CSE 5243 Intro To Data Mining
TED Talks Data Analysis - Homework 1

Amanpreet Singh & Sarav Rajavelu

According to Wikipedia, **"TED** (**Technology, Entertainment, Design**) is a media organization which posts talks online for free distribution, under the slogan 'ideas worth spreading'".

This is a dataset of 2500 TED Talks and their attributes up until September 2017. Following attributes are available in the dataset:

TED Main Dataset

- **name:** The official name of the TED Talk. Includes the title and the speaker.
- **title:** The title of the talk
- **description:** A blurb of what the talk is about.
- **main_speaker:** The first named speaker of the talk.
- **speaker_occupation:** The occupation of the main speaker.
- **num_speaker:** The number of speakers in the talk.
- **duration:** The duration of the talk in seconds.
- **event:** The TED/TEDx event where the talk took place.
- **film_date:** The Unix timestamp of the filming.
- **published_date:** The Unix timestamp for the publication of the talk on TED.com
- **comments:** The number of first level comments made on the talk.
- **tags:** The themes associated with the talk.
- **languages:** The number of languages in which the talk is available.
- **ratings:** A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)
- **related_talks:** A list of dictionaries of recommended talks to watch next.
- **url:** The URL of the talk.
- **views:** The number of views on the talk.

TED Transcripts Dataset

1. **url:** The URL of the talk
2. **transcript:** The official English transcript of the talk.

	comments	duration	film_date	languages	num_speaker	published_date	views
count	2550.000000	2550.000000	2.550000e+03	2550.000000	2550.000000	2.550000e+03	2.550000e+03
mean	191.562353	826.510196	1.321928e+09	27.326275	1.028235	1.343525e+09	1.698297e+06
std	282.315223	374.009138	1.197391e+08	9.563452	0.207705	9.464009e+07	2.498479e+06
min	2.000000	135.000000	7.464960e+07	0.000000	1.000000	1.151367e+09	5.044300e+04
25%	63.000000	577.000000	1.257466e+09	23.000000	1.000000	1.268463e+09	7.557928e+05
50%	118.000000	848.000000	1.333238e+09	28.000000	1.000000	1.340935e+09	1.124524e+06
75%	221.750000	1046.750000	1.412964e+09	33.000000	1.000000	1.423432e+09	1.700760e+06
max	6404.000000	5256.000000	1.503792e+09	72.000000	5.000000	1.506092e+09	4.722711e+07

Table 1. Summary statistics of the numeric columns in the main dataset

The Pandas describe function displays descriptive statistics about the numerical attributes in the dataset. The attributes in the dataset are of varied types ranging from integers, text, to arrays.

Comments:

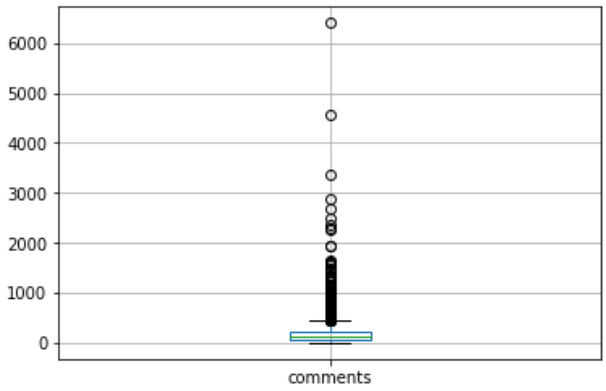

Fig 1. Boxplot of the distribution of comments

The 90th percentile for comments is just **393.09** which suggests that most talks get fewer than 400 comments. Only around 10% talks get more than 400 comments.

	title	main_speaker	views	comments
96	Militant atheism	Richard Dawkins	4374792	6404
0	Do schools kill creativity?	Ken Robinson	47227110	4553
644	Science can answer moral questions	Sam Harris	3433437	3356
201	My stroke of insight	Jill Bolte Taylor	21190883	2877
1787	How do you explain consciousness?	David Chalmers	2162764	2673

Table 2. Talks that received the most comments.

Views:

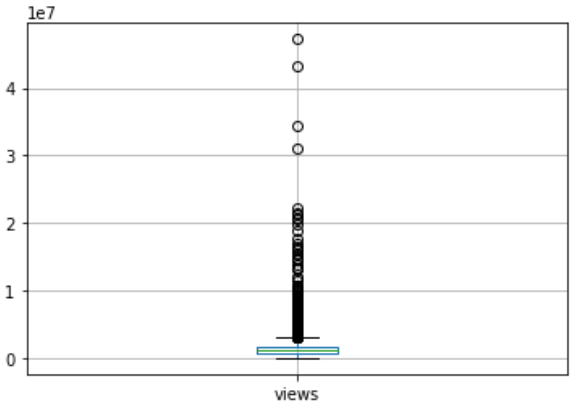

Fig 2. Boxplot of the distribution of view

Similarly, the 90th percentile for views is around 3,051,913 comments which means that **most talks received less than 3.05 million views**. Just 10% talks received more than 3.05 million views with the most viewed talk receiving 47 million views.

	title	main_speaker	views
0	Do schools kill creativity?	Ken Robinson	47227110
1346	Your body language may shape who you are	Amy Cuddy	43155405
677	How great leaders inspire action	Simon Sinek	34309432
837	The power of vulnerability	Brené Brown	31168150
452	10 things you didn't know about orgasm	Mary Roach	22270883

Table 3. Talks that received the most views.

Duration:

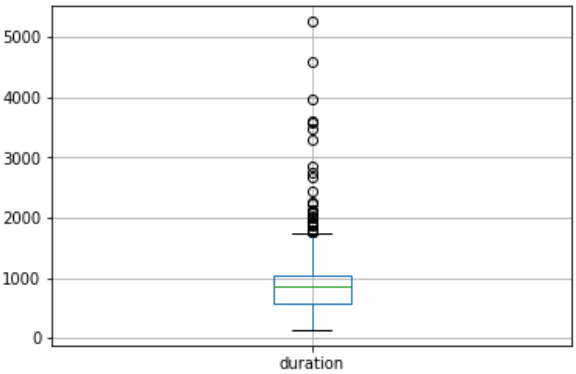

Fig 3. Boxplot of the distribution of duration in seconds

The duration of talks has a better spread with **most talks lasting less than 20 minutes** (90th percentile). The **longest talk went on for 87 minutes**! This talk titled “Parrots, the universe and everything” by Douglas Adams would actually have covered “everything”.

Correlation:

The correlation between attributes of the dataset is presented below:

	comments	duration	film_date	languages	num_speaker	published_date	views
comments	1.000000	0.140694	-0.133303	0.318284	-0.035489	-0.185936	0.530939
duration	0.140694	1.000000	-0.242941	-0.295681	0.022257	-0.166324	0.048740
film_date	-0.133303	-0.242941	1.000000	-0.061957	0.040227	0.902565	0.006447
languages	0.318284	-0.295681	-0.061957	1.000000	-0.063100	-0.171836	0.377623
num_speaker	-0.035489	0.022257	0.040227	-0.063100	1.000000	0.049240	-0.026389
published_date	-0.185936	-0.166324	0.902565	-0.171836	0.049240	1.000000	-0.017920
views	0.530939	0.048740	0.006447	0.377623	-0.026389	-0.017920	1.000000

Table 4. Correlation matrix

Comments and Views:

A moderate correlation of 0.53 is found between comments and views which we expected but on closer inspection higher views doesn’t always mean higher comments.

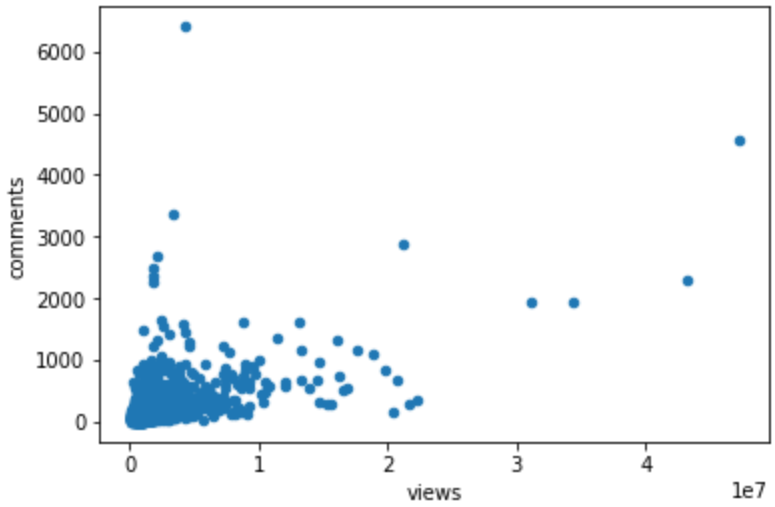

Fig 4. Comments vs Views scatter plot

There doesn’t seem to be any definite correlation between comments and views as is apparent in this plot. Talks viewed a lot might have a few comments and talks not viewed as much might have a lot of comments.

For example, Adam Driver’s talk “My journey from Marine to actor” has been viewed over 3 million times but received just 36 comments. On the contrary, Senator Diane J. Savino’s talk on “The case for

same-sex marriage" has been viewed less than 300,000 times but has received 649 comments which is more than 3 times the number of comments a talk receives on average.
We find that the controversial nature of Sen. Diane J. Savino's talk provokes more interaction in the comments.

Similarly, we observe **that the talk that was most commented on, "Militant atheism" by Richard Dawkins has over 6,400 comments while it has just a tenth of the number of views of the top viewed talk.** The emotion provoking nature of this talk is realized in its description: "Richard Dawkins urges all atheists to openly state their position -- and to fight the incursion of the church into politics and science. A fiery, funny, powerful talk."

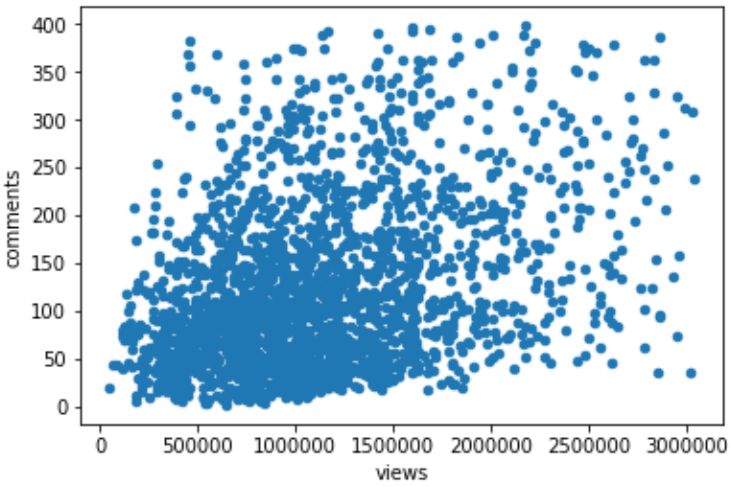

Fig 5. Comments vs Views scatter plot up to the 90 percentile of views and comments

A better representation of the relationship between comments and views can be visualized in the plot above.

Views and languages

We find that according to our expectation, number of views and the number of languages the talk is available in, show a moderately positive correlation of 0.37. **TED Talks with more than 10 million views are available in at least 28 languages.**

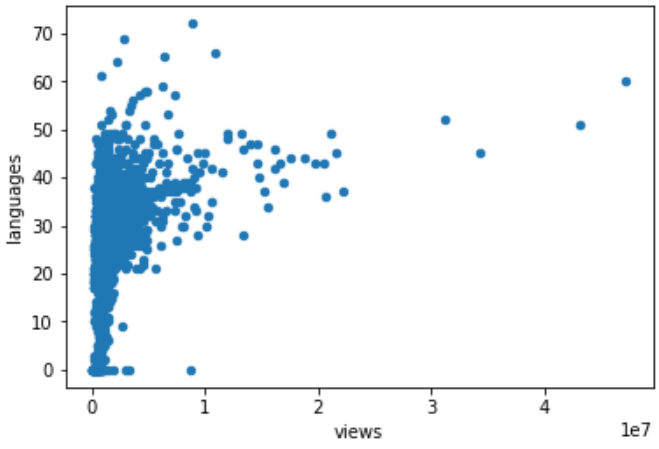

Fig 6. Languages vs Views scatter plot

Ratings:

One of the more interesting attributes of a talk is the ‘ratings’ attribute which provides an estimate of the subjective qualities of a talk.
Users can select any of the following qualities they found in the talk:

- Inspiring
- Informative
- Fascinating
- Persuasive
- Beautiful
- Courageous
- Funny
- Ingenious
- Jaw-dropping
- OK
- Unconvincing
- Longwinded
- Obnoxious
- Confusing

Subsequently, each talk will have a score for each rating quality which shows how many users found the talk funny, inspiring etc.
To find out the funniest talk, the most inspiring talk, the most unconvincing talk etc, we could find the top rated quality amongst all talks. But, this would give us a set of results that are biased towards talks that have been viewed more and hence have been rated more. To remove this bias, **we normalize ratings by the number of views** and find the maximum rating amongst all talks. This gives us an estimate of how funny or how obnoxious each user finds a particular talk.
A few results are enumerated below:

	title	Funny	views
675	Lies, damned lies and statistics (about TEDTalks)	0.002509	2212944
941	Gotta share!	0.002056	357454
194	Juggle and jest	0.002027	807628
1341	An animated tour of the invisible	0.001962	336430
764	This is broken	0.001696	955329

Table 5. The funniest talks

	title	Beautiful	views
972	Building a park in the sky	0.009493	704205
237	"Kounandi"	0.004001	82488
1193	The secret life of plankton	0.001933	197120
209	"M'Bifo"	0.001770	294936
791	A message to gay teens: It gets better	0.001762	278672

Table 6. The most beautiful talks

	title	Fascinating	views
1193	The secret life of plankton	0.002039	197120
1173	Deep ocean mysteries and wonders	0.001686	277544
1044	The divided brain	0.001624	648251
1171	The cockroach beatbox	0.001316	303986
1341	An animated tour of the invisible	0.001296	336430

Table 7. The most fascinating talks

	title	Jaw-dropping	views
108	How PhotoSynth can connect the world's images	0.003086	4772595
148	This is Saturn	0.001892	2627709
117	New insights on poverty	0.001584	3243784
137	Luke, a new prosthetic arm for soldiers	0.001320	1575699
16	Improvising on piano, aged 14	0.001226	1628912

Table 8. The most jaw-dropping talks

	title	Inspiring	views
914	Transplant cells, not organs	0.006747	620231
791	A message to gay teens: It gets better	0.003387	278672
1301	When a reporter becomes the story	0.002923	144044
2171	Why gun violence can't be our new normal	0.002819	1096198
1916	Why we all need to practice emotional first aid	0.002688	4984884

Table 9. The most inspiring talks

We expected some correlation between a few qualities found in the ratings.

We expected positive correlation between:

- Jaw-dropping and Fascinating : 0.49
- Confusing and Unconvincing : 0.57
- OK and Long-winded : 0.59
- Beautiful and Fascinating : 0.18

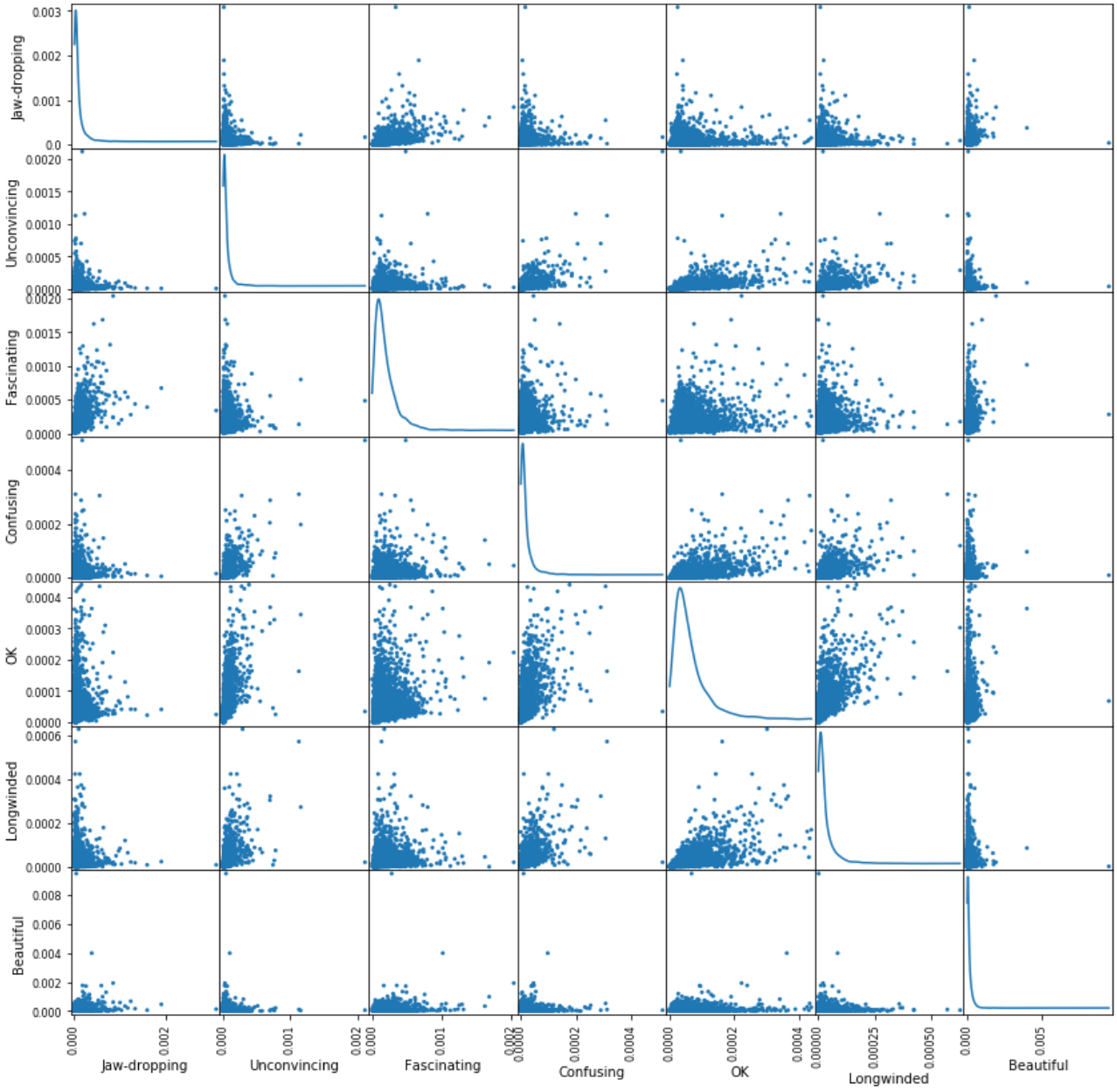

Fig 7. Pairwise correlation plot of select ratings

And a negative correlation between:

- Persuasive and Unconvincing : 0.26
- Informative and Confusing : 0.18
- Funny and OK : 0.25

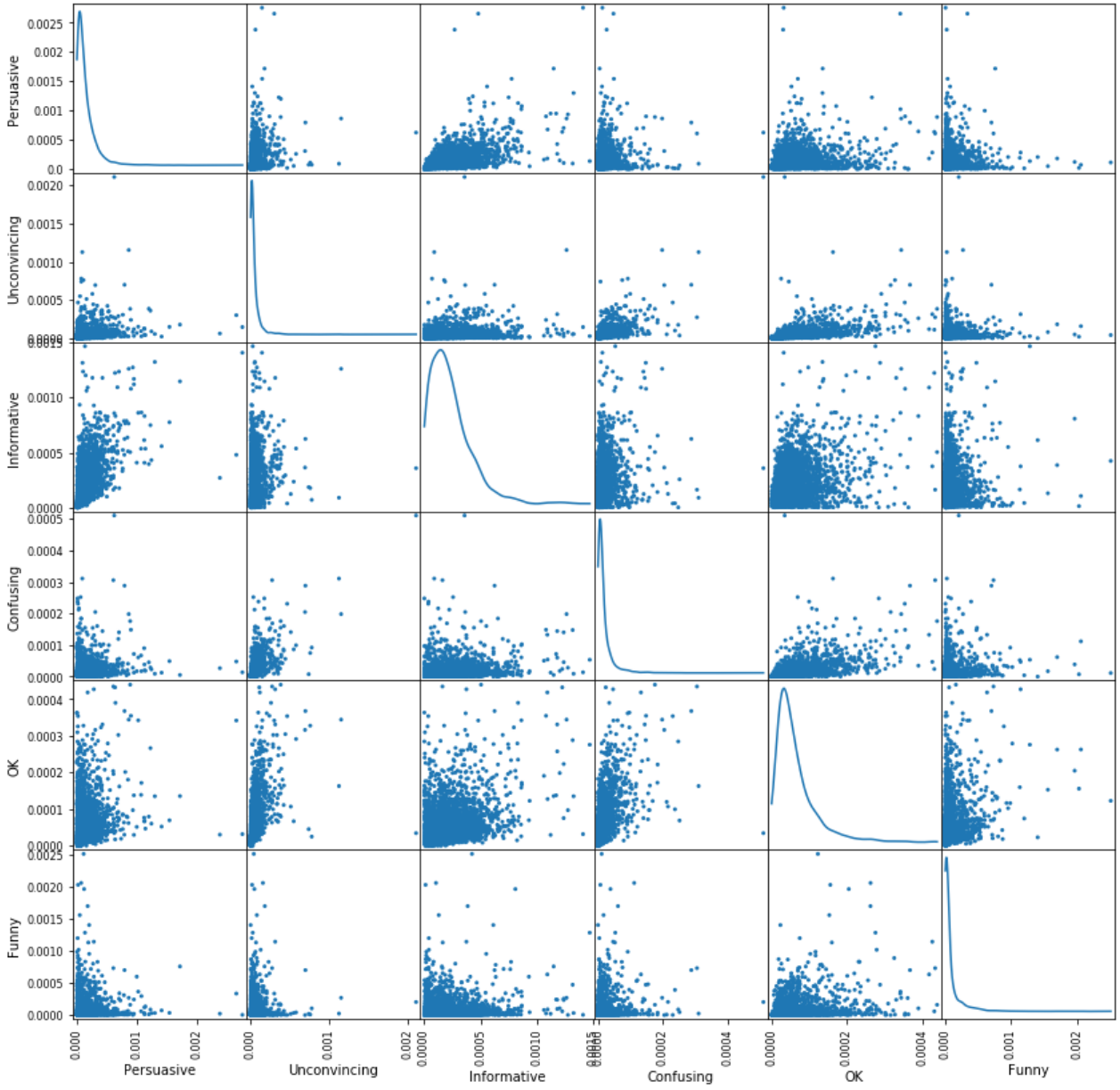

Fig 8. Pairwise correlation plot of select ratings

Although our expectations with positive correlation were met, we were expecting a more significant trend with negative correlation. Especially, **persuasive and unconvincing not having a negative correlation is surprising.**

Tags:

The tag attribute which provides list of all topic the talk pertains to. Looking at how each tag has faired across the years could show a shift in focus of the TED talks.

We start by taking the 10 most popular tags in 2016-17 and looking at their trends across all the years.

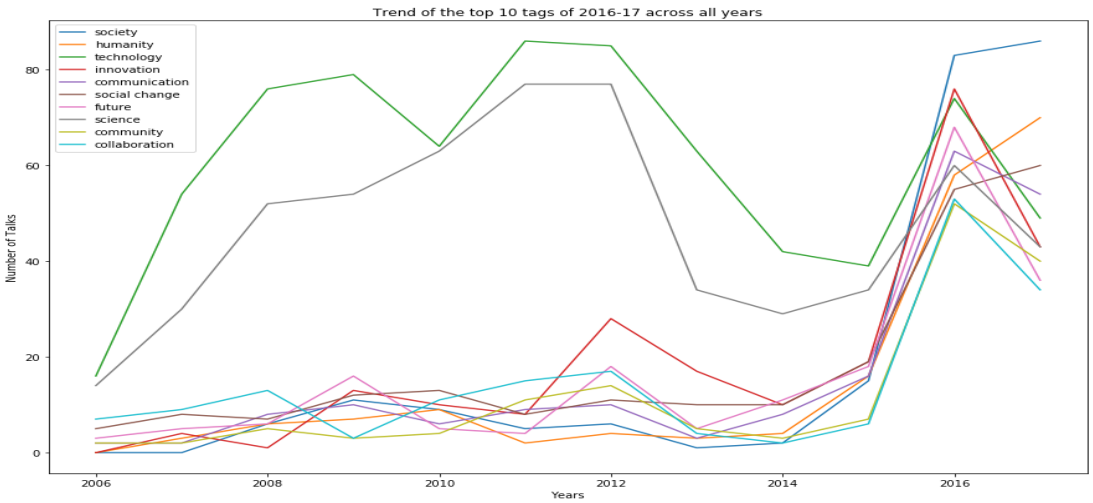

Fig 9. Trend of the Top 10 tags of 2016-17

We notice that two topics, i.e., technology and science have always been popular, whereas topics like social change, society, future and community have only grown in prominence in the last 3 years. There has been an increase in discussion on topic surrounding society and a recent spike in discussions around technology due to the rise of AI, blockchain etc.

We then compare the previous trend with the top 10 most popular tags of all time.
Looking at Fig 10, we see that interest in topics such as culture, entertainment and business has tapered off towards 2017.

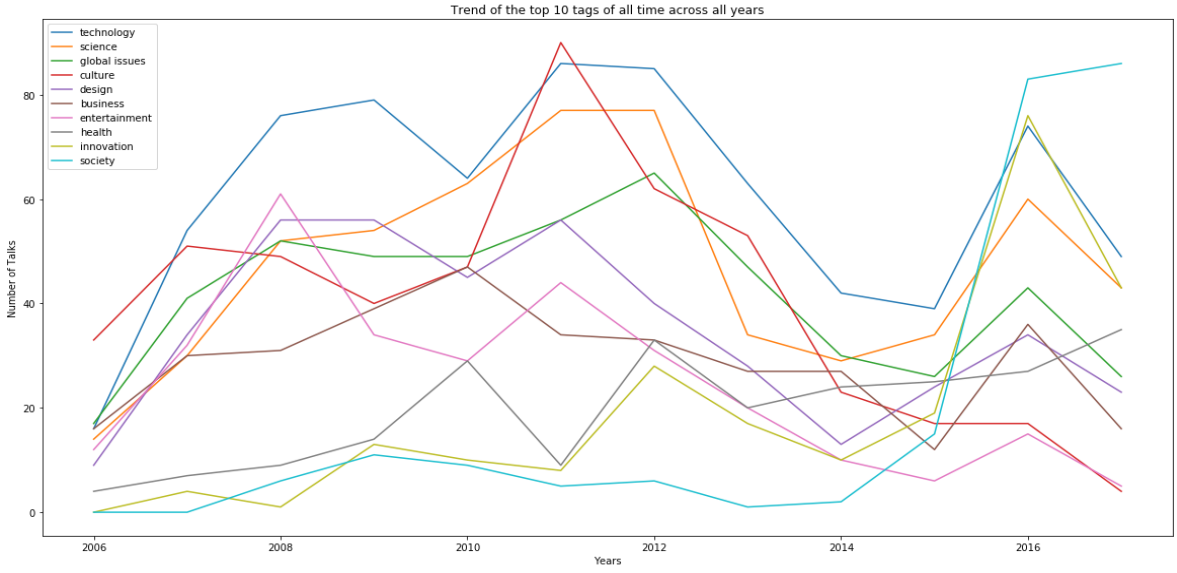

Fig 10. Trend of the Top 10 tags of All-Time

TED Transcript Dataset

The TED Transcript data contains a transcript of the words spoken during the TED talks.

We decided to calculate the words per minute spoken during each talk to see if there was a **correlation between how fast the speakers spoke and the kind of ratings** that the talk received. We're interested in a few particular ratings such as 'Confusing', 'Longwinded', 'Persuasive' that could possibly be selected due to the pace of the speaker.

We calculated the number of words spoken during the talk for each talk, divided by the duration of the talk in minutes. Below is the five point summary of the words per minute (**wpm**) column.

As we can see range of wpm varies from 0.2 to 254. We notice that 0.2 wpm does not appear to be plausible. Checking the transcript for the document we see that it contains only the text ‘(Music)(Applause)‘.

Upon further inspection we notice that there are more such talks that contain no words spoken by the speaker (Talk #304 is of a artist beatboxing).

	wpm
count	2467.000000
mean	153.403986
std	31.926969
min	0.200000
25%	138.138095
50%	155.846154
75%	172.763305
max	254.400000

Table 10. Summary statistics of word per minute (wpm)

	transcript	words
1796	(Music)(Applause)	2
146	(Applause)(Music)(Applause)	3
908	(Music)(Applause)(Music)(Applause)	4
1919	(Music)(Music) (Applause)(Applause)	4
1365	(Mechanical noises)(Music) (Applause)	4
2301	(Guitar music starts)(Cheers)(Cheers)(Music ends)	7
1173	(Music)(Applause)(Music)(Applause)(Music)(Appl...	8
2000	(Guitar music starts)(Music ends)(Applause)(Di...	20
304	Let's just get started here.Okay, just a momen...	21
574	(Music)(Applause)(Music)(Music) (Applause)(Mus...	25

Table 11. Transcripts with no content

Omitting these talks from our wpm analysis and we get a range from 10 to 254 wpm.
We did not find any correlation between the wpm and the votes for each rating category the talk received. We decided to bucket the wpm column. Setting wpm between 140 and 175 as ‘Optimal’ for speech, greater than 160 as ‘Fast’ and the rest as ‘Slow’. To check whether classes of speech pace affected the average proportion of each rating. But we could not find any relation between the pace of speech and how people rated talks.

Document Term Analysis:

We then proceeded to convert the transcripts into a document term matrix of each word. We omitted stop words from the matrix using the ‘stopwords’ defined in Python’s NLTK package.

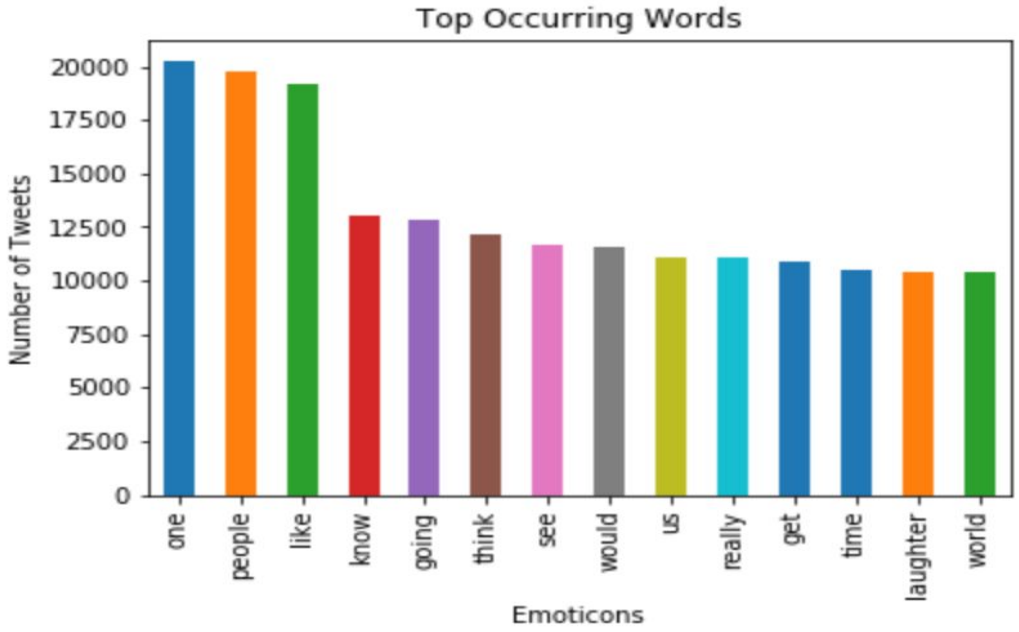

Fig 11. Top occuring words in the transcripts

Audience Engagement:

While parsing through the transcripts we noticed that certain audience responses, like when they applauded or laughed was capture in the transcripts. We expect to find the talks that have the highest audience laughter occurrence with the majority funny vote. (Majority vote for each talk is the rating with the with highest number of votes)

	top_rating	title_y
578	OK	All things are Moleeds
21	Funny	Nerdcore comedy
96	Funny	A comic sendup of TED2006
557	Funny	A one-man world summit
416	Funny	A theory of everything
405	Funny	Cool tricks your phone can do
444	Jaw-dropping	10 things you didn't know about orgasm
2033	Funny	This is what happens when you reply to spam email
0	Funny	Do schools kill creativity?
270	Fascinating	Close-up card magic with a twist
2	Funny	Simplicity sells
331	Persuasive	Tidying up art
191	Jaw-dropping	Juggle and jest
2249	Courageous	A political party for women's equality
715	Funny	Did you hear the one about the Iranian-American?

Table 12. Talks with the most audience laughter

The above list is sorted by decreasing counts of the instances the audience laughed. As we can see 7 out of the top 10 have received the 'Funny' majority vote.
We carried out the same analysis for instances when the audience applauded. However, we found no trend in the top most applauded talks.

Word Cloud:

A visualization to highlight textual data points with most occurences.

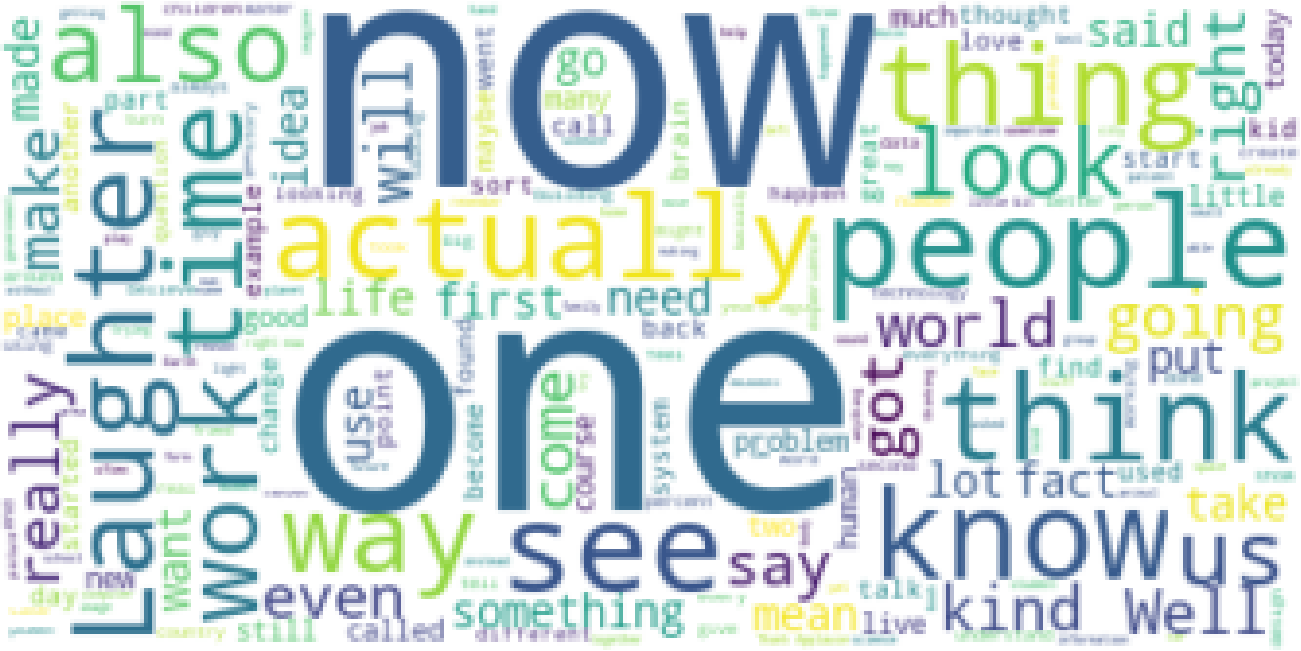

Fig 12. Word cloud