Capstone Project 1 – Final Report

ANNAPOORANI | SPRINGBOARD | DATA SCIENCE CAREER TRACK

Table of Contents

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.

$$similarity\ (doc1, doc2) = cos(\theta) = \quad doc1doc2\ /\ |doc1||doc2$$

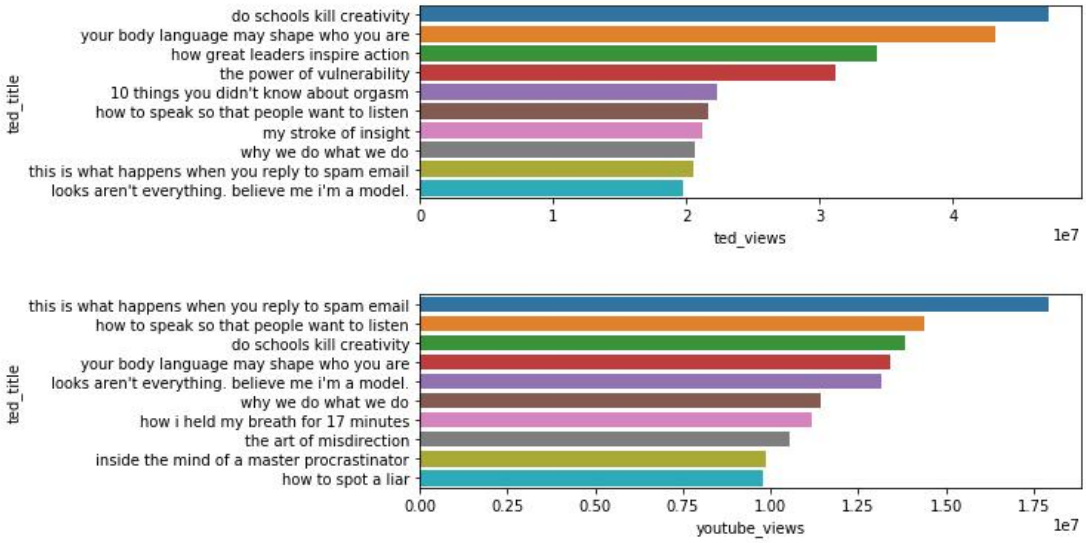
- Then all the data frames are merged into single data frame and exported to csv file.
- Each Talk have several ratings described in TED data like Beautiful, Long Winded, OK, Inspirational etc. Same ways they have different tags in YouTube data. So, they are

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.

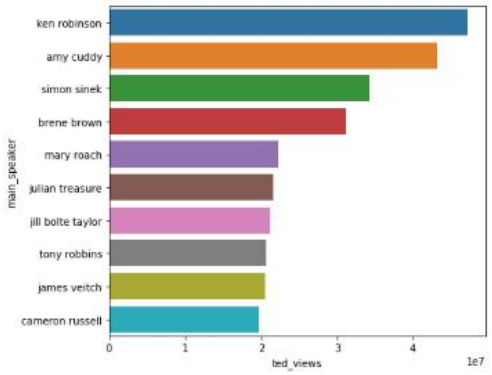
TED Title Winners

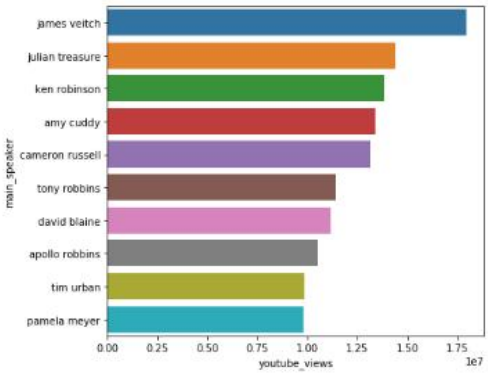


The above plot gives us the top viewed TED talks in TED.com and YouTube.

- 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

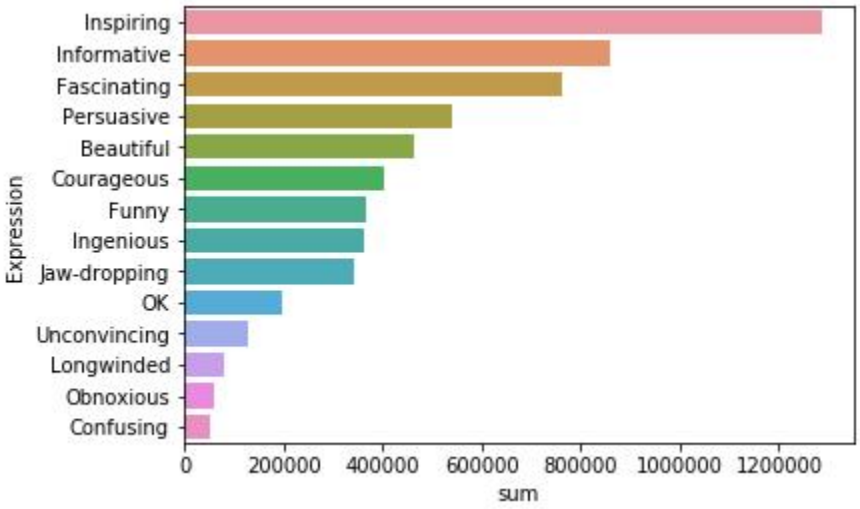




TED Speakers are equally popular as TED Talks.

- **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

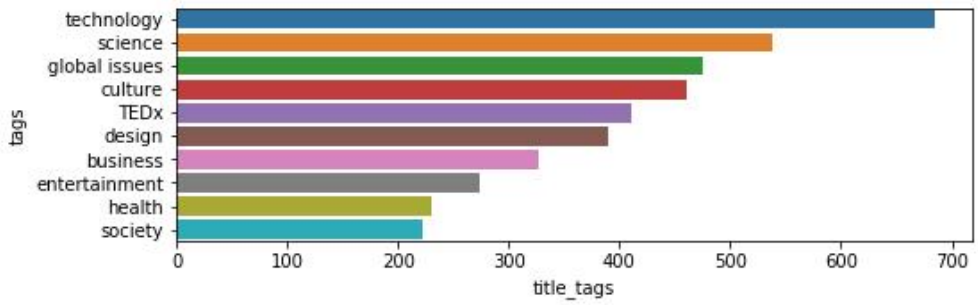


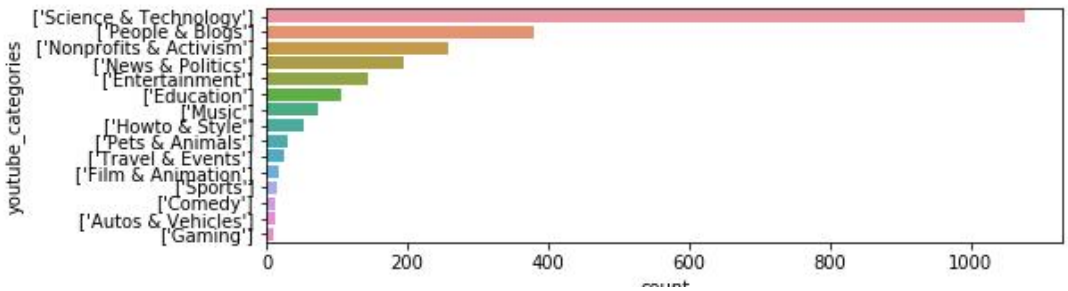
TED allows viewers to rate the talk not by numbers but by words. Viewers are allowed to choose three words to express their response to talk. Above plot gives us the top words chosen by users to rate the videos.

- 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

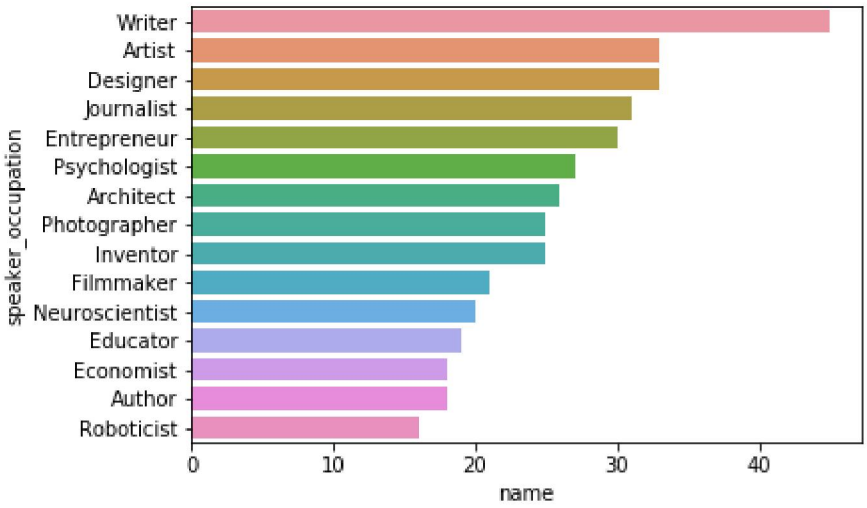




TED segregated their videos based on tags and YouTube based on Categories.

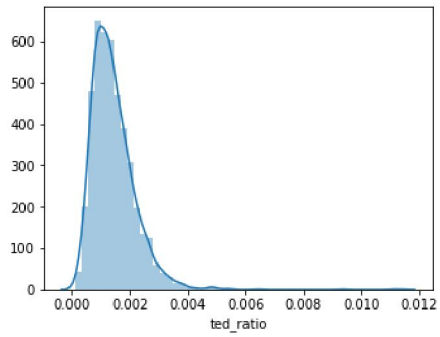
- No of Talks tagged as Technology and Science are large numbers, same as in Science and Technology Categories.
- TED organizes more talks in Science & Technology and Global issues

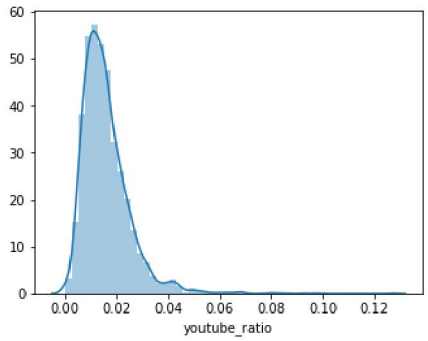
Diverse Occupation of speakers



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

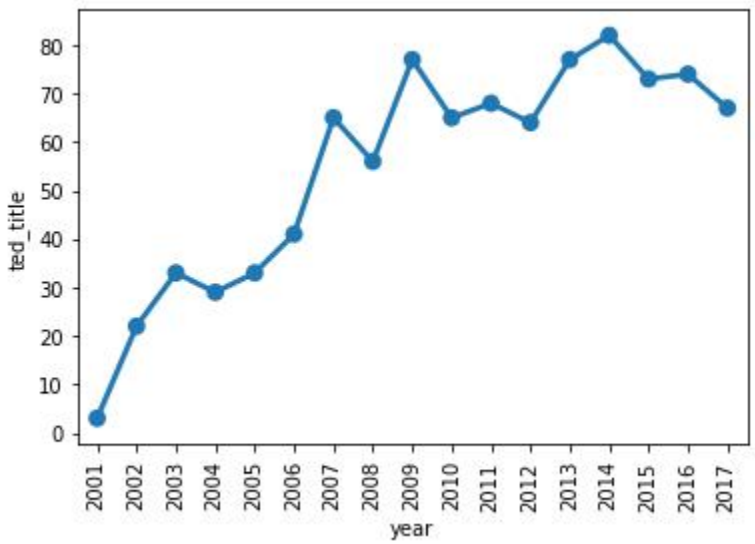


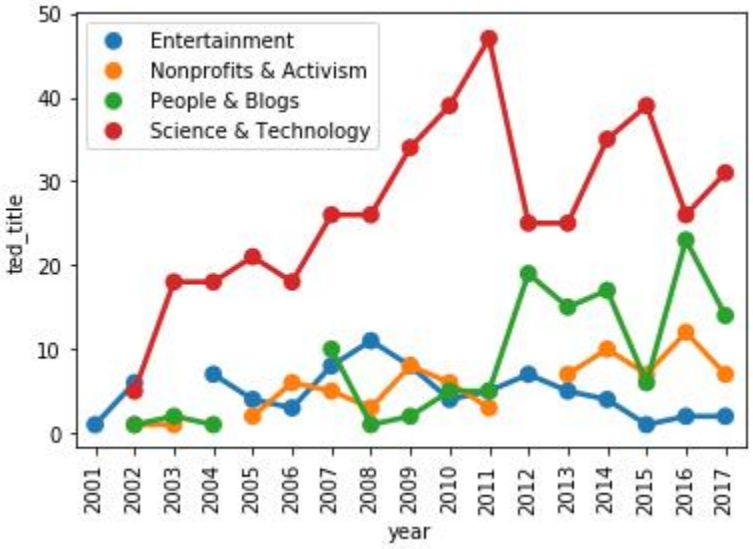


Though the count of views in TED.com and YouTube touches millions, but the number of people ready to comment or rate is very less.

- 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.

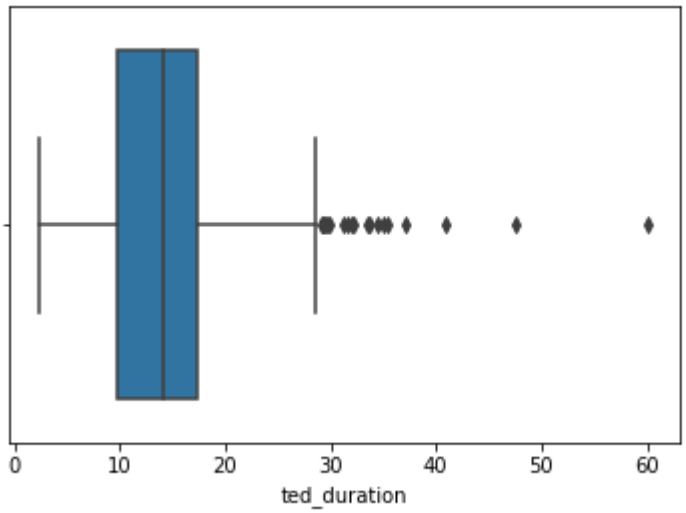
TED Talks growth over years





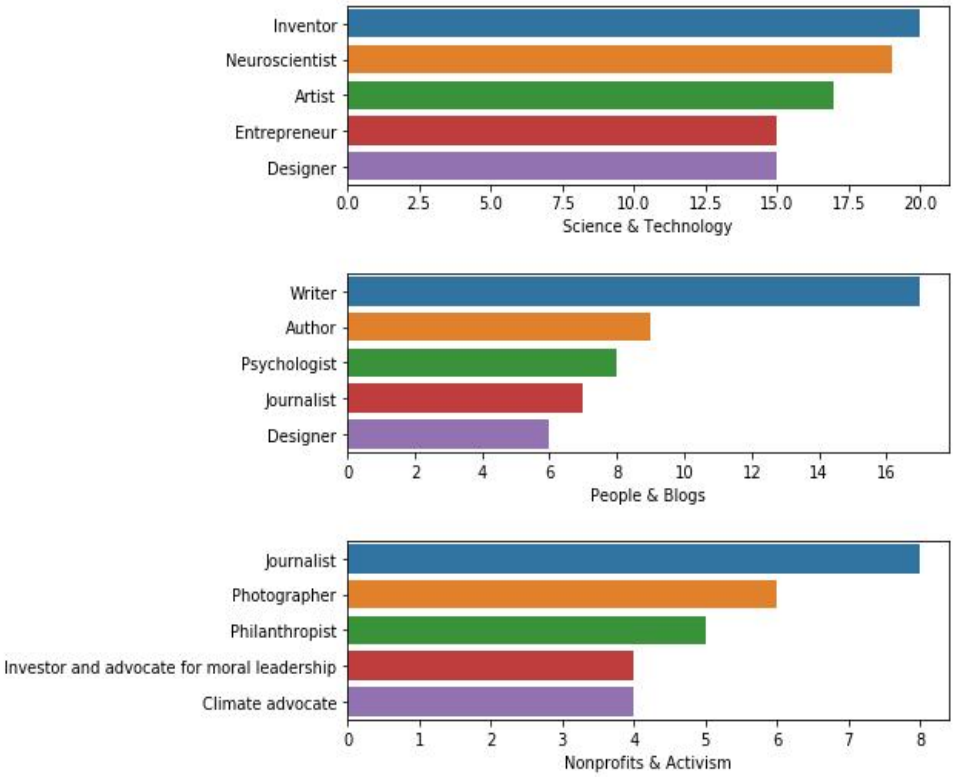
- **Number of talks was increased steadily from 2001 to 2007**, after that number of talks remain between 60 - 80.
- Considering YouTube categories spread over the years of TED talk, **Science & Technology** remains the top number of talks
- In **TED 2011** event **Science and Technology talks are high** 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.

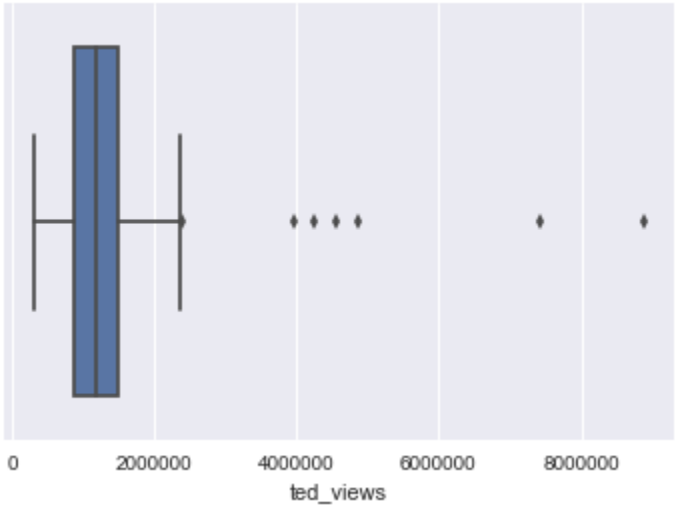
Top speaker occupation among Categories



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**

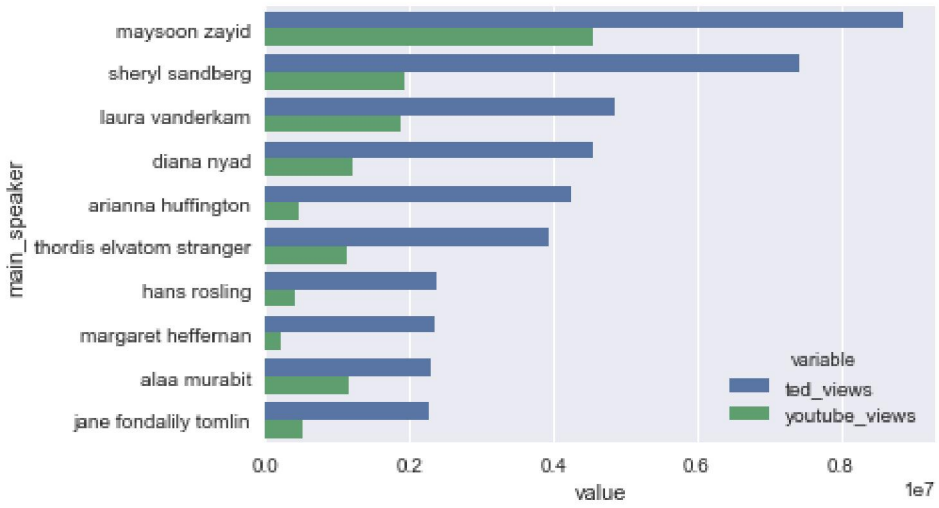
Great response for TED Women



The distribution on TED Women views are shown above

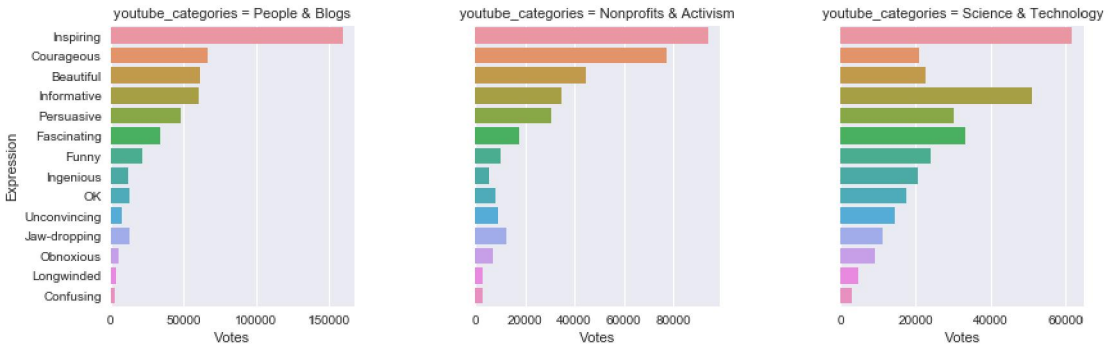
- Most of the TED Women talks have 1 Million to 1.8 Million. The range is pretty good as the count of views is consistent among all the talks

Best Women Speakers



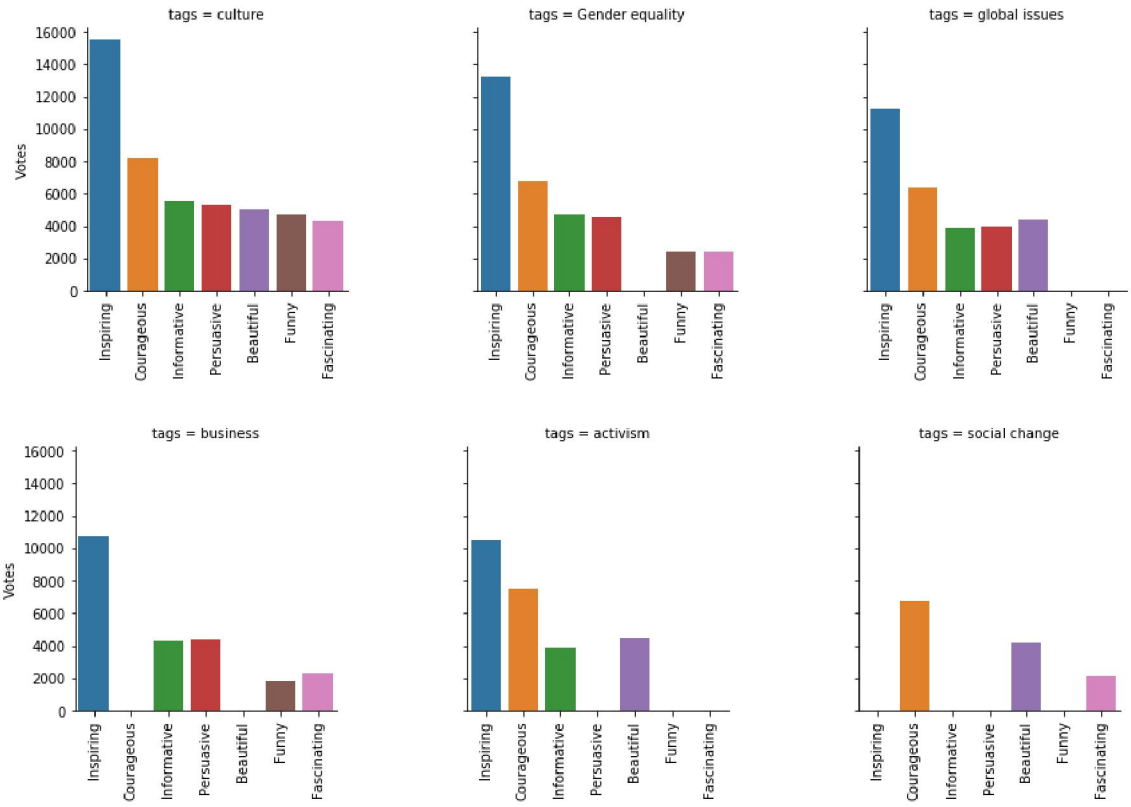
- **Maysoon Zayid, Sheryl Sandberg, Laura Vanderkam and Diana Nyad are the top speakers both in TED.com and YouTube**. These are the top talks overall.
- Interestingly fact is that usually from our above plot usually **Youtube have more views that TED.com, but for TEDWomen talks TED.com have more views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

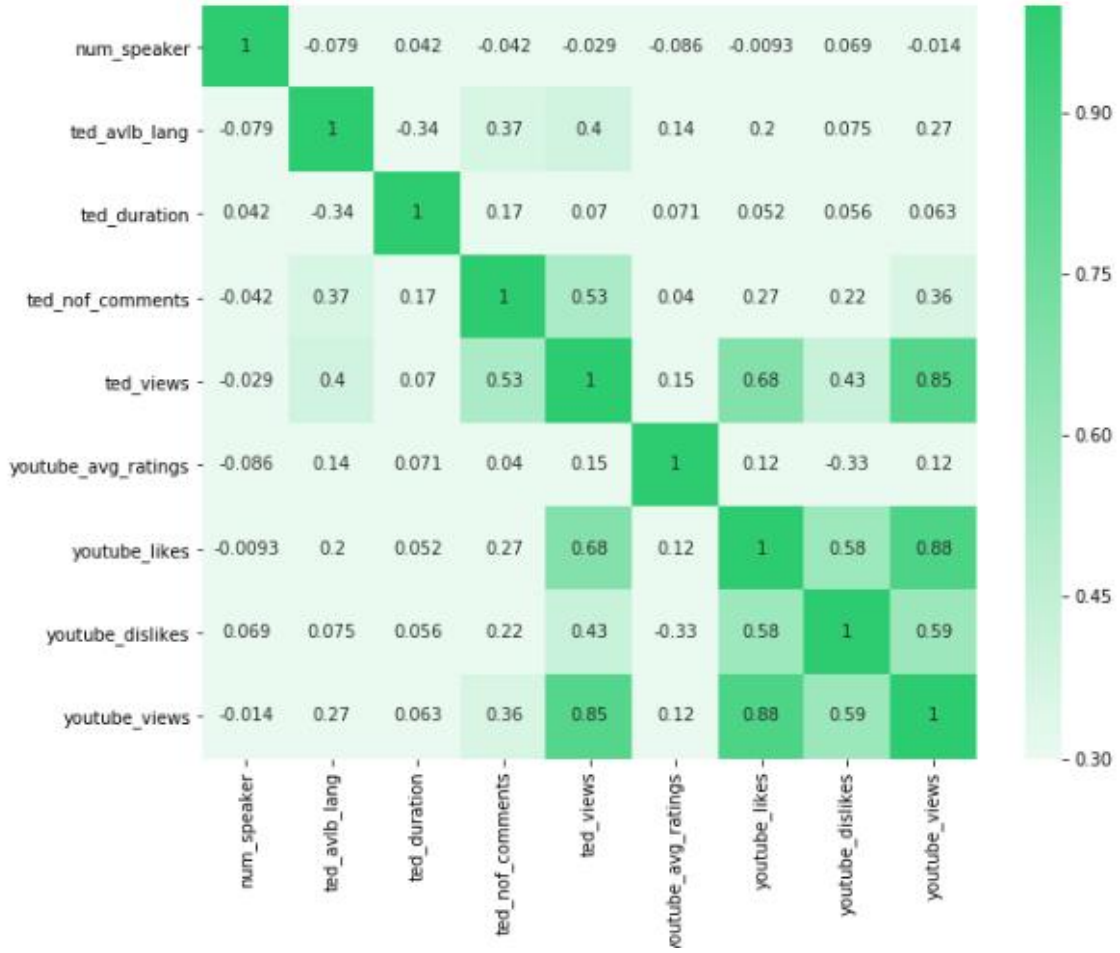


Considering top tags and analyzing the ratings of each

- 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 speakers is not correlated with any other variables.

Analysis of TED Text Scripts

NLP is a branch of data science that consists of systematic processes for analyzing, understanding, and deriving information from the text data in a smart and efficient manner. By utilizing NLP and its components, one can organize the massive chunks of text data, perform numerous automated tasks and solve a wide range of problems such as – automatic summarization, machine translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation etc.

Kaggle TED Data Set contains the scripts of all Talks. By applying NLP techniques, we can uncover hidden features behind the text that gives us a way for

- Recommendation systems
- Topic Modelling
- Clustering.

TEXT PREPROCESSING:

As Data Wrangling is an important step in Data Science process, similar way text pre-processing is required for any text analysis. They are as follows

- Removing Accented characters
- Expanding Contractions
- Removing Special Characters
- Removing Stop Words
- Lemmatization
- Stemming
- Removing unnecessary White spaces

Removing Accented Characters

Accented Characters are the characters have accents or symbols above them. Replacing them with normal characters is important before analysis. Examples of accented characters are á, à, â, é, è, ê, í, ì, î, ó

Expanding Contractions

Contractions are common in English Language. Contractions are like aren't, isn't, they've, they're. For Semantic analysis expanding them help to identify the negation effects in text and negative sentiments.

Removing Special Characters

TED Scripts are the text from a talk, there is a chance for several special characters. Removing the special characters are essential for further analysis.

Lemmatization

Lemmatization consider the morphological forms of words. Example lemmatization consider 'studies', 'studied ', 'studying' are considered as the root word 'study'. This helps to identify all these words as a single word and finding frequency based on them, instead of considering each as separate identity.

Removing White Spaces

While typo there are lot of chances for having unwanted white spaces. Sometimes words with and without spaces are considered as different words

Removing Stop words

Text contains stop words like 'the', 'an', 'he', 'is', 'was' those words are just fillers, but when analyzing sentiments those words does not have any impact so those can be removed. For current analysis the words 'no' and 'not' are removed from stop words list as they will clearly identify the negative sentiments.

TOKENIZATION:

TText needs to be converted to numbers before we apply machine language algorithms to it. Basic models are Bag of words, converting each text to a number and counting the number of occurrences.

tf-idf stands for **Term Frequency - Inverse Document Frequency**. Term frequency gives the frequency of the word in each document. It is the ratio of number of times the word appears in a document compared to the total number of words in that document. Inverse Document Frequency used to calculate the weight of rare words across all documents in the corpus.

Recommendation Systems

Recommendation systems is a well-known feature in all websites. Netflix, YouTube and Amazon are famous examples. Suggestion can be made based on the similarity. Hence, we need to find the similarity between the scripts to build the similarity system. There are several similarity measures available most prominent are **Jaccard, Cosine, Euclidean distance and Manhattan distance.**

Cosine similarity metric finds the normalized dot product of the two attributes. By determining the cosine similarity, we would effectively try to find the cosine of the angle between the two objects. The cosine of 0° is 1, and it is less than 1 for any other angle.

It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

```
from sklearn.metrics.pairwise import cosine_similarity
similar =cosine_similarity(matrix)
similar_df = pd.DataFrame(similar)
```

By means of above code similarity matrix are created for each transcript. As mentioned above similarity value of 1 denotes more similar the texts are.

```
#data['title','similar_talks'][12]

print ("The recommended talks for title: {} are \n\n {} ".format(data['title'][12],data['similar_talks'][12]))
```

```
The recommended talks for title: My wish: Help me stop pandemics are

 HIV and flu -- the vaccine strategy,  Lessons from the 1918 flu,  How we'll stop polio for good,  The case for optimism
```

```
print ("The recommended talks for title: {} are \n\n {} ".format(data['title'][1],data['similar_talks'][1]))
```

```
The recommended talks for title: Averting the climate crisis are

 Design and discovery,  A one-man world summit,  A climate solution where all sides can win,  New thinking on the climate crisi
s
```

From the above results e.g. **for the talk "Help me stop pandemics are" the recommendations are "HIV and flu – Vaccine strategy", "Lessons from the 1918 flu" and "how we'll stop polio for good". The results are very convincing same for the climate crisis talk.**

Topic Modelling

Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.

LDA and EMF are two well-known examples of topic models. Most of the times EMF gives better results than LDA.

LDA

Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. It uses two probability values: P (word | topics) and P (topics | documents).

Applying LDA with 10 topic model and extracting top 5 words the results are

```
{0: ['women', 'brain', 'music', 'data', 'water'],
 1: ['god', 'book', 'building', 'creativity', 'writing'],
 2: ['ca', 'language', 'ok', 'community', 'audience'],
 3: ['universe', 'stars', 'earth', 'planet', 'space'],
 4: ['song', 'oh', 'music', 'film', 'yeah'],
 5: ['god', 'force', 'education', 'push', 'oh'],
 6: ['design', 'ok', 'designers', 'building', 'music'],
 7: ['happiness', 'fuel', 'happy', 'design', 'waste'],
 8: ['news', 'god', 'answers', 'google', 'dollars'],
 9: ['music', 'ends', 'starts', 'africa', 'black']}
```

NMF

Non-negative Matrix Factorization (NMF) that strongly resembles Latent Dirichlet Allocation (LDA) which we covered in the previous section, Whereas LDA is a probabilistic model capable of expressing uncertainty about the placement of topics across texts and the assignment of words to topics, NMF is a deterministic algorithm which arrives at a single representation of the corpus. For this reason, NMF is often characterized as a machine learning algorithm. Like LDA, NMF arrives at its representation of a corpus in terms of something resembling “latent topics”.

Applying NMF with 10 topic model and extracting top 5 words the results are

```
{1: ['god', 'book', 'stories', 'oh', 'art'],
 2: ['music', 'play', 'sound', 'song', 'ends'],
 3: ['women', 'men', 'girls', 'woman', 'sex'],
 4: ['brain', 'brains', 'cells', 'body', 'activity'],
 5: ['water', 'earth', 'planet', 'ocean', 'species'],
 6: ['countries', 'africa', 'government', 'global', 'dollars'],
 7: ['cancer', 'cells', 'patients', 'disease', 'cell'],
 8: ['kids', 'children', 'education', 'students', 'teachers'],
 9: ['city', 'design', 'cities', 'building', 'buildings'],
 10: ['data', 'information', 'computer', 'machine', 'internet']}
```

COMPARISON RESULTS OF LDA AND NMF

Considering an example performing **LDA** results in

```
Topic distribution for document #8:
 [[0.93190255 0.00756637 0.00756637 0.00756637 0.00756637 0.00756637
  0.00756637 0.00756637 0.00756637 0.00756652]]
Relevant topics for document #8:
 [0]

Transcript:
 It's wonderful to be back. I love this wonderful gathering. And you must be wondering, "What on earth? Have they put up the wr
ong slide?" No, no. Look at this magnificent beast, and ask the question: Who designed it?This is TED; this is Technology, Ente
rtainment, Design, and there's a dairy cow. It's a quite wonderfully designed animal. And I was thinking, how do I introduce th
is? And I thought, well, maybe that old doggerel by Joyce Kilmer, you know: "Poems are made by fools like me, but only G ...
```

Considering an example performing NMF results in

```
Topic distribution for document #8:
 [[0.06924094 0.00939016 0.         0.0490575  0.02995617 0.00534906
  0.         0.03283779 0.01871856 0.01609445]]
Relevant topics for document #8:
 [0 3 4 7 8 9]

Transcript:
 It's wonderful to be back. I love this wonderful gathering. And you must be wondering, "What on earth? Have they put up the wr
ong slide?" No, no. Look at this magnificent beast, and ask the question: Who designed it?This is TED; this is Technology, Ente
rtainment, Design, and there's a dairy cow. It's a quite wonderfully designed animal. And I was thinking, how do I introduce th
is? And I thought, well, maybe that old doggerel by Joyce Kilmer, you know: "Poems are made by fools like me, but only G ...
```

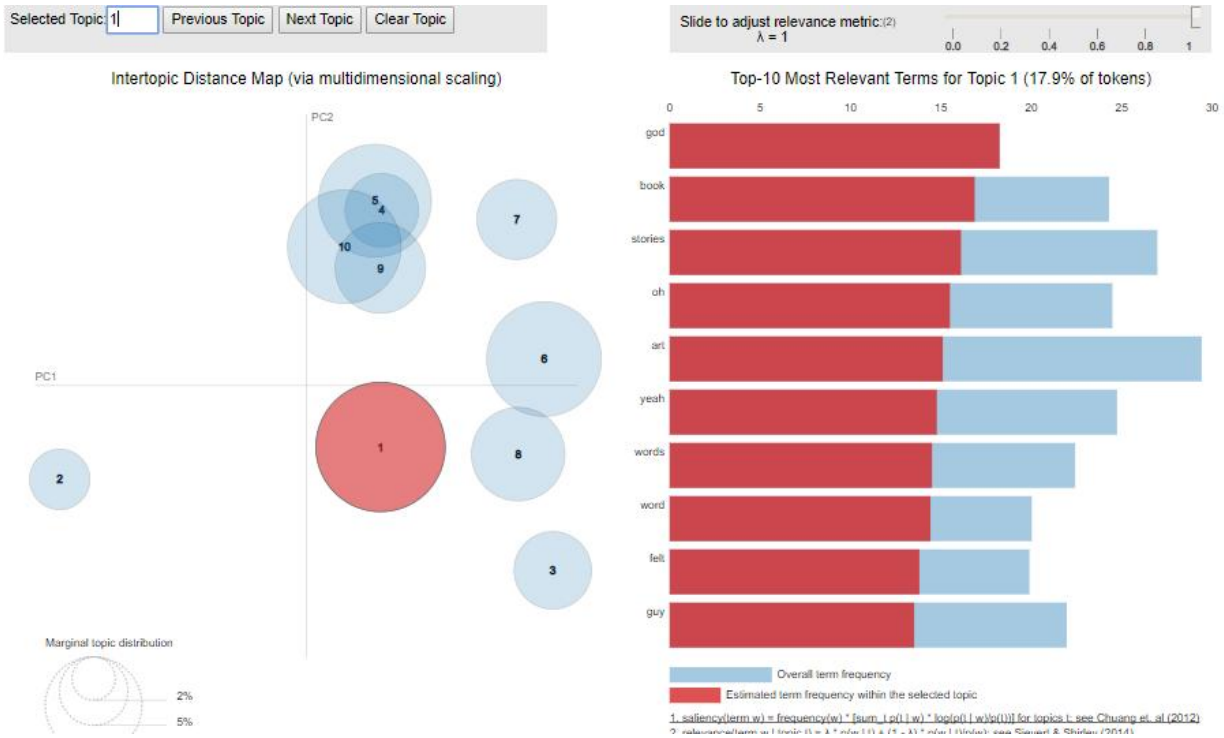
According to NMF relevant topics for the above talk are 0,3,4,7,8,9. The **0th topic according to NMF is 'god', 'book', 'art', 'stories' and 3rd topic is 'brain', 'cell', 'body', 'activity'** which are relevant to the tags given in **tags** column **'God', 'atheism', 'brain'**.

According to LDA relevant topic for above talk is **topic 0 that is 'women', 'brain', 'music', 'data', 'water' but actual tags are 'God', 'atheism', 'brain'**

So **NMF performs better than LDA**

TOPIC MODELLING VISUALIZTION

pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an **interactive web-based visualization.**



From the above visualization the specific topics can be chosen to identify the top topics. Also, the bubble diagram shows the clusters of topics and how closely they are related. The **cluster number 2 is related to topics like music, sound, song and videos that's why it stood out. The Cluster 6 and 8 have overlap since cluster 6 topics are global, countries, economy, social etc. and cluster 8 are kids, children, education, food.**

Clustering

Clustering is the grouping of objects together so that objects belonging in the same group (cluster) are more similar to each other than those in other groups (clusters). There are two best-known clustering approaches **K-Means and Hierarchical Clustering.**

K-MEANS

By K-Means Clustering a data set can be segregated into K distinct, non-overlapping clusters. K needs to be decided before the algorithm application.

Let C1,...,CK denote sets containing the indices of the observations in each cluster.

These sets satisfy two properties:

1. C1 ∪ C2 ∪...∪ CK={1,...,n}. In other words, each observation belongs to at least one of the K clusters.

2. Ck ∩ Ck1 = ∅ for all k1! = k

Algorithm

- Step 1: Randomly assign a value to k
- Step 2: Compute the centroid of each cluster
- Step 3: Assign each observation to the cluster with closest centroid
- Step 4: Repeat steps 2 and 3 till the clusters stop changing

K = 10

```
: 3                                  Greening the ghetto
  7                Behind the design of Seattle's library
  10          My wish: A call for open-source architecture
  18                    Organic design, inspired by nature
  50                     The hidden world of shadow cities
  73                       How architecture can connect us
  89                               Cradle to cradle design
  102                   The ghastly tragedy of the suburbs
  103                            Human-centered design
  112                        A memorial at Ground Zero
  160                     The sticky wonder of gecko feet
  174                              Treat design as art
  183                               A song of the city
  189                              Building uniqueness
  196             On the verge of creating synthetic life
  199                          My days as a young rebel
  203                     My green agenda for architecture
  227                  Designing objects that tell stories
  258                          The wonder of Zulu wire art
  262                       A tour of modern architecture
  297                Things I've learned in my life so far
  299    The Blur Building and other tech-empowered arc...
  306                          Design and the Elastic Mind
  358                  Organic algorithms in architecture
  359                                        Ways of seeing
  363                  Great design is serious, not solemn
  387                      My underground art explorations
  409                          Can design save newspapers?
  411                              The Airstream, restyled
  446                     A supercharged motorcycle design
                                     ...
  1801   Social maps that reveal a city's intersections...
  1810                                           Happy maps
  1829   Got a wicked problem? First, tell me how you m...
  1851   Why the buildings of the future will be shaped...
  1863   How to revive a neighborhood: with imagination...
  1894                       Magical houses, made of bamboo
  1895   Why city flags may be the worst-designed thing...
  1979   Design at the intersection of technology and b...
  2037          Why great architecture should tell a story
  2079                        How Airbnb designs for trust
  2117           Pirates, nurses and other rebel designers
  2153                     What can we learn from shortcuts?
  2159   How Syria's architecture laid the foundation f...
  2162       When we design for disability, we all benefit
  2169     A project of peace, painted across 50 buildings
  2200                     Architecture that's built to heal
  2213             Immigrant voices make democracy stronger
  2237   4 ways to build a human company in the age of ...
```

K = 20

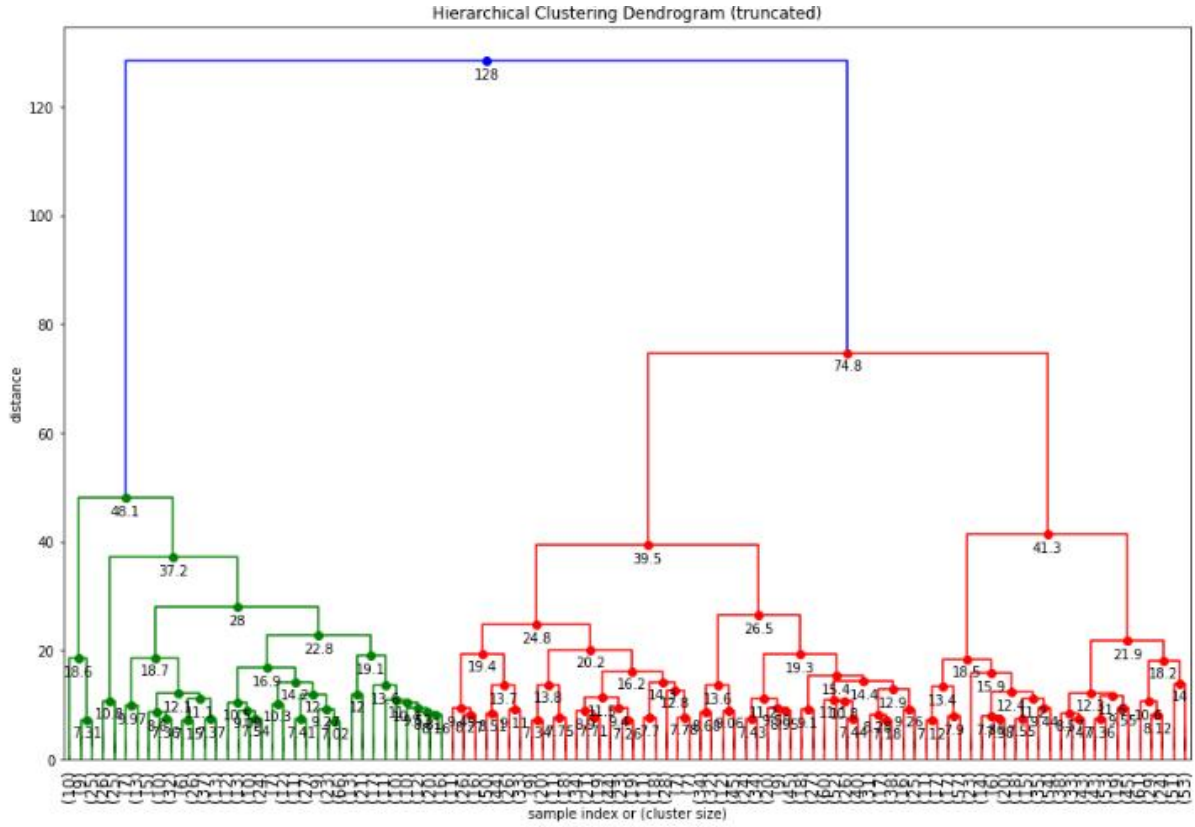
24	Happiness in body and soul
101	"Black Men Ski"
131	Patient capitalism
169	Tales of passion
289	What security means to me
449	A passionate, personal case for education
541	The surprising spread of Idol TV
562	Photographing the hidden story
584	Embrace your inner girl
643	Radical women, embracing tradition
782	Women, wartime and the dream of peace
790	A feminine response to Iceland's financial crash
791	A call to men
793	New data on the rise of women
798	Why we have too few women leaders
805	A test that finds 3x more breast tumors, and w...
819	Drawing on humor for change
824	Social media and the end of gender
829	Mother and daughter doctor-heroes
834	Inspiring a life of immersion
836	On being a woman and a diplomat
889	The mothers who found forgiveness, friendship
906	Art in exile
942	See Yemen through my eyes
967	Compassion and the true meaning of empathy
1028	What we learn before we're born
1057	"Women of Hope"
1068	Women entrepreneurs, example not exception
1111	Listening to shame
...	...
1956	An invitation to men who want a better world f...
1962	This tennis icon paved the way for women in sp...
1965	Why gender equality is good for everyone — men...
1968	How I stopped the Taliban from shutting down m...
1984	Why medicine often has dangerous side effects ...
2003	How I'm working for change inside my church
2014	The US needs paid family leave -- for the sake...
2017	The untapped genius that could change science ...
2024	A hilarious celebration of lifelong female fri...
2073	Teach girls bravery, not perfection
2163	3 lessons on success from an Arab businesswoman
2188	How women wage conflict without violence
2239	It's time for women to run for office
2244	The urgency of intersectionality
2249	A political party for women's equality
2268	What will you tell your daughters about 2016?
2275	The lies we tell pregnant women
2281	What happens when you have a disease doctors c...
2282	How online abuse of women has spiraled out of ...
2300	How racism harms pregnant women -- and what ca...

When k =10, mostly the clusters involving topics as Women but some talks related to science and art also included. After considering k=20, the clustering becomes meaningful, the talks about **Women** are clearly grouped.

HIERARCHICAL CLUSTERING

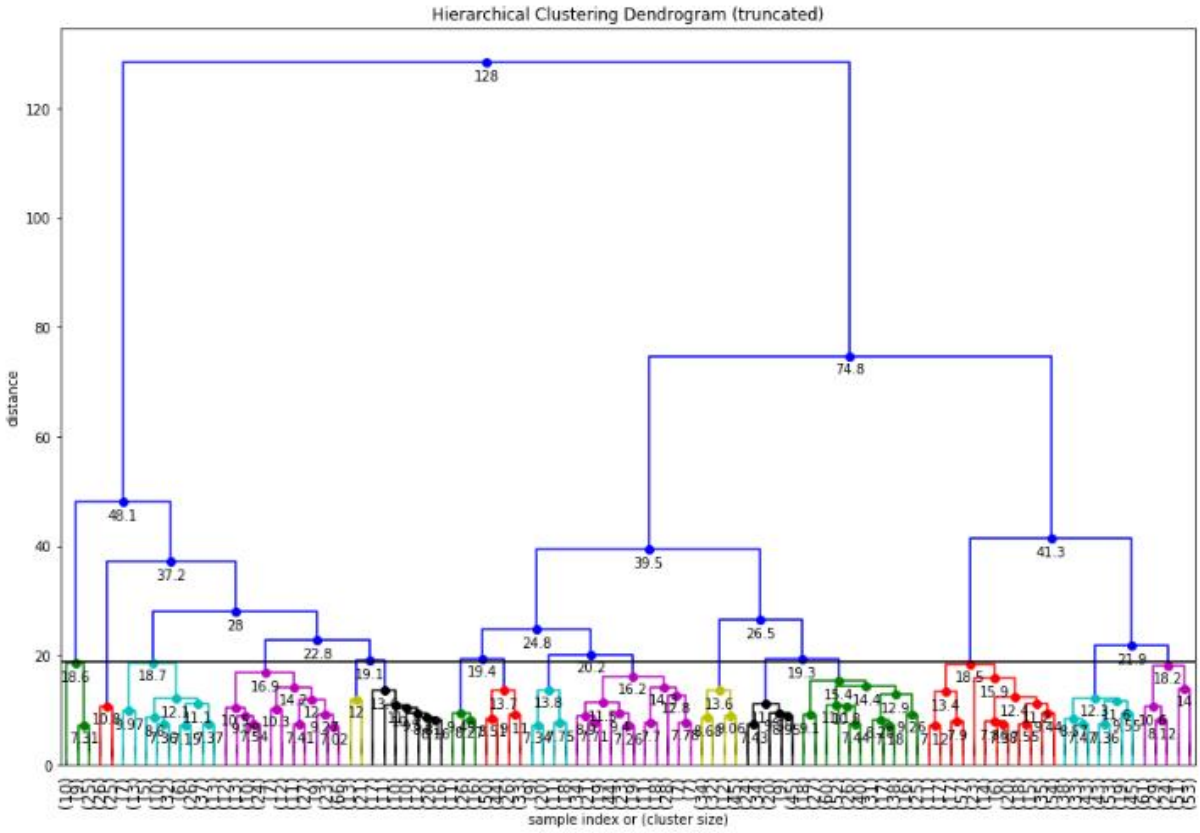
Rather than choosing clusters in k-means, hierarchical clustering considers each data point as one cluster. Next data point will be added to the previous cluster if it is close by. Process will be repeated till we get one giant cluster. The history tree thus formed is called a **Dendrogram.**

To understand the distance between clusters we can visualize only part of the tree by parameter **truncated mode**.

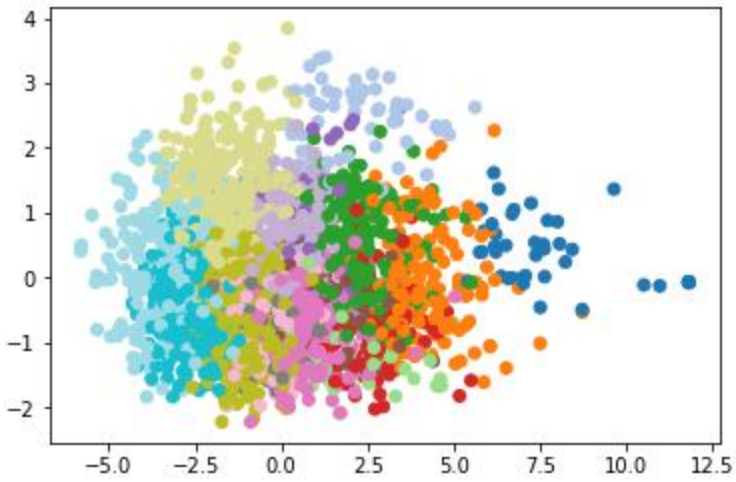


In the above dendrogram, when considering left part of it we can see **cluster at height of 10.9, merge to the next cluster at height of 37.2**, so distance between clusters is 26.3. It can be easily decided to cutoff the cluster at a height of 10.9, but if we analyze further **the left most cluster height of 18.6 merge to a cluster only at a height of 48.1, difference leads to 29.5.**

So, let us consider the **cutoff distance be 19 i.e. max_d = 19**



Using Scatter plot, we can visualize the clusters after applying PCA



Cluster depicting the issues and improvements in different parts of the world (Geography)

```
final['title'][final.cluster == 11]
```

```
4                          The best stats you've ever seen
33                    How mobile phones can fight poverty
36                          How to rebuild a broken state
51           Global priorities bigger than climate change
62                          My wish: Rebuilding Rwanda
108                   Salvation (and profit) in greentech
109                 Want to help Africa? Do business here
115                           New insights on poverty
125                      Africa's cheetahs versus hippos
128                                 Why invest in Africa
136                          Aid for Africa? No thanks.
152                   A commodities exchange for Ethiopia
230                                 The "bottom billion"
291                          Health and the human mind
298              Politics and religion are technologies
342                                   The future of cars
373             A solar energy system that tracks the sun
439             Insights on HIV, in stunning data visuals
450                   Why we're storing billions of seeds
482                          Wiring a web for global good
493                    Let my dataset change your mindset
511                          Photographs of secret sites
513                      Mapping the future of countries
547                          Asia's rise -- how and when
548                    Transition to a world without oil
552                   Global ethic vs. national interest
592                            How to expose the corrupt
648                   Social experiments to fight poverty
```

Cluster depicting the Topics related to Women

```
final['title'][final.cluster == 2]
```

```
449              A passionate, personal case for education
541                        The surprising spread of Idol TV
562                          Photographing the hidden story
643                     Radical women, embracing tradition
782                  Women, wartime and the dream of peace
791                                           A call to men
793                          New data on the rise of women
798                       Why we have too few women leaders
819                             Drawing on humor for change
824                     Social media and the end of gender
829                      Mother and daughter doctor-heroes
836                        On being a woman and a diplomat
889          The mothers who found forgiveness, friendship
906                                           Art in exile
967             Compassion and the true meaning of empathy
1068            Women entrepreneurs, example not exception
1111                                    Listening to shame
1150                   A teen just trying to figure it out
1188                  Women should represent women in media
```

Cluster depicting the Topics related to Technology and Data (Computers, Big Data, Robots, Smartphones etc.)¶

```
final['title'][final.cluster == 5]
```

```
397                                        The next web
433                                The mathematics of war
453              A university for the coming singularity
606         Is Pivot a turning point for web exploration?
610                       The year open data went worldwide
717                        The beauty of data visualization
742                                   The quantified self
815                Visualizing the medical data explosion
816                                  Silicon-based comedy
955                  Are we filtering the wrong microbes?
963                           Beware conflicts of interest
1001                                   Art made of storms
1142                              Texting that saves lives
1170               Revealing the lost codex of Archimedes
1244               The rise of human-computer cooperation
1406     If cars could talk, accidents might be avoidable
1478               Better baby care -- thanks to Formula 1
1610                      How data will transform business
1634    Your social media "likes" expose more than you...
1659                           Comics that ask "what if?"
1716                                 Own your body's data
1745                               Big data is better data
1843    How we found the worst place to park in New Yo...
```

Conclusion & Future work

NLP techniques like topic modelling, similarity findings help in building recommendation systems and customizing search tags. Clustering helps in placing the talk in right groups based on text analysis. As the number of samples increases the time consumption also increases, so reducing dimensionality should be taken to consideration. Text pre-processing also crucial like considering which stop words needs to be removed and performing stemming and lemmatization.

Future work can be extended in identifying the top rating talks based on text scripts. Applying word2vec for vectorization and other Deep Learning techniques.