TED Talks Data Analysis based on Views, Comments and Tags by Year

Yuying Wang, University of Washington, SOC 225

Data Introduction:

The data used in this study comes from the official TED website. It contains information about all audio-video recordings of TED Talks uploaded to TED.com from June 2006 to September 2017.

People watch TED Talks because they are interested in the contents involved. TED Talk has become a social platform where viewers implicitly express their ideas through the number of views and comments. So it is a useful source from which we can study what topic are people most concerned about, how people react to different social topics and recognize their variation by each year.

Therefore, I selected the following variables from the dataset:

- Number of Views
- Number of Comments
- Published Date
- Content Tags

We will start with correlation and then extend to covariation.

Question 1:

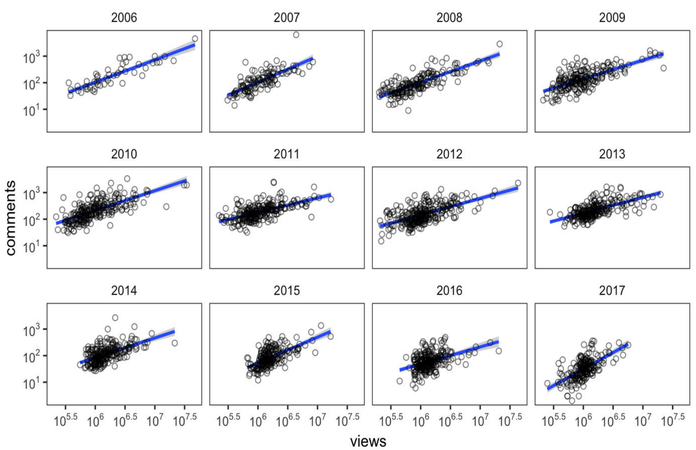
What is the relationship between comments and views?

If we look at talks ranked Top 10 in views and talks ranked Top 10 in comments, we find four of them are matching:

- Do schools kill creativity?
- How great leaders inspire action
- My stroke of insight
- Your body language may shape who you are

The outcome implies that views and comments may have a moderate correlation. Next we take a closer look at the data.

Linear relationship between views and comments



Note: Log transformation is used here to help display a linear relationship.

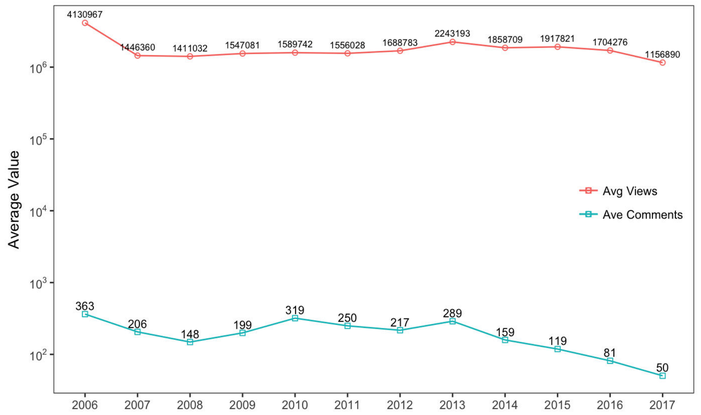
A positive linear relation between views and comments is shown in the graph, but their correlation and steepness of the slope do differ by time.

Year 2006, 2008, 2009 and 2012 indicate a stronger correlation between views and comments with points evenly distributed along the straight line.

Question 2:

How did people' attitudes towards TED Talks change over time? Could they explain the change in correlation differed by year?

Change in average comments and average views from 2006 to 2017



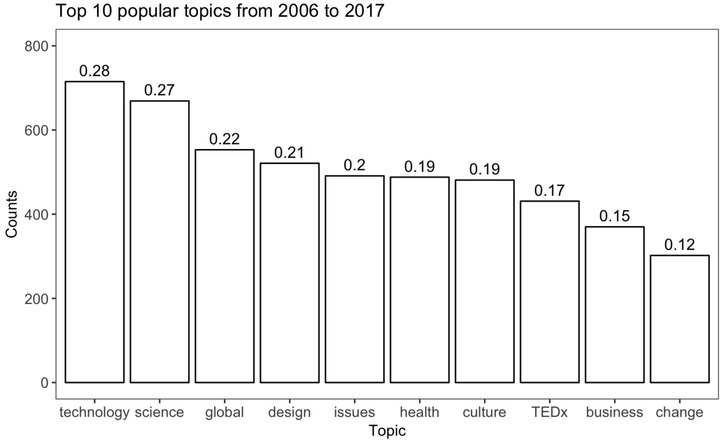
Average number of views and comments is used to explain how people see TED Talks. The graph indicates that the weaker correlation in year 2007, 2014 and 2016 can be explained by non-corresponding change in average numbers of views and comments.

For example, the change in average number of views from 2013 to 2014 is significantly smaller than the change in average number of views by comparing their decreasing slopes, and that could lead to the weaker correlation in 2014.

Question 3:

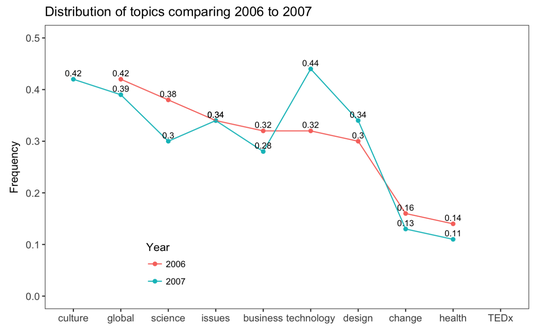
What are the most popular topics in TED Talks and how often do they appear?

Top 10 popular topics from 2006 to 2017



Did the distribution of topics published each year have any influence on the number of views and comments?

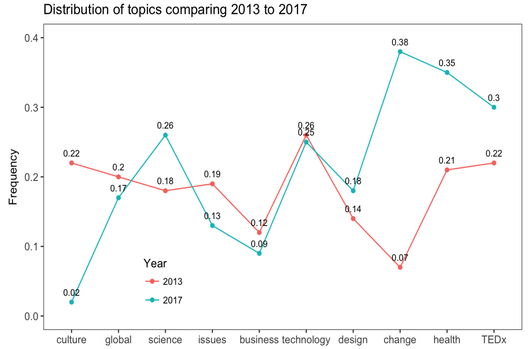
Distribution of topics comparing 2006 to 2007



After comparing topics' appearing rates between years, we see that each topic has a varying distribution and that difference may have effects on the average number of views and comments.

In the examples, we might conclude the increased proportion of 'Technology' accounted for decreasing average views and comments from 2006 to 2007; and the increased proportion in 'Change' and decrease in 'Culture' accounted for decreasing views and comments from 2013 to 2017.

Distribution of topics comparing 2013 to 2017



Discussion:

The data was scraped from the official TED Website and is available to use under the Creative Commons License. Moreover, no TED users information is included in the dataset so privacy is not an essential problem in this study.

From the results of previous questions, we see that variables are related to each other:

- Certain topics produce contents people are more interested in or concerned about in a specific year;
- They boost the number of views and trigger more discussions;
- The corresponding change in number of views and comments then result in the different degree of correlation between them.

However, more precise examinations on the data shall be done before coming up with any solid conclusion. It is possible that some variable is overlooked and factors other than those included in the dataset also have a significant influence on the questions studied. Additionally, looking at only one dataset to answer the questions sets limitation in the study. Improvements could be made by combining data from different sources such as real time audience response reflected from the video.