644 |
| and i | 17138 |
| i am | 15498 |
| this is | 15453 |
| we are | 15358 |
| to the | 14335 |

Next

123456…100

Previous
1-10 of 1,565,765 rows

```
bigrams_separated <- bigrams %>%
  separate(bigram,c("word1","word2"),sep=" ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!grepl("[0-9]",paste(word1,word2,sep=" ")))

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort=TRUE)

bigram_counts
```

A tibble:454,394 × 3

| word1<br><chr> | word2<br><chr> | n<br><int> |
|---|---|---|

| | | |
|---|---|---|
| climate | change | 811 |
| laughter | applause | 777 |
| health | care | 580 |
| million | people | 555 |
| social | media | 421 |
| billion | dollars | 378 |
| million | dollars | 344 |
| york | city | 332 |
| billion | people | 324 |
| public | health | 314 |

Next

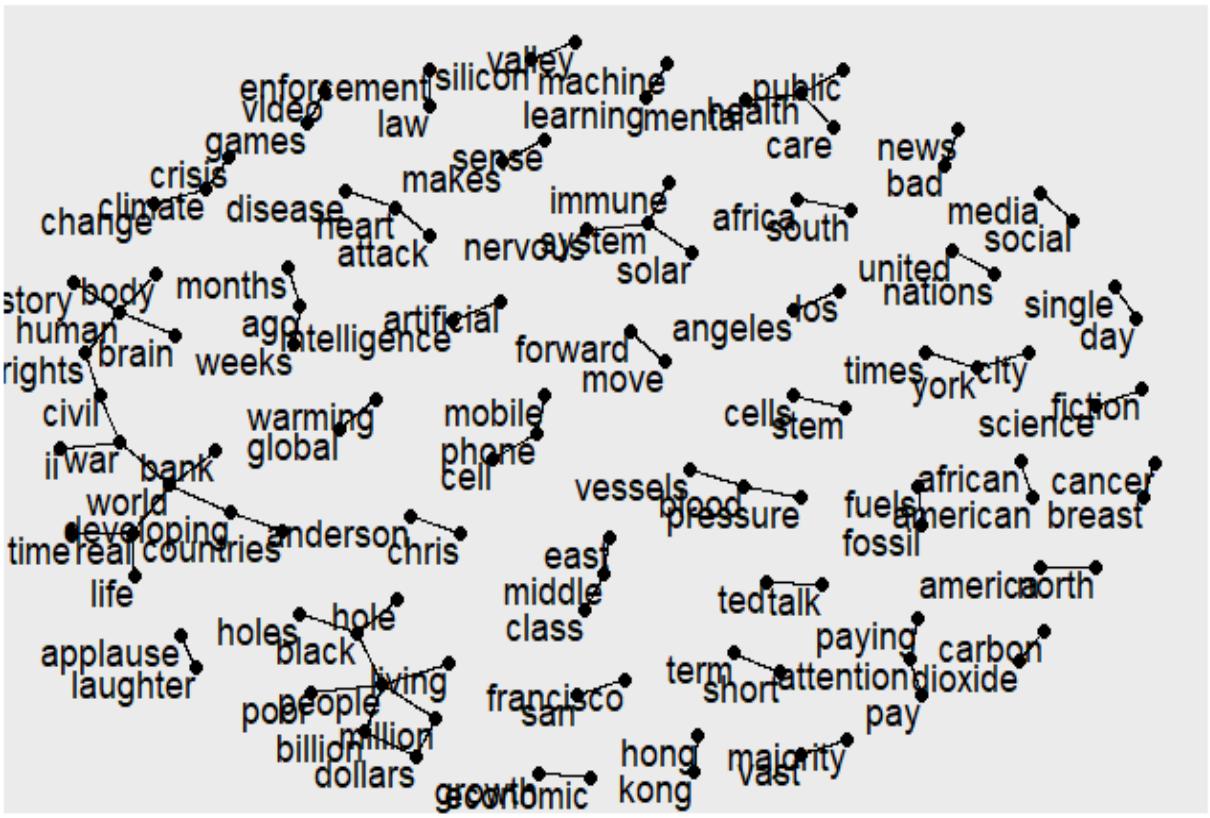
123456…100

Previous

1-10 of 454,394 rows

```
library(igraph)
bigram_graph <- bigram_counts %>%
  filter(n>120) %>%
  graph_from_data_frame()

library(ggraph)
ggraph(bigram_graph,layout="fr")+
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label=name),vjust=1,hjust=1)
```




```
a <- grid::arrow(type="closed",length=unit(.15,"inches"))

ggraph(bigram_graph,layout="fr")+
  geom_edge_link(aes(edge_alpha=n),show.legend=FALSE,
arrow=a,and_cap=circle(.07,"Inches"))+
  geom_node_point(color="lightblue",size=4)+
  geom_node_text(aes(label=name),vjust=1,hjust=1)+
  theme_void()
```


In this graph, I initially set the filter at n>120; however, I noticed that the arrows were obfuscated by the bubbles and text. To fix this problem, the graph above shows only bigrams that appeared in the transcripts over 150 times, with the darkness of the arrow representing a stronger correlation (ie. how often the word behind the arrow preceded the word that it's pointing to.).

## Findings

1. **Are there any common themes that can be identified among Ted Talks?**

This first question relies on all three types of analyses that I used: the simple word count, sentiment analysis, and visualizing bigrams with word networks. After tidying the data by eliminating stop words, people, world, life, and human stuck out as being some of the top 10 words in the word count. This tells us at the very least that the protagonists of Ted Talks and the stories they tell are people sharing their experience with others around the globe.

To further examine the transcripts to find out what the types of experiences that the TED speakers share with their audience, the sentiment analysis on the top 20 "positive" and "negative" words revealed that the most frequent words identified using the Bing sentiment lexicon were led by "love", which was much more frequent than all the others. Looking at the

rest of the words on the positive side, we see “free”, “support”, “modern”, and “trust”. These words can allude to social progression in a more accessible society as things that would resonate with the audience in a positive way. As for the negative graph, we can find words like “cancer”, “death”, “fear” and “poor”. Cancer, being the most frequent words of these four, is something that transcends social class, borders, and unfortunately would resonate with a lot of the audience. The same can be said about death, fear, and for many, poor.

As for the bigrams, a simple count() function of the top 10 revealed pairs like “climate change”, “health care”, “social media”, and “public health” in descending order. This tells us these are the most common themes in the Ted Talks.

2. **Does word-usage tend to be more more positive (ie. motivating/inspiring) in Ted Talks?**

In the sentiment analysis section, the graph displaying the top 20 positive and negative words does offer some incredible insight; but, there are some nuances which should be taken into consideration when categorizing words. For example, while “pretty” can be synonymous with beautiful, which is undoubtedly positive, it can also be used as an adverb if we were to call something “pretty awful”. Similarly, patient (patience) can be considered a positive virtue, yet seeing as though we have the word “cancer” on the negative side, patient could also be a noun in the medical sense where it does not carry a positive connotation.

Unlike on the positive side, the 20 words that appear on the negative side can clearly be considered negative in most contexts. That being said, we can still take note of how there are much more instances of the positive words than the negative ones in the transcripts, with love almost exceeding a count of 4,000, in a collection of 4298 Ted Talks. After taking consideration of the words that are concretely positive or negative, the Ted Talks are almost evenly split, with slightly more “positive” words. This might, however, be an inherent flaw in using the Bing Sentiment Lexicon since words are categorized in a binary fashion where homonyms are treated as the same word, despite having very opposite sentiments.

3. **Can we use bigrams to explore whether topics discussed in Ted Talks are relevant to a global audience?**

To answer this question, I will be referencing both word network graphs along with the output for *bigram_counts*. The top ten bigrams did reveal helpful bigrams previously discussed in question 1, but they also contained less helpful ones like “laughter applause”, “million/billion people”, and “million/billion dollars”. So, in order to visualize much more bigrams at once and visualize clusters of topics, a simple bigram graph with a lower word-count filters (ie. displaying more bigrams). The first of the two network graphs contained several clusters that were related to medical topics, which is not too surprising with “health care” and “public health” being among the top 10 frequent bigrams. Pairs like “heart disease”, “blood pressure”, and “immune system”, and “stem cell” clearly belong to the medical-related topics.

The word network with the light-blue node points was helpful to display how often one word in the pair was preceding the word that the arrow was pointing at, with a darkness of the arrow related to a higher correlation. By formatting the second graph to display only bigrams with counts over 120, it is easier to view while still displaying great results. As far as technology goes, we see that “machine learning”, “social media”, “artificial intelligence”, and “cell phone” are all frequently discussed in the Ted Talks and all very much relevant to us today. “Climate change” was the most frequent bigram in all of the Ted Talks so it came to no surprise that additional bigrams like “global warming”, “fossil fuels”, and “carbon dioxide” are also present in the word network graphs.

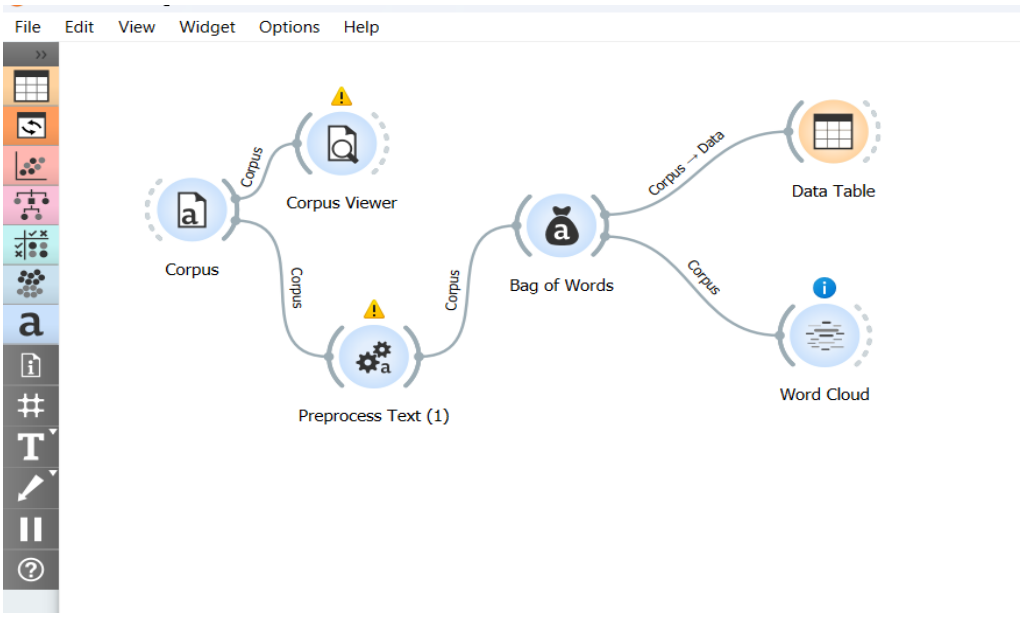
Bigram analysis helps explain that the common themes shared across Ted Talks are all relevant to billions of people, which is why these events have been so successful among their global audience.

## What else could be done?

An additional column in the data set that shows when each Ted Talk would be helpful to evaluate how common topics evolved across the years and run sentiment analysis on Ted Talks belonging to certain years to see how major events might have impacted what people were talking about. Since the bigram graph revealed that economic growth was a common topic, I would like to be able to see how economy-related Ted Talks addressed issues following the 2008 financial crisis and other years of economic recessions. The COVID-19 pandemic was of course a global event, but could we see whether health care and other medical-related bigrams were as frequent in the years prior to the pandemic? In conclusion, one more column devoted to recording the year that each Ted Talk was posted, would have been extremely useful in answering other interesting questions.

**Work Shown:**

My RStudio work will be shown in the .Rmd file in the zipped folder that I will upload with this document on canvas.


File
Edit
View
Widget
Options
Help
Corpus
Corpus Viewer
Corpus
Corpus
Preprocess Text (1)
Corpus
Bag of Words
Corpus → Data
Data Table
Corpus
Word Cloud