Capstone Project 1 – Milestone Report



***Annapoorani | Springboard | Data Science Career Track***

# Problem Statement

TED talks are very famous in recent years. The ideas spreads through TED talks are amazing. TED talks cover most of the people interested topics includes lifestyle, technology, arts and so on. This project will focus on

* Creation of Recommendation engine for the viewers based on the current selection
* Sentiment Analysis of the talk transcripts
* Predict the ratings of the talks
* Topic Modelling

# Client

The non-profit organization TED will be beneficial by this project. This will help them in the following ways

* Improve user experience by recommendation systems
* As customer response to the talks are analyzed, they can plan the events and marketing based on customer taste
* To come up with new topics of interests.

# Data

TED talk data collected from Kaggle.

* <https://www.kaggle.com/rounakbanik/ted-talks>
* https://www.kaggle.com/goweiting/ted-talks-transcript

The first data contains the details about the talk and the next one is the transcripts and the features from the YouTube.

# Data Wrangling

Success of analysis depends upon how the data is cleaned up as usable for analysis with columns as separate features and each row as single observation. As it is highlighted always as Data Wrangling is time consuming, wrangling for this project also was very challenging.

As mentioned in Data section the goal is to combine TED data and YouTube data for analysis.

## **Main challenges** in merging YouTube data and TED DATA:

* There is no any common field like Video ID.
* The details are only the title name and speaker names.
* Titles are not exactly alike in both dataset.
* As there is a chance that one speaker delivered more than one titles, we cannot match only with speaker names, so merging based on titles is the best bet.
* The format of the title is completely different, TED data contains title alone, but YouTube data have 'title|speaker' or 'speaker|title' as formats.

## STRATEGY:

* YouTube likes and dislikes contains **NA values**. As those talks didn’t have significant number of views they are **replaced by zeros**.
* The **Columns are renamed** for better usage and unwanted columns are dropped.
* Using **pandas. series. String** functions like **strip, replace and concatenate** the texts in the titles are cleaned.
* As YouTube have no separate columns for title, speaker and title are separated into new columns using **merge and split** functions.
* First titles with exact match of words are matched by merging based on TED and YouTube titles.
* Second rows of speakers with only one talk are filtered and merged based on speaker names.
* But the real hurdle was merging the titles of same talks but described with different words, so using **nltk package the words are tokenized**.
* To find the similarity between words**, cosine similarity** which is popular to match similar words with good degree of accuracy is used to merge based on similarity values.

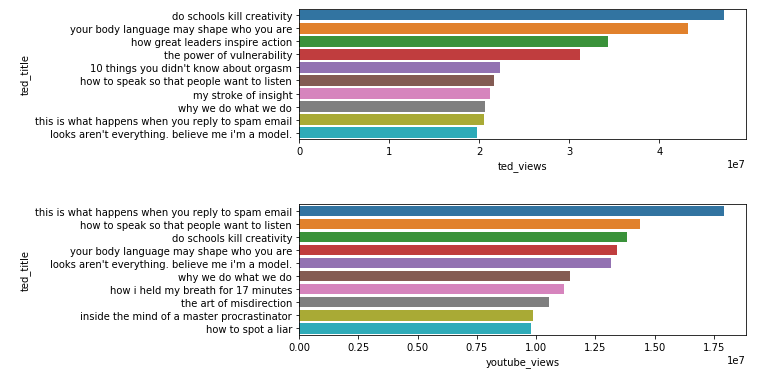
***similarity (doc1, doc2) = cos(θ) = doc1doc2 / |doc1||doc2***

* Then all the data frames are merged into single data frame and exported to csv file.
* Each Talk have several ratings described in TED data like Beautiful, Long Winded, OK, Inspirational etc. Same ways they have different tags in YouTube data. So, they are separated using **Counters applied using iterrows** and stored as separate data frames for further analysis.

# Exploratory Data Analysis

Pictures worth Thousand words. EDA helps in visualizing data in different angles.

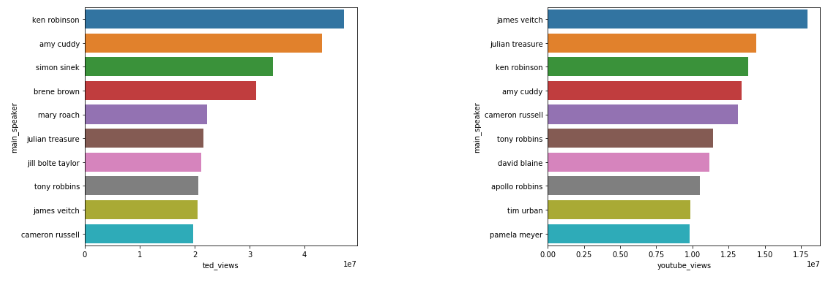
### TED Title Winners



The above plot gives us the top viewed TED talks in TED.com and YouTube.

* Interestingly the ranking orders are not same in both websites (YouTube and TED). Also, the number of views also differs
* Six talks from top TED views are also listed in top YouTube Views category. So those talks are clear winners.
* But the number of views clearly represents the mission of TED **'Ideas Worth Spreading'**

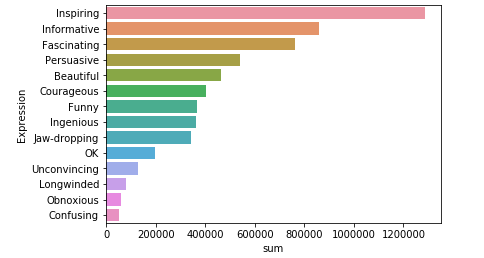
### Best Speakers



TED Speakers are equally popular as TED Talks.

* **Ken Robinson, Amy Cuddy, Simon Sinek, Julian Treasure, James Veitch and Cameron Russell** are top speakers in both TED.com and YouTube.
* But the order differs greatly in YouTube. James Veitch's 'this is what happens when reply to spam email' is mainly attract all the category of people throughout the world as this may be also the question searched on YouTube, that gives edge.

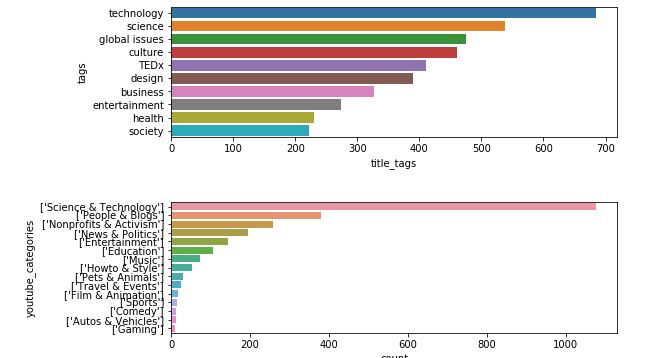
### Impact of TED Talk



TED allows viewers to rate the talk not by numbers but by words. Viewers are allowed to choose three words to express their response to talk. Above plot gives us the top words chosen by users to rate the videos.

* It seems most of the viewers wants the talk to be **Inspiring**
* There is also clear border between the positive and negative expressions. As **OK, Unconvincing, Longwinded, Obnoxious, Confusing** are not kind of good comments are only in lesser numbers compared to positive comments.
* By above two points most of the TED talks are **best**, also people are not willing to express their negative feedback.

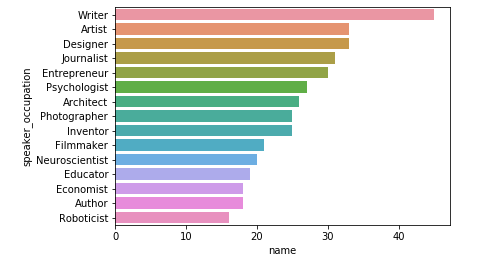
### Exploring Categories



TED segregated their videos based on tags and YouTube based on Categories.

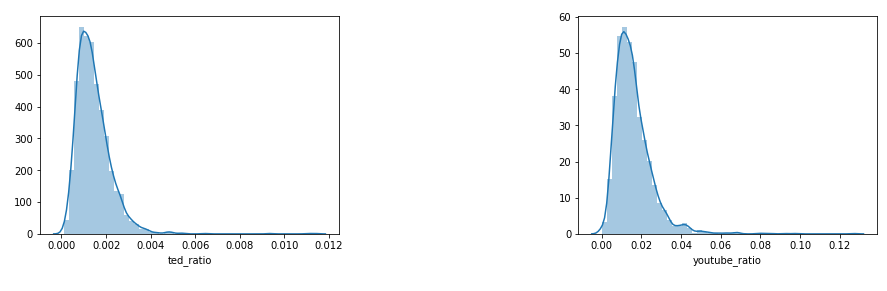
* No of Talks tagged as Technology and Science are large numbers, same as in Science and Technology Categories.
* TED organizes more talks in Science & Technology and Global issues

### Diverse Occupation of speakers



**Writers** performed most of the Talks in TED events. As writers are the best with words also known for their charming ways to attract readers they topped the list.

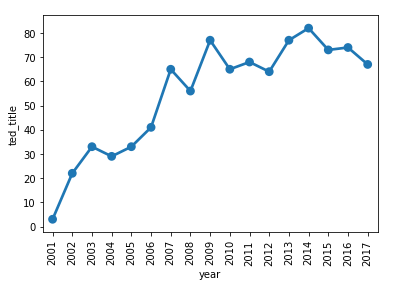
### Viewers willingness to comment

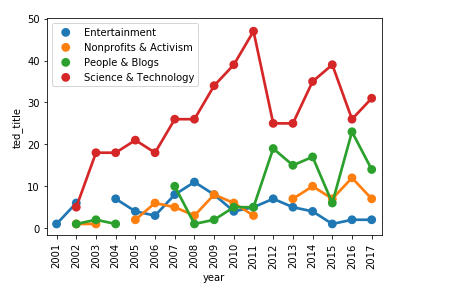


Though the count of views in TED.com and YouTube touches millions, but the number of people ready to comment or rate is very less.

* From above plots No of people commented vs Total views in **TED.com is only 0.2 % whereas YouTube is 2%**
* More comments from people will give the clear picture on the feedback of the Talks.

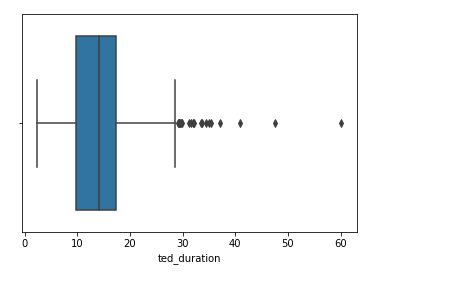
### TED Talks growth over years





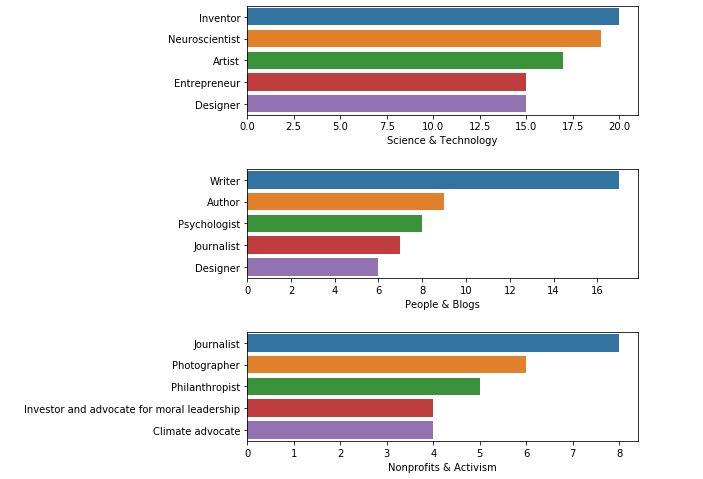
* **Number of talks was increased steadily from 2001 to 2007**, after that number of talks remain between 60 - 80.
* Considering YouTube categories spread over the years of TED talk, **Science & Technology** remains the top number of talks
* In **TED 2011** event **Science and Technology talks are high** but there is significant reduction in other categories.
* On contrary **2012 and 2013** have very less number of technology talks, but increase in **People & Blogs** category

Duration Matters



Most of the TED talk duration are between 9 minutes to 18 minutes.

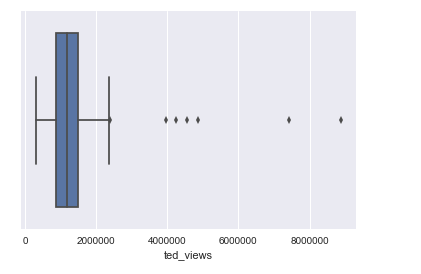
### Top speaker occupation among Categories



### Among the top Categories we explored different speaker occupations

### **Inventor** tops in **Science & Technology** Category, without any surprise **Writers tops People & Blogs**. Who can talk well on **Activism is our Journalists** and about **Non-profits is Philanthropists**

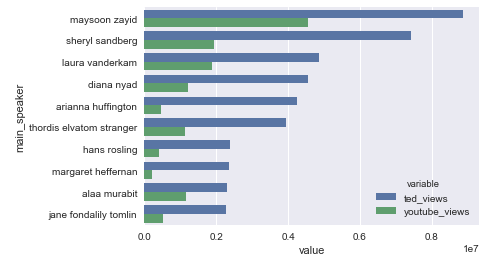
# Great response for TED Women



The distribution on TED Women views are shown above

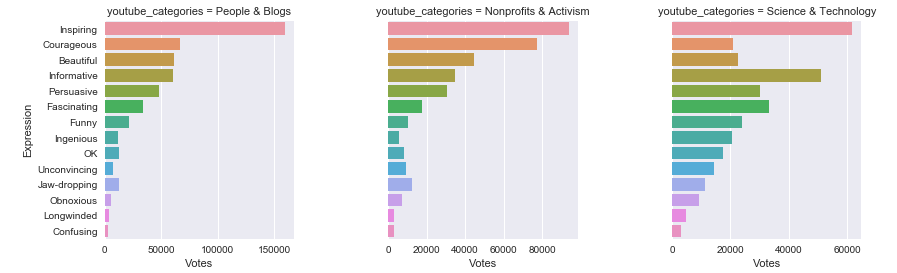
* Most of the TED Women talks have 1 Million to 1.8 Million. The range is pretty good as the count of views is consistent among all the talks

### Best Women Speakers

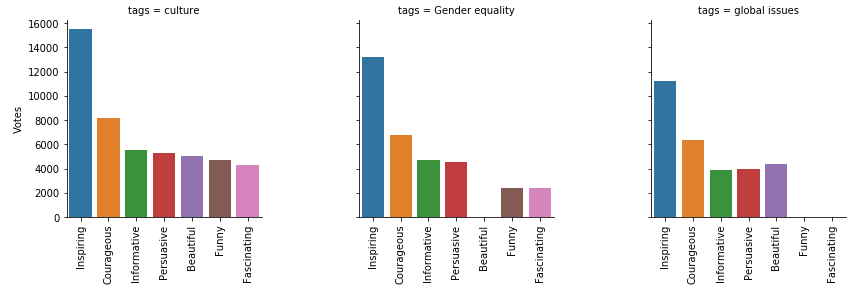


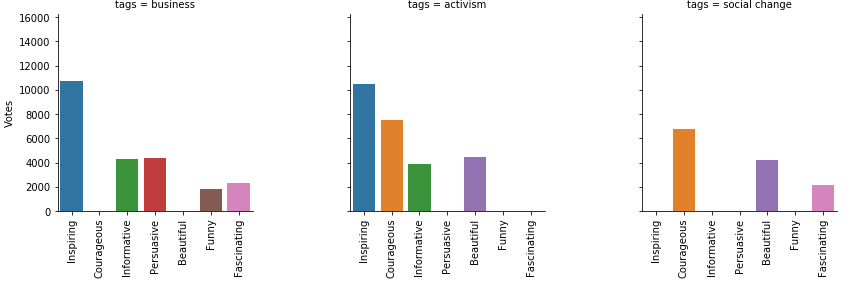
* **Maysoon Zayid, Sheryl Sandberg, Laura Vanderkam and Diana Nyad are the top speakers both in TED.com and YouTube**. These are the top talks overall.
* Interestingly fact is that usually from our above plot usually **Youtube have more views that TED.com, but for TEDWomen talks TED.com have more views that Youtube.**
* Youtube have only **half the views** as compared to TED.com

### Top 3 Categories are Science and Technology, Non-Profits & Activism and People & Blogs



* Even though Science & Technology category have more talks, there are more **positive votes for People and Blogs Category.**
* Most people felt the **Science & Technology category as Informative** equivalent to Inspiring. **Negative expressions are greater for Science & technology category compared to other two categories**.
* **People & Blogs have 23 talks lesser than other two but yet this category is clear winner.**



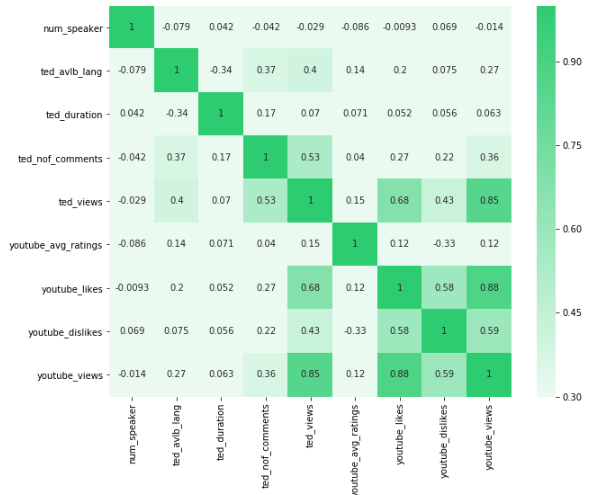


Considering top tags and analyzing the ratings of each

* As most of the Women expects inspirational and bold Women leaders as their role models, it is very clear from above in all the tags Inspiration and Courage remains the top ratings.
* This is TEDWomen talk so we ignored the tag 'Women' as it will occur in all these talks

# Statistical Analysis

Numbers speaks correlation



Heatmap is used to find correlations between the numerical variables in Data frame

* **Youtube Likes are highly correlated with YouTube Views and TED views**
* Talks available in several **languages** show little edge in **number of views**
* Number of speaker is not correlated with any other variables.