TEXT ANALYTICS OF TED TALKS

CLIENT

- 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.

PROBLEM STATEMENT

- 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

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 feature from the YouTube.

Data Wrangling - Challenges

MAIN CHALLENGES IN MERGING YOUTUBE DATA AND TED DATA:

- No 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.

Data Wrangling - Strategy

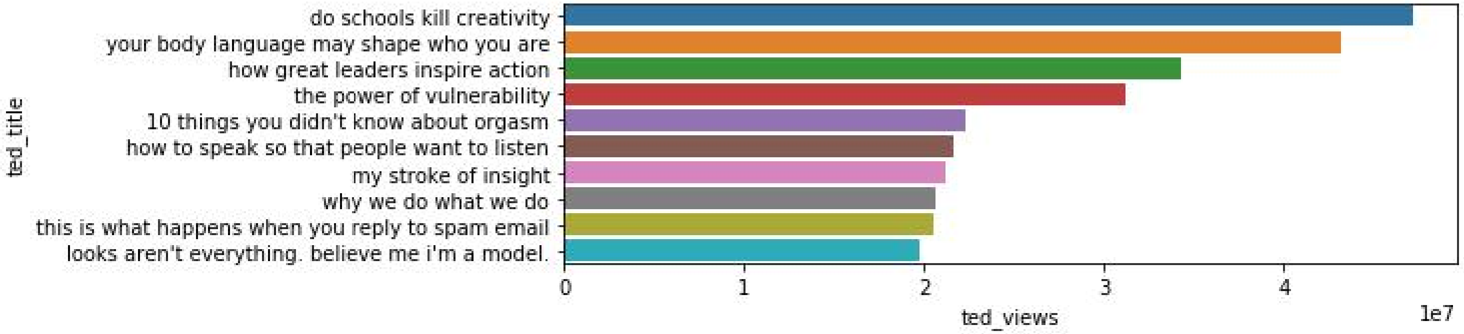
- Using **pandas. series. String** functions like **strip, replace and concatenate** the texts in the titles are cleaned.
- As YouTube have no separate columns for title, speaker and title are separated into new columns using **merge and split** functions.
- First titles with exact match of words are 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.

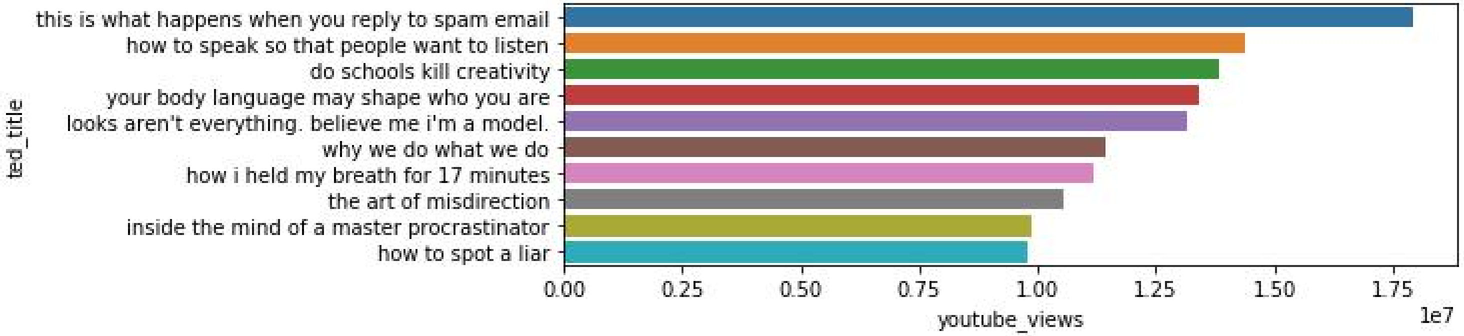
Exploratory Data Analysis

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

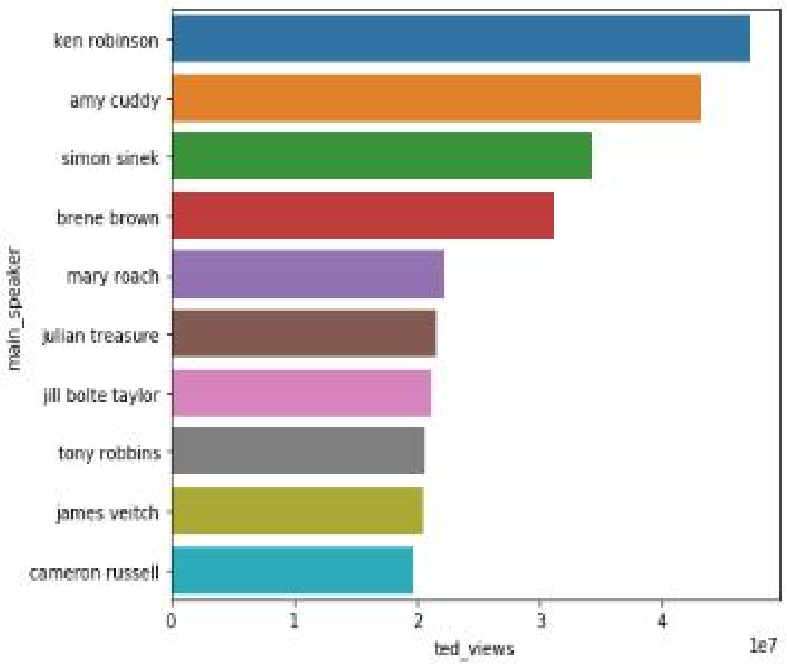
- EDA is performed as a comparison between YouTube and TED data
- Impact of Talk on viewers
- Categories with most views
- TED talk familiarity with viewers over years

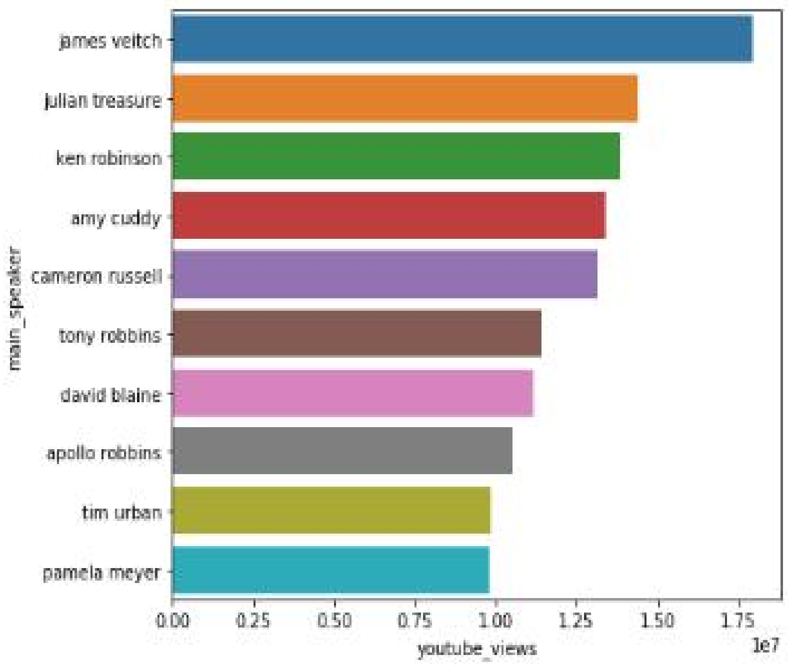
TITLE WINNERS



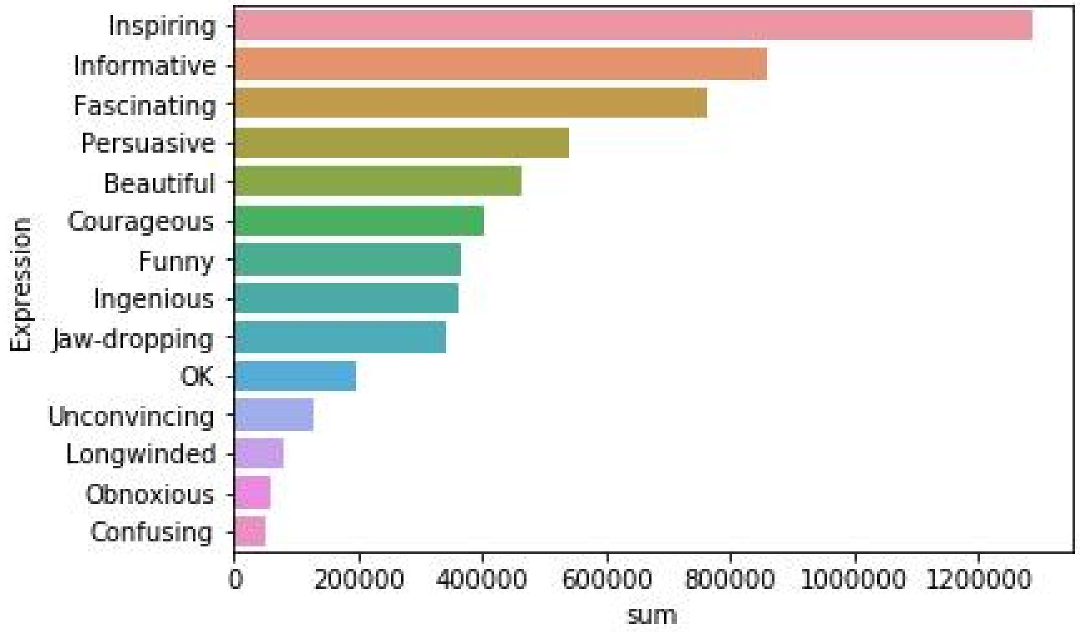


Best Speakers

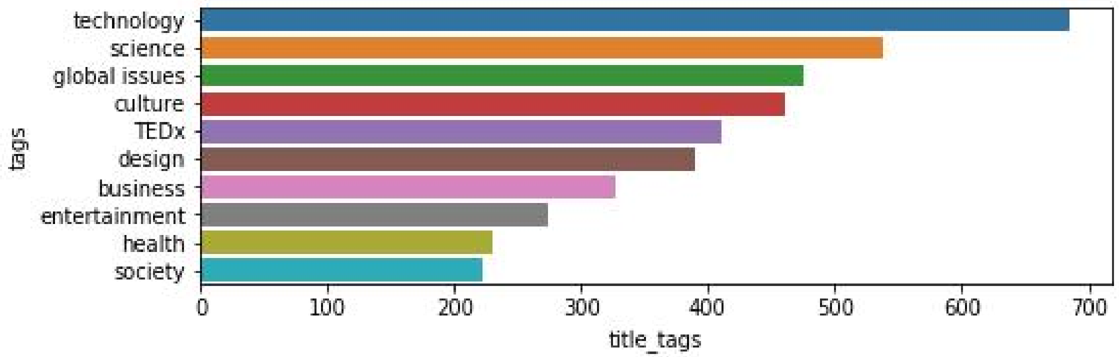


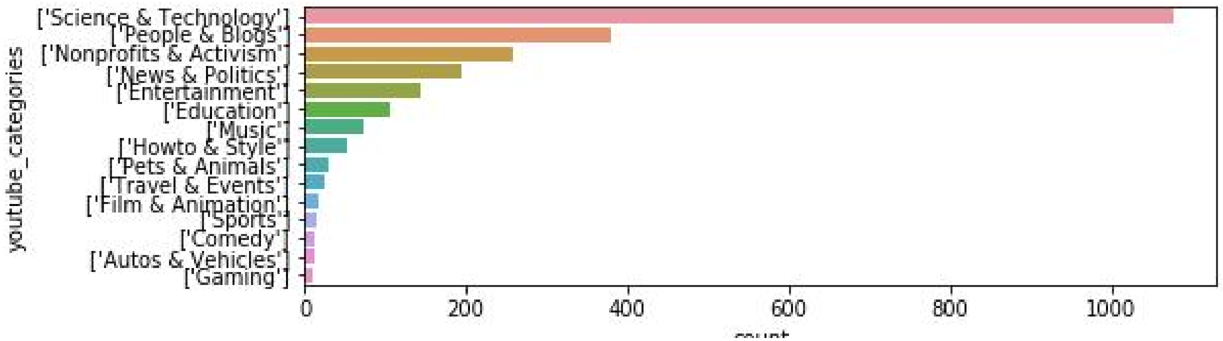


Impact of TED Talk

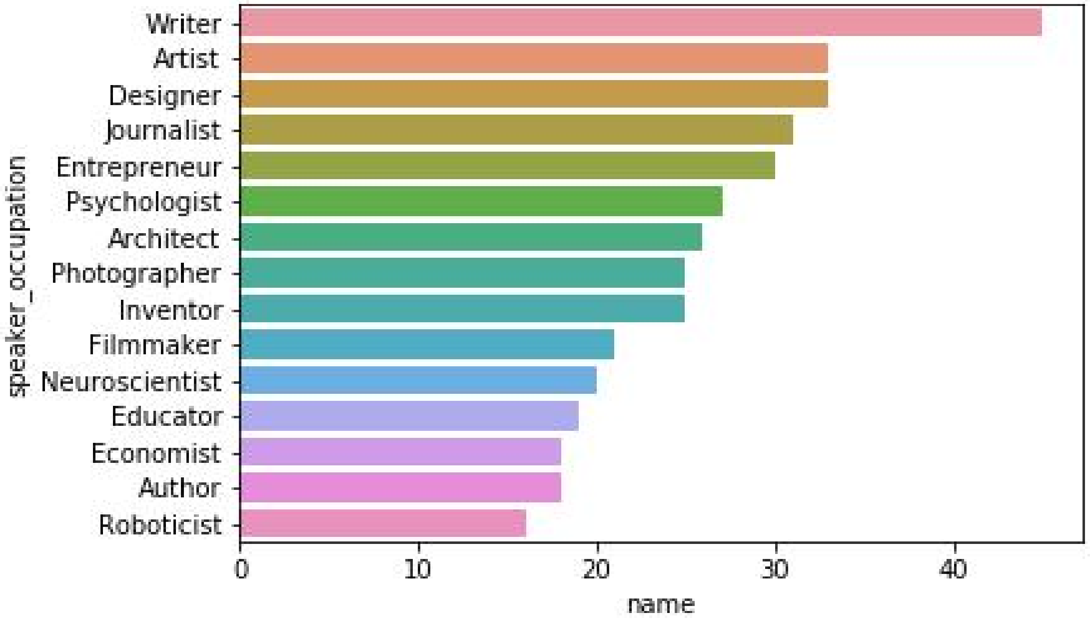


Exploring Categories

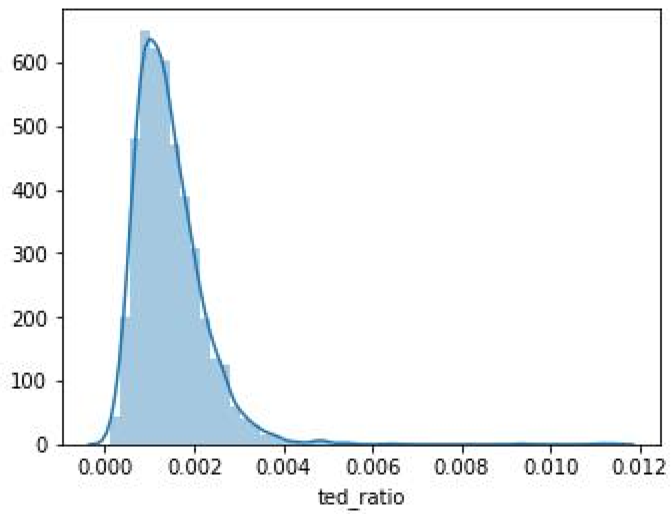


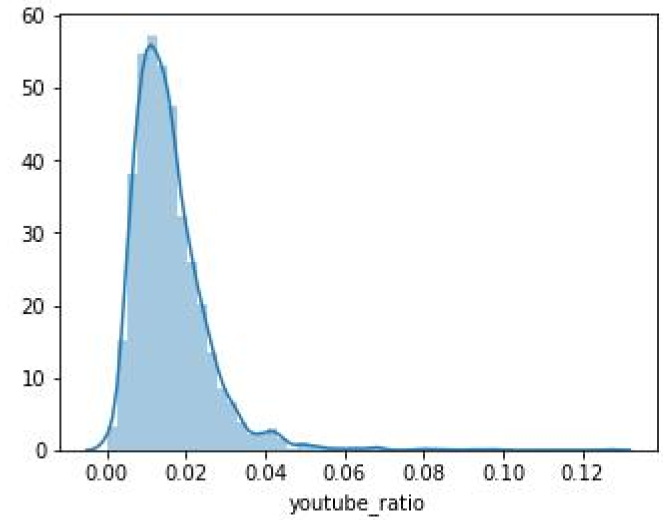


Diverse Occupations of speakers

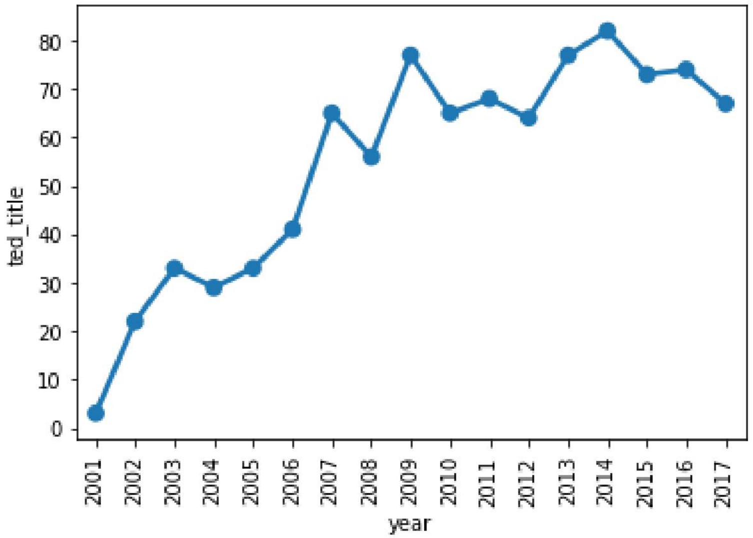


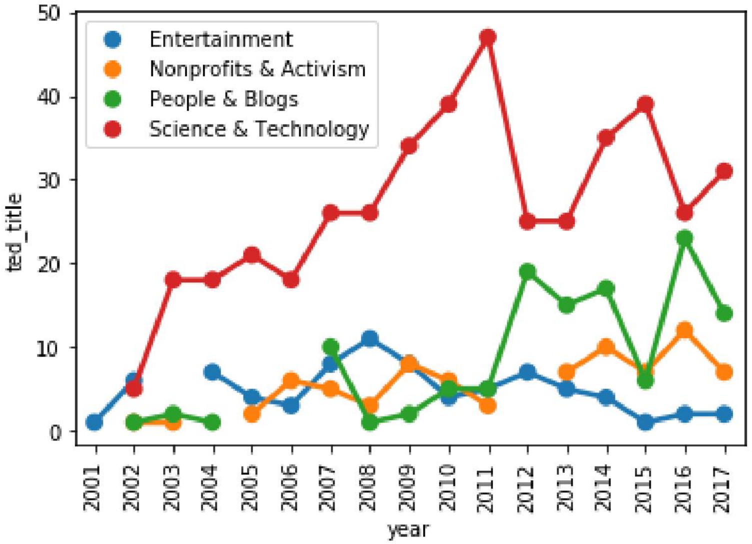
Viewers willingness to comment



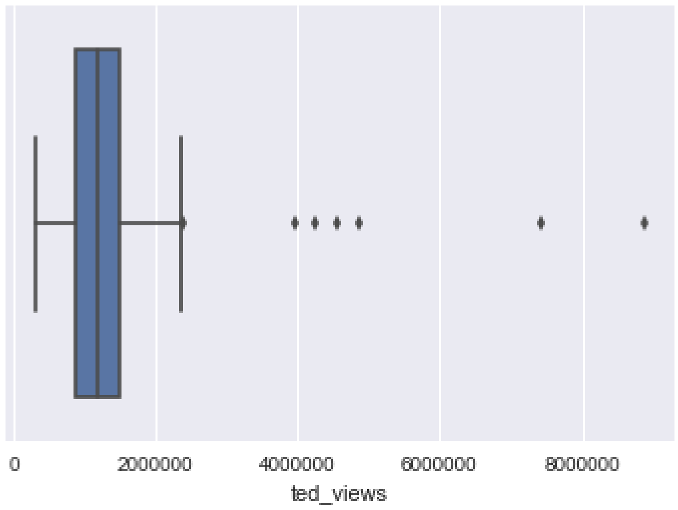


TED Talk growth over years



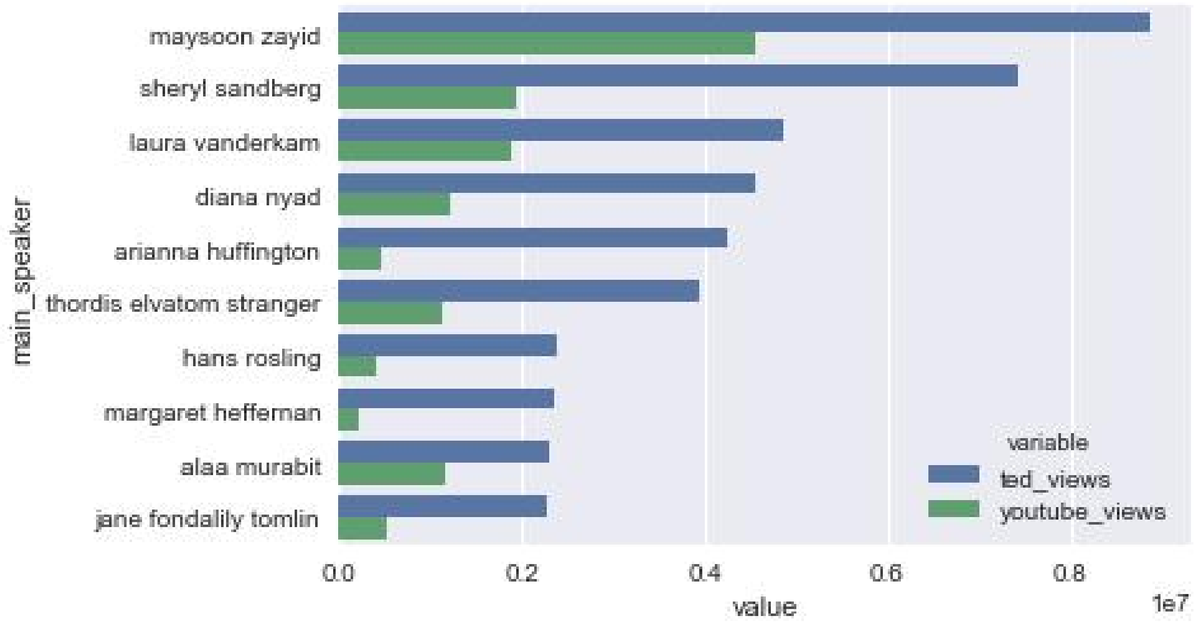


Great Response for TED Women

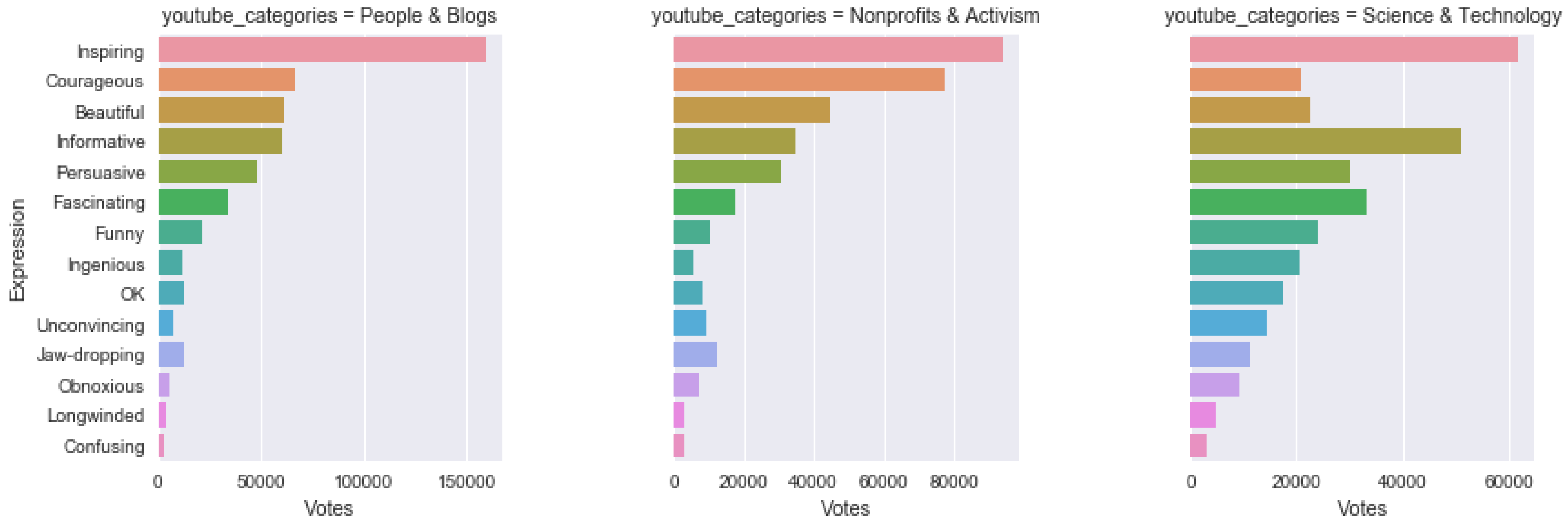


Most of the TED Women talks have 1 Million to 1.8 Million.
The range is pretty good as the count of views is consistent among all the talks

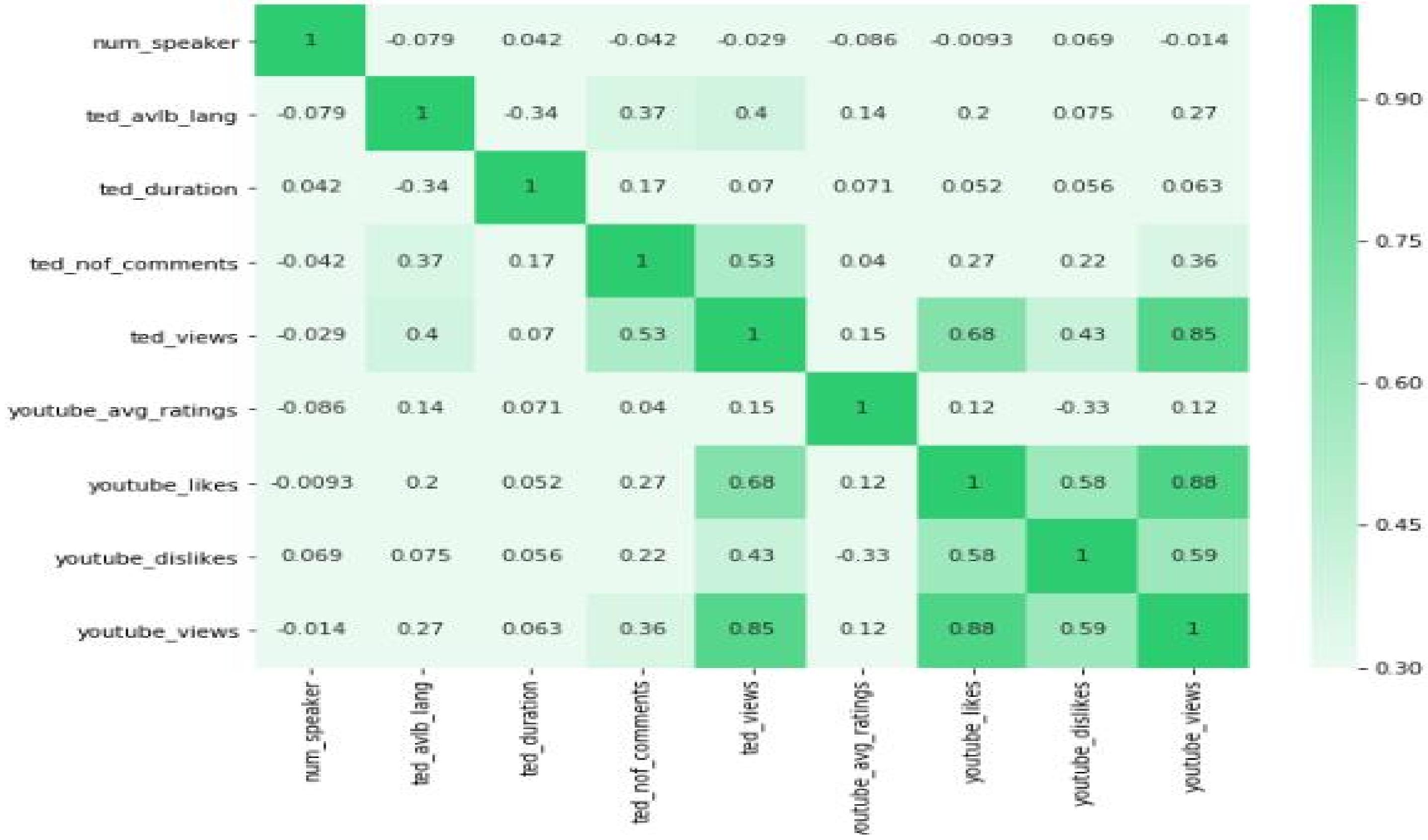
Best Women Speakers



Top 3 Categories - Impact



Numbers speaks correlation



TED Scripts – Text Analytics

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 Pre-Processing

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

Tokenization TF - IDF

TF-IDF stands for **Term Frequency - Inverse Document Frequency**

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Recommendation System

- Recommendation system will be built based on a similarity measure. 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.

Results

```
#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
```

Topic Modelling - LDA

- **Latent Dirichlet Allocation** (LDA) uses two probability values:

  P (word | topics) and P (topics | documents).

- 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”.

LDA - NMF

LDA

```
{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

```
{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']}
```

Comparing Results

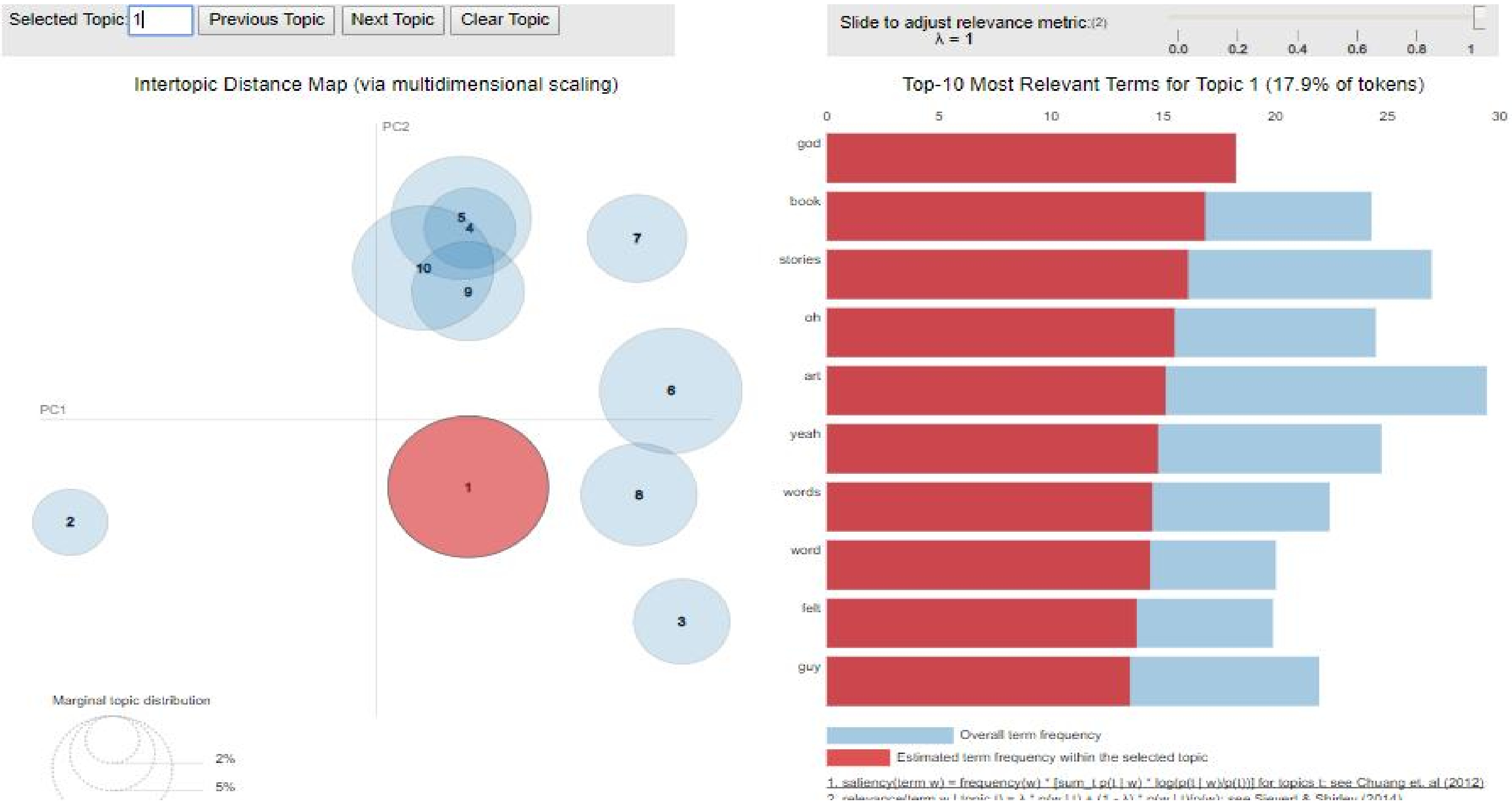
LDA

```
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 ...
```

NMF

```
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 ...
```

Visualization



Clustering

Best-known clustering approaches **K-Means and Hierarchical Clustering**

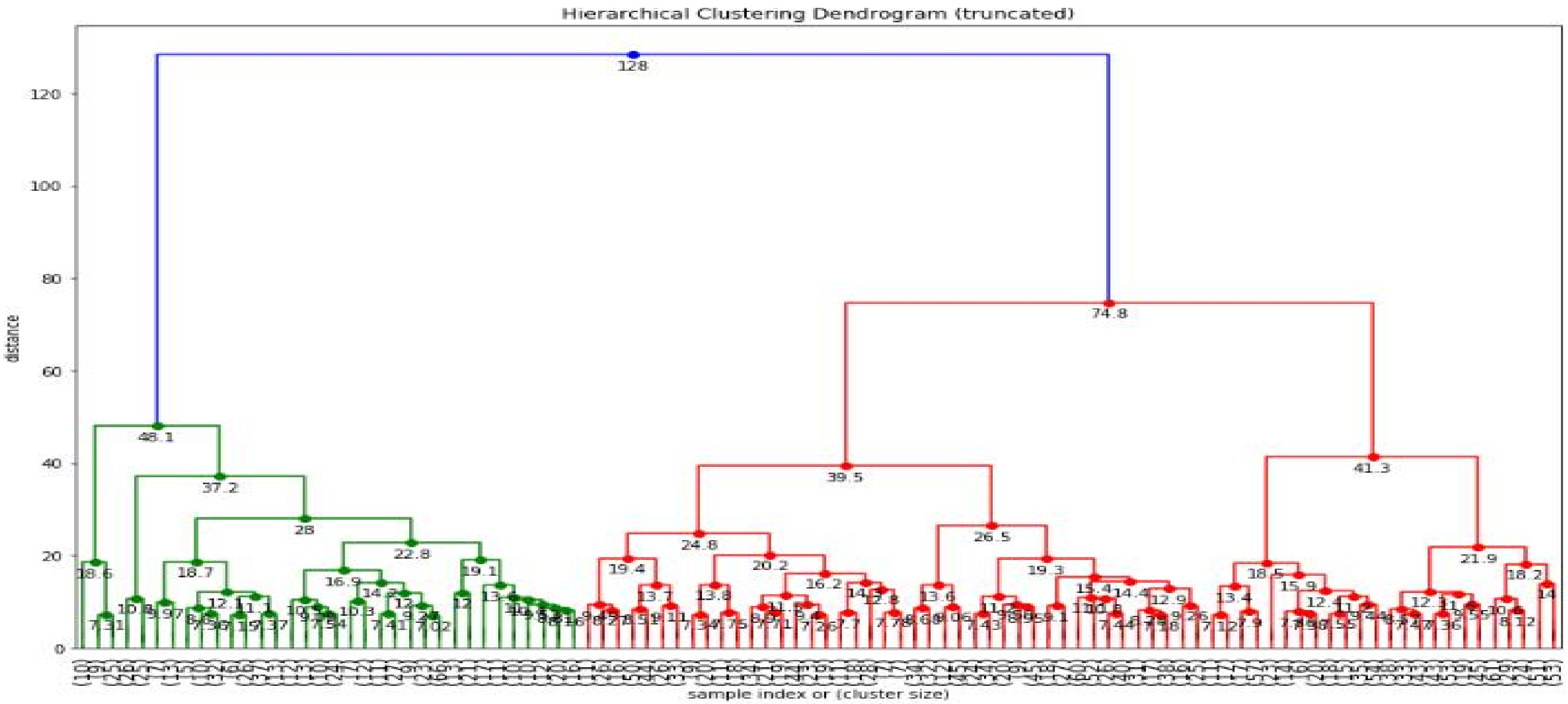
K-MEANS

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.

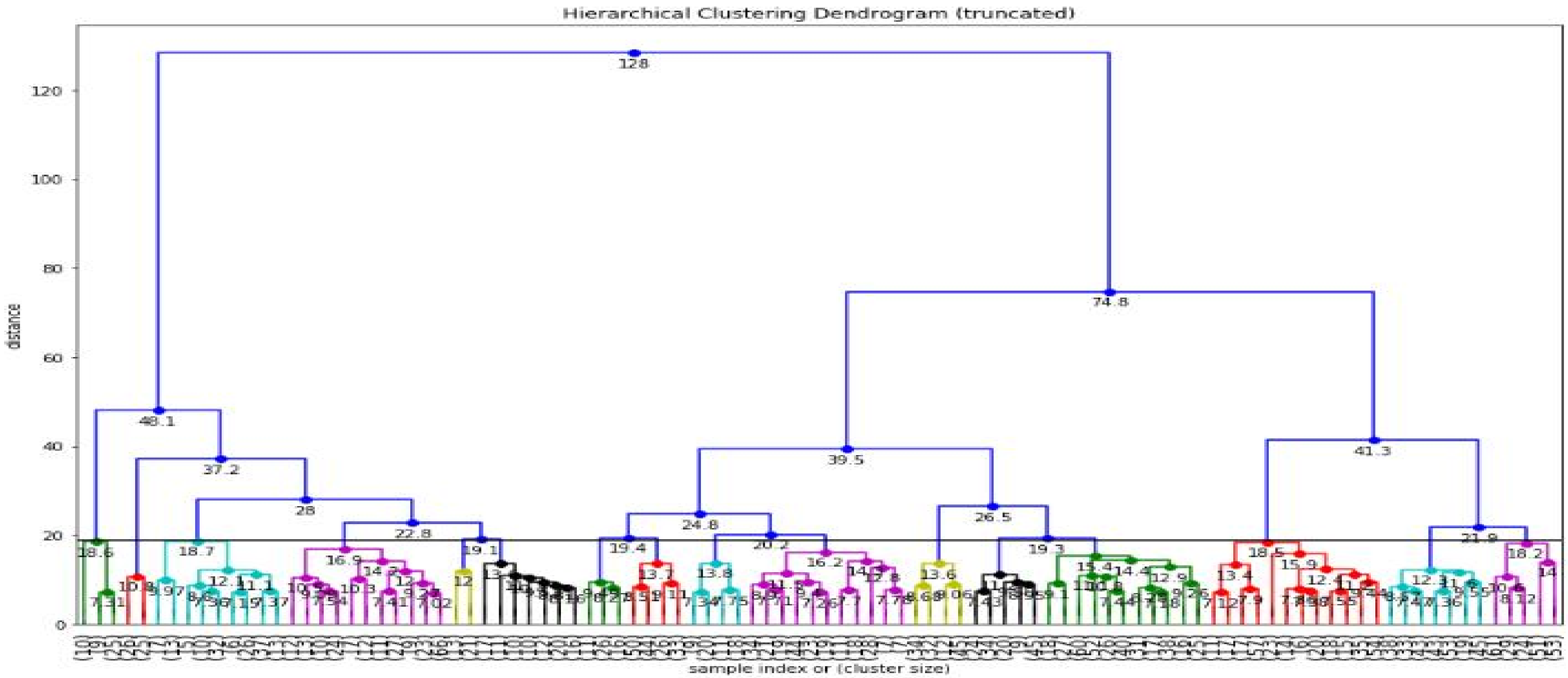
Hierarchical clustering

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.**

Truncated Mode



Cut-off Distance = 19



Cluster - 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

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