ORIGINAL RESEARCH PAPER
Volume - 9 | Issue - 9 | September - 2020 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

EXPLORATORY DATA ANALYSIS AND TOPIC MODELLING ON TED TALKS

Media

Parichay Pothepalli

Tower-15, 301, The Close North, Nirvana Country, Gurgaon, 122018

ABSTRACT

TED (Technology, Entertainment, Design) is a non-profit organization that influences the audience across the globe to deep dive into thinking .The short, powerful talks in more than 100 languages, from great inspired achievers engage the curious people and change their way of perception about issues on science, entertainment, business, technology, global concerns and various other topics. Why do some TED Talks get more views, go viral? What makes a TED talk the change maker in outlook, attitude and behaviour? What intrigues and influences people? This paper aims to analyze the various drivers behind maximum view count of certain TED talks from the start of 2006 till the end of June 2020. The analysis takes into consideration various parameters such as the speaker's profession, chosen topic, his/her transcript, number of views, comments, tags to name a few. This analysis will help an aspiring TED Talker identify the drivers and plan a video that will attract more view counts. This research paper uses Exploratory Data Analysis with a special emphasis on Natural Language Processing using the native Latent Dirichlet Allocation model from Gensim and the LDA Mallet. Analysis of the TED Talk data suggests that the content is a prime driving factor rather than the public speaking abilities thereby making the view count less predictable.

KEYWORDS

Exploratory Data Analysis (EDA), Latent Dirichlet Allocation (LDA), TED Talks, Topic Model

INTRODUCTION

There is a plethora of online platforms today that provide entertainment, information, knowledge and global affairs. One such platform is TED (Technology, Entertainment, Design) Talks, that operates under the slogan "ideas worth spreading" bringing renowned experts from various walks of life, who give their invaluable insights from their past experiences, irrespective of their profession.

These videos have been offered for free online viewing since 2006 and have been viewed over one billion times worldwide till date. Some of the TED Talks rack up millions of views, go viral, while others are viewed less. In the current world of Data Analytics, looking for patterns, analyzing the talks to gain meaningful insights into the success criteria of having maximum view counts of a TED Talk will be helpful for aspiring TED Talkers. This research paper has used the data derived from Kaggle. It assesses the factors behind a high view count of any TED Talk, if there is any correlation between variables such as number of comments, tags, type of TED platforms, duration of the video, month/day of publishing to name a few.

RESEARCH ELABORATION AND DATA ANALYSIS

The data (4609 TED Talks) from the year of 2006 till June 2020 has been extracted from the public dataset domain of Kaggle. For the purpose of analysis and correlation, specific parameters such as the number of comments, number of subtitled videos, duration of the talk, number of tags, day of publishing, name of the speaker, speaker's profession, the transcript, have been chosen to understand the drivers for increased view count of TED talks. Pre-built Latent Dirichlet Allocation (LDA) and LDA Mallet (Java-based console application) is used for Topic Modelling. LDA takes into consideration a range of different topics thereby generating words based on their probability distribution. Gensim is a Python library with good tools that work flexibly by tweaking more parameters and Mallet regardless of being a console application is much more user friendly. Exploratory Data Analysis and Topic Modelling using Python packages such as Numpy, Pandas, Seaborn, Gensim, NLTK, re and Spacy have been used for data analysis.

1. Exploratory Data Analysis

Exploratory Data Analysis is the one of the first key steps done on the data set of TED Talks in this research paper. The Seaborn heatmap correlation matrix in Figure 1 suggests that the comments have a correlation of 0.5 with view count indicating a relatively higher positive correlation. The correlation between view count and the number of subtitled videos is weak (0.28), while the view count with respect to duration of a TED Talk is the least.

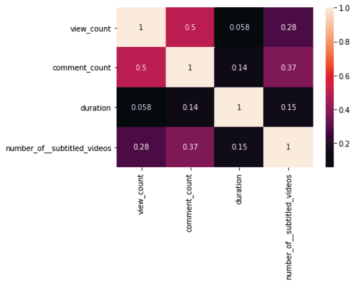

Figure 1 Heatmap Correlation Matrix

1.1 Comment count as a parameter

If the TED talks are sorted on the basis of highest comment count (Figure 2 given below), the TED talk name "Militant Atheism" has the highest number of comments but does not have maximum view count. This may be due to the type of controversial topic that openly seeks comments on atheism. The TED talk by Sir Ken Robinson garners the highest views yet it has the second highest comments.

	talk_name	speaker_name	view_count	comment_count	number_of_subtitled_videos
226	Militant atheism	Richard Dawkins	5837159	6456.0	42
0	Do schools kill creativity?	Sir Ken Robinson	65678748	4952.0	61
195	Science can answer moral questions	Sam Harris	6403424	3430.0	39
650	How do you explain consciousness?	David Chalmers	2937716	3006.0	34
16	My stroke of insight	Jill Bolte Taylor	26728715	2986.0	47

Figure 2 TED Talk Comments and views

To further understand this relationship, the outliers on the basis of Z-score have been removed in the scatterplot (Figure 3). A Linear regression line using the "Spearman coefficient" is placed in the graph. As per this graph, the correlation between both the parameters is 0.5. It is likely that one expects that with a greater number of comments there also will be a greater number of views. However, Figure 2 (above) proves that there is no predictive correlation between the parameters view count and comment count.

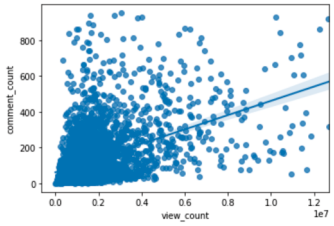

Figure 3 Scatterplot View Count vs Comment Count

1.2 Duration as a parameter

Data for duration of the TED Talk is extracted (in seconds) and almost all the talks range between 60- 1080 seconds or 1-18 minutes. The scatterplot in Figure 4 shows a normal distribution around 1000 seconds. The videos after 2000 seconds show a steep decline in the number of view counts indicating a trend that the audience is not attracted towards very long TED Talks. Short videos tend to keep the viewer gripped towards the topic. The longest TED Talk -"Parrots, the Universe, and Everything" lasted for 87 minutes, garnered only 0.5 million views, which is 0.007 % of the views garnered by the highest viewed TED Talk.

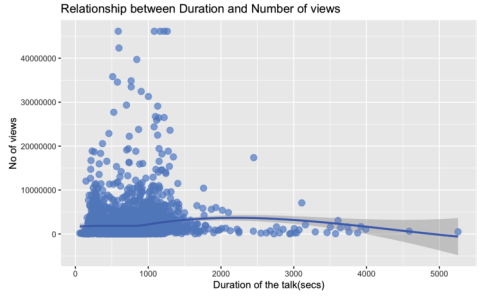

Figure 4 Scatterplot-Duration vs View Count

1.3 Publishing Month and Day as a parameter

The timestamp of TED Talks collected from 2006 till June 2020 on the basis of the month and day of publishing a video. From Figure 5A, it can be observed that February though having the least number of days has the highest number of TED talks being featured. This can be attributed to the reason that a large number of official TED conferences are held in February and thus are published on the website within the same month.

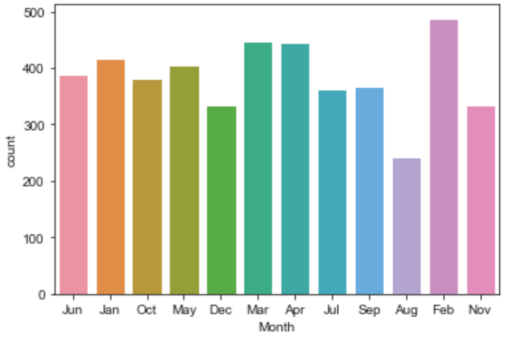

Figure 5A TED Talk View Count as per Month of Publishing (2006-June 2020)

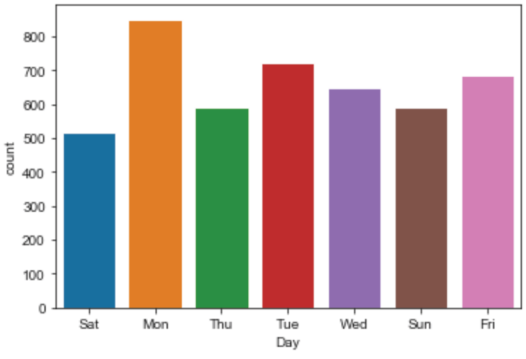

Figure 5B TED Talk View Count as per Day of Publishing (2006-June 2020)

From the above Figure 5B, TED Talks published on Saturdays seem to have much less views and Monday has the highest number of views followed by Tuesday and Friday. More than 800 talks were released on a Monday and an approximate of 700 TED talks on a Tuesday.

1.4 TED Talk Platform as a parameter

TED releases its talks on various platforms such as TED Stage Talks, TEDx Talk, TED-ED (Education platform) Originals to name a few.

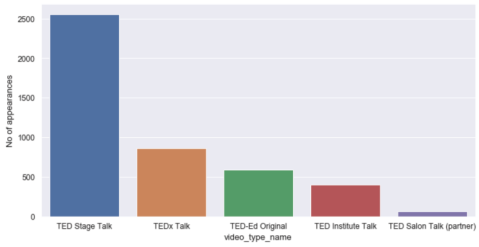

Figure 6A Number of appearances on various TED Platforms

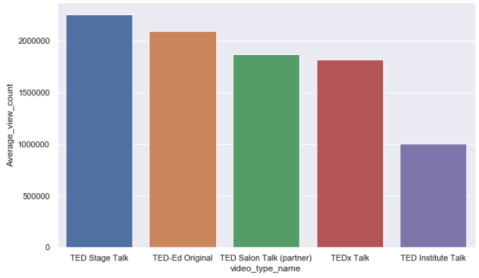

Figure 6B Average view count on various TED Platforms

Considering all the TED talks from 2006 to June 2020, it is observed (Figure 6A) that TED Stage Talks is the most popular platform followed by TEDx talks and TED-Ed originals. But does TED Stage Talk platform garner more views than any other platform? Figure 6B analyses that TED Stage Talk has the maximum average count of 2.25 Million views and the least approximate average count of 1 Million view count was noted on TED Institute Talks. An interesting observation is that though the number of appearances for TEDx Talks is high, its average view count is lower than TED-Ed Originals.

1.5 TED Event as a parameter

There are unique event names for each of the TED talks featured. Figure 7 shows that the TED talks that have the highest view counts have not always featured in the same event. Analysing the top 10 most viewed TED talks, it can be understood that the event in which it is featured has no impact on the number of view counts.

	talk_name	speaker_name	view_count	event	video_type_name
0	Do schools kill creativity?	Sir Ken Robinson	65678748	TED2006	TED Stage Talk
1	This is what happens when you reply to spam email	James Veitch	59725446	TEDGlobal>Geneva	TED Stage Talk
2	Your body language may shape who you are	Amy Cuddy	57734063	TEDGlobal 2012	TED Stage Talk
3	How great leaders inspire action	Simon Sinek	50494918	TEDxPuget Sound	TEDx Talk
4	The power of vulnerability	Brené Brown	48503432	TEDxHouston	TEDx Talk
5	How to speak so that people want to listen	Julian Treasure	42330489	TEDGlobal 2013	TED Stage Talk
6	Inside the mind of a master procrastinator	Tim Urban	39714672	TED2016	TED Stage Talk
7	The next outbreak? We're not ready	Bill Gates	35814459	TED2015	TED Stage Talk
8	My philosophy for a happy life	Sam Berns	34858496	TEDxMidAtlantic 2013	TEDx Talk
9	Looks aren't everything. Believe me, I'm a model.	Cameron Russell	34572281	TEDxMidAtlantic	TEDx Talk

Figure 7 TED Event w.r.t Highest View Counts

1.6 Profession of TED Speakers as a parameter

TED Talk speakers with a certain profession may choose to address concerns that may not be related to their primary professional expertise. Figure 8A shows the 10 most popular professions and writers are found to be the leading speakers, followed by entrepreneurs and journalists, artists, architects, designers and others. The Boxplot in Figure 8B also reveals that writers have the highest average view count followed by Educators and entrepreneurs.

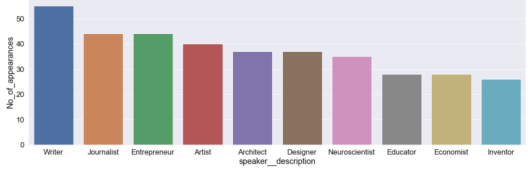

Figure 8A Speaker's Profession vs Number of appearances

The IQR (Inter-Quartile Region) of the view count is marked by the boundaries of the boxes for each of the professions. The baseline in the box plot measures 0 percentile and it is noticed that the talks delivered by Neuroscientists and Educators have a higher minimum view count than compared to topics delivered by either a writer, journalist or an

entrepreneur. Reading the 100th percentile gives us insights that writers, neuroscientists, educators have a far better maximum view counts than those compared to artists, journalists and architects. The dots above 100th percentile in the boxplot represent the outliers which explain that some entrepreneurs, journalists, educators and designers have outperformed their peers from the same industry and became popular with exceptional view counts.

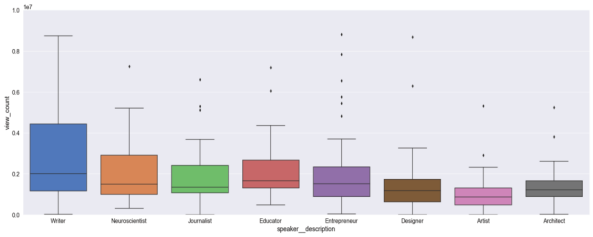

Figure 8B Boxplot of speaker's Profession vs Number of views

1.7 TED Talk Tags as a parameter

Multiple tags are associated with a specific TED talk. Some of the tags are science, culture, society, education, business and more. These tags (Figure 9A) can be diverse and are used to browse through related TED talks associated with the head and tail of talk names.

	talks__tags	talk__name
0	[children, creativity, culture, dance, educati...	Do schools kill creativity?
1	[comedy, curiosity, communication, humor, tech...	This is what happens when you reply to spam email
2	[body language, brain, business, psychology, s...	Your body language may shape who you are
3	[TEDx, business, entrepreneur, leadership, suc...	How great leaders inspire action
4	[TEDx, communication, culture, depression, fea...	The power of vulnerability
...	...	...
4604	[leadership, military, business]	Where are we trying to end up?
4605	[life, poetry, humanity, spoken word]	We are not mud
4606	[computers, science, technology]	Designing materials one atom at a time
4607	[journalism, education, social change, society...	Why student journalists should be protected fr...
4608	[music, performance, mining, singer]	A tribute to West Virginia coal miners

Figure 9A Talk tags w.r.t Talk name

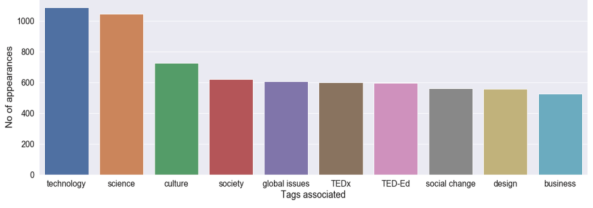

Figure 9B Tags associated vs Number of appearances

An analysis of Figure 9B shows the number of tags occurring in the TED talks dataset is maximum for "Technology" followed by "science" and "culture" during the period 2006- June 2020. Interestingly generic tags such as "TEDx" and "TED-Ed" are also used. The TED Talks which garnered an average view count of more than 2 Million views have tags in the order of Culture, Business followed by TED-Ed, TEDx and Science respectively. (Figure 9C)

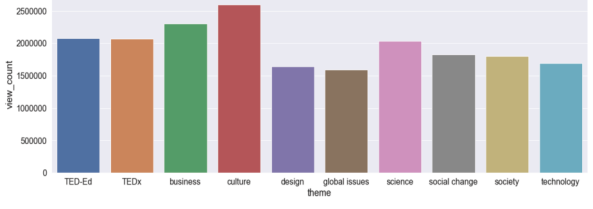

Figure 9C Tags w.r.t average view count

A trending change in the usage of tags over the years has been observed in figure 9D. This may be attributed to the various trending events or current affairs during the respective year of publishing of videos. Topics on "Society" have been of great interest to speakers during the years 2016-17 while talks related to "social change" have seen a rise since 2014. "Science" and "Technology" have been the most spoken and well received tags by the audience right from 2006 till date

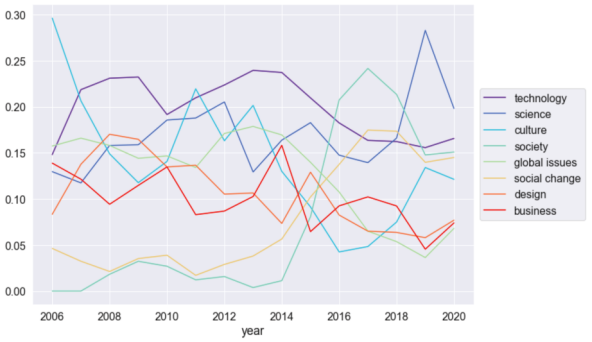

Figure 9D Change in trends of tags over the years

2. Transcript Analysis

Topic Modelling has been used to understand the broad topics under which the transcripts of all the TED talks can be classified. Analysis of TED transcripts will help in prediction of factors that influence view counts. The steps involved are Data Pre-Processing, Data Lemmatization, creating a Document Term Matrix and Corpus followed by Topic Modelling of the transcript using Gensim in-built LDA Model and the LDA Mallet.

2.1 Data Pre – Processing

After importing the data, Text pre-processing tools like the Gensim package- "Gensim.corpora", Gensim.utils" and "Gensim models" in Python have been used. After the text is obtained, text normalization was initiated i.e... converted all letters to lowercase, numbers into words, removed punctuations, white spaces, abbreviations, stop words and an additional set of insignificant stop words (Figure 10) from the TED talk data. Using the Natural Language Toolkit (NLTK) package the given text is further split into tokens.

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['go','let','think','people','come','look
                   "across", "after", "afterwards", "again"
                   "amongst", "amoungst", "amount",  "an",
                   "at", "back","be","became", "because","b
                   "besides", "between", "beyond", "bill",
                   "de", "describe", "detail", "do", "done"
                   "enough", "etc", "even", "ever", "every"
                   "first", "five", "for", "former", "forme
                   "hasnt", "have", "he", "hence", "her", "
                   "however", "hundred", "ie", "if", "in",
                   "least", "less", "ltd", "made", "many",
                   "must", "my", "myself", "name", "namely"
                   "nothing", "now", "nowhere", "of", "off"
```

Figure 10 Sample of stop words extended from the default

2.2 Data Lemmatization

After the data has been pre-processed into tokens, it is then converted into bigrams and trigrams using the Phrases () module in Gensim. Bigrams are then defined by the frequency by which two words occur together frequently, for modelling purposes. Lemmatization is a process where different forms of the word are converted into their root form, to make topic modelling easier and more coherent. In the definition of the function only the "Noun", "Adj", "Verb" and "Adverb", are passed, as other parts of the speech would not help in generating a good topic model and result in a low coherence score. (Figure 11)

```
def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):
    """https://spacy.io/api/annotation"""
    texts_out = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append([token.lemma_ for token in doc if token.pos_ in allowed_postags])
    return texts_out
```

Figure 11 Function for Data Lemmatization.

2.3 Document Term Matrix and Corpus

After Data Lemmatization, a dictionary of all the words used in the 4609 TED Talks was generated. For each of the above word used along with its frequency a document term matrix was created. This presented a corpus wherein Gensim was used in creating a unique id for all the words in the dictionary. Figure 12 shows the mapping of the word ID and the word frequency.

```
In [24]: [[(id2word[id], freq) for id, freq in cp] for cp in corpus[:1]]

Out[24]: [[('ability', 2),
           ('abstract', 1),
           ('academic', 4),
           ('accord', 1),
           ('achievement', 1),
           ('adult', 1),
           ('advice', 1),
           ('affection', 1),
           ('afford', 1),
           ('afterward', 1),
           ('agent', 1),
           ('agree', 1),
           ('alien', 1),
           ('allow', 1),
           ('anniversary', 1),
```

Figure 12: Corpus generated using TED Talks transcript

2.4 Topic Modelling

Topic Modeling-a process to extract the hidden topics from large TED Talk datasets. Corpus and the term document matrix are ready for Topic Modelling. This paper has used Gensim pre-built Latent Dirichlet Allocation (LDA) and LDA Mallet (Java-based console application) for topic modelling.

2.4.1 Gensim pre-built LDA for Topic Modelling

Upon numerous iterations the optimum number of topics for LDA in the Gensim package was 19 topics. The alpha value is set to "auto" and the random state="123"(Figure13)

```
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                            id2word=id2word,
                                            num_topics=19,
                                            random_state=123,
                                            update_every=1,
                                            chunksize=100,
                                            passes=10,
                                            alpha='auto',
                                            per_word_topics=True)
```

Figure 13 LDA model for Topic Modelling

Topics were generated by the combined contribution of keywords. Each key word contributes a certain weightage and can be grouped under a generic topic as mentioned in Figure 14. Drawing topics from Figure 14, the keywords can be grouped under a generic topic as mentioned in Table 1 below.

```
pprint(lda_model.print_topics())

[(0,
  '0.011*"number" + 0.010*"understand" + 0.010*"idea" + 0.009*"important" + '
  '0.009*"human" + 0.007*"course" + 0.007*"study" + 0.005*"reason" + '
  '0.005*"science" + 0.005*"group"'),
 (1,
  '0.018*"applaud" + 0.017*"trumpet" + 0.011*"achievable" + 0.008*"solo" + '
  '0.006*"groan" + 0.006*"native_american" + 0.005*"downright" + '
  '0.004*"psychotherapy" + 0.002*"mute" + 0.000*"rebel"'),
 (2,
  '0.314*"water" + 0.022*"drink" + 0.017*"tea" + 0.011*"pump" + 0.008*"pipe" + '
  '0.005*"stream" + 0.004*"impairment" + 0.004*"inconvenience" + '
  '0.004*"beverage" + 0.000*"ecosystem"'),
 (3,
  '0.044*"cell" + 0.037*"body" + 0.018*"animal" + 0.016*"human" + 0.012*"gene" '
  '+ 0.012*"dna" + 0.011*"protein" + 0.011*"smell" + 0.010*"molecule" + '
  '0.009*"tissue"'),
 (4,
  '0.125*"voice" + 0.065*"noise" + 0.040*"listen" + 0.033*"ear" + 0.022*"tone" '
  '+ 0.020*"silence" + 0.020*"downtown" + 0.017*"quiet" + 0.017*"loud" + '
  '0.015*"beatboxe"'),
 (5,
  '0.034*"patient" + 0.029*"health" + 0.023*"disease" + 0.022*"drug" + '
  '0.018*"cancer" + 0.018*"doctor" + 0.015*"treatment" + 0.015*"vaccine" + '
  '0.014*"medical" + 0.014*"virus"'),
```

Figure 14 Topics under the LDA Mode

Table 1 Topics obtained using LDA model

	TOPIC-3	TOPIC-7	TOPIC-8	TOPIC-10	TOPIC-11	TOPIC-13	TOPIC-14
Key words for Topics	Cell	Percent	Light	Water	Life	Country	School
	Body	Money	Planet	Fish	Love	Government	Child
	Animal	Company	Star	Planet	Live	War	Kid
	Human	Dollar	Universe	Ocean	Old	Political	Student
	DNA	Business	Earth	Animal	Friend	Power	Learn
	Gene	Market	Space	Plant	Home	Public	Education
	Protein	Economy	Wave	Tree	Realize	State	Community
Generic Topic Name	Biological Science	Business and Economics	Space and the Universe	Biosphere/ Earth	Quality Life and Compassion	Political Science	Education and Community

Upon topic modelling, the Perplexity of the LDA model was defined. Perplexity is a measurement of how well an LDA model predicts a sample. A low perplexity indicates that the probability distribution is good at predicting the sample. Using the log_perplexity function on the corpus of TED talk the following result was obtained. A low score such as -10.63 suggests a good LDA model (Figure 15). In order to further understand the topics generated using the keywords the coherence score is checked. The topic coherence is applied to the top N words from the topic. It is defined as the average / median of the pairwise word-similarity scores of the words in the topic.

```
print('\nPerplexity: ', lda_model.log_perplexity(corpus))

Perplexity:  -10.630326561309214
```

Figure 15 LDA Perplexity score of LDA model

The "c_v" coherence scores to check for the LDA model reflect a score of 0.5131 (Figure 16) indicating distinct topics and the top N words from the topic define the topic succinctly.

```
coherence_model_lda = CoherenceModel(model=lda_model, texts=data_lemmatized, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)

Coherence Score:  0.5131781419269746
```

Figure 16 Coherence score of LDA model

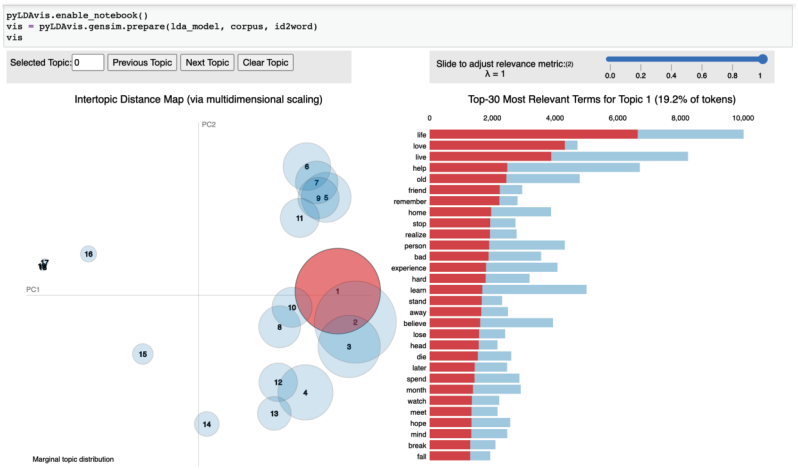

Figure 17 PyLDAvis visualization of LDA model

To graphically understand the topics modelled refer to PyLDAvis visualization (Figure 17). Each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent the topic. The big bubbles spread across all the quadrants with minimum overlaps suggest a very good model for TED talk transcripts. To compare and produce the best model, the same corpus is modelled using the LDA Mallet.

Table 2 Topics obtained using LDA Mallet model

	TOPIC-3	TOPIC-7	TOPIC-8	TOPIC-10	TOPIC-11	TOPIC-13	TOPIC-14
Key words for Topics	Live	Technology	Government	Music	Food	Woman	Energy
	Love	Computer	Power	Language	Water	Man	Power
	Life	Information	Political	Art	Animal	Sex	Air
	Friend	Machine	Country	Image	Eat	Black	Fuel
	Realize	Build	Public	Voice	Plant	Kill	Speed
	Die	Robot	State	Artist	Tree	Violence	Drive
	Fear	Device	War	Film	Fish	Community	Engine
Generic Topic Name	Human Life	Information Technology	Legislature	Artwork	Earth and Biosphere	Community	Energy and Fuel

Result-Topic Modeling

The topics generated using the LDA Mallet and Gensim pre-built LDA are almost similar. A coherence score "c_v" of 0.5025 (Figure 18) was generated using LDA

Mallet which is lower than the score generated for Gensim pre-built LDA model. Hence, Gensim pre-built LDA model is being considered an ideal model for the analysis of this research paper on TED Talk transcripts.

```
ldamallet = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus, num_topics=22, id2word=id2word)

coherence_model_ldamallet = CoherenceModel(model=ldamallet, texts=data_lemmatized, dictionary=id2word, coherence='c_v')
coherence_ldamallet = coherence_model_ldamallet.get_coherence()
print('\nCoherence Score: ', coherence_ldamallet)

Coherence Score:  0.502525874395684
```

Figure 18 Coherence score of the LDA Mallet package.

On the basis of the Gensim pre-built LDA model each TED Talk is assigned a dominant topic. (Figure 19) An analysis of the average view count for each dominant topic is presented in Figure 20. The key takeaway from the graph is, topic 14 (Education and Community) garners the highest average view count of 3 Million views followed by topic 11 (Quality Life and Compassion) at 2.7 Million views. The least average view count is observed for Topic 10 and 13 i.e. Biosphere/Earth and Political Science respectively. The analysis using Gensim pre-built LDA for Topic Modelling suggests that the highest average TED Talk view counts are garnered in the order of dominant topics related to Education and Community followed by Quality Life and Compassion, Biosphere/Earth and Political Science respectively. On observing the Figure 21 above, the analysis of this research paper is in total sync with the dominant topic of most viewed TED Talks.

TED_2

	Document_No	Dominant_Topic	Keywords	Text
0	0	14.0	school, child, kid, student, learn, education,...	Good morning. How are you?(Audience) Good.It's...
1	1	11.0	life, love, live, help, old, friend, remember,...	A few years ago, I got one of those spam email...
2	2	0.0	number, understand, idea, important, human, co...	So I want to start by offering you a free no-t...
3	3	0.0	number, understand, idea, important, human, co...	How do you explain when things don't go as we ...
4	4	11.0	life, love, live, help, old, friend, remember,...	So, I'll start with this: a couple years ago, ...
...	...	...	...	...
4086	4086	3.0	cell, body, animal, human, gene, dna, protein,...	Malaria is still one of the biggest killers on...
4087	4087	0.0	number, understand, idea, important, human, co...	After a harrowing chase, Ethic, Hedge, and th...
4088	4088	11.0	life, love, live, help, old, friend, remember,...	C'est la première sortie de mes années collège...
4089	4089	14.0	school, child, kid, student, learn, education,...	It's just after sunrise, and 16-year-old Mori...
4090	4090	6.0	music, idea, create, image, art, language, col...	The art of movies has existed for more than 10...

Figure 19 Dominant Topic for head and tail of TED Talk data

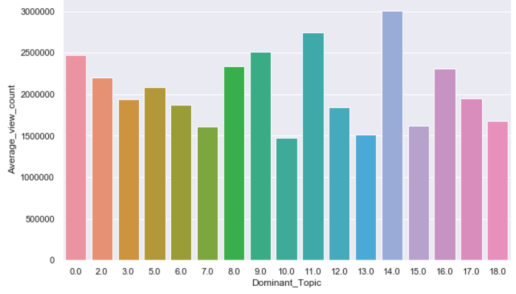

Figure 20 Dominant Topic w.r.t Average view count received.

2.4.2 LDA Mallet (Java-based console application) Topic Modeling

Mallet is an open source toolkit (a Java based package) used for NLP, document classification, clustering, topic modeling, and many other machine learning applications to text for an LDA model. Running LDA Mallet on the transcript of the TED talks, the below topics (Table 2) have been obtained.

	talk_name	speaker_name	view_count	published_timestamp	Dominant_Topic	Topic_Perc_Contrib
0	Do schools kill creativity?	Sir Ken Robinson	65678748.0	2006-06-27	14.0	0.2403
1	This is what happens when you reply to spam email	James Veitch	59725446.0	2016-01-08	11.0	0.4163
2	Your body language may shape who you are	Amy Cuddy	57734063.0	2012-10-01	0.0	0.3527
3	How great leaders inspire action	Simon Sinek	50494918.0	2010-05-04	0.0	0.3255
4	The power of vulnerability	Brené Brown	48503432.0	2010-12-23	11.0	0.5504
5	How to speak so that people want to listen	Julian Treasure	42330489.0	2014-06-27	0.0	0.3719
6	Inside the mind of a master procrastinator	Tim Urban	39714672.0	2016-03-15	11.0	0.4522
7	The next outbreak? We're not ready	Bill Gates	35814459.0	2015-04-03	5.0	0.3751
8	My philosophy for a happy life	Sam Berns	34858496.0	2018-03-28	11.0	0.4126
9	Looks aren't everything. Believe me, I'm a model.	Cameron Russell	34572281.0	2013-01-16	11.0	0.3421

Figure 21 Dominant Topic for most viewed TED Talks

CONCLUSION

TED Talks is a powerful platform that is instrumental in influencing the audience across the globe. Exploratory Data Analysis on various parameters along with a special emphasis on Natural Language Processing using the native Latent Dirichlet Allocation model from Gensim and the LDA Mallet provides the following insights which will help an aspiring TED Talker identify the drivers and plan a video that will attract more view counts. February is clearly the most popular month for TED Talks and Monday is the most popular day for publishing. Analysis shows that a short video duration keeps the viewer gripped and caters to the average person's span of attention, which is around 1000 seconds/ 17 minutes. TED Stage Talks can be used to present a talk as it is not only the most popular platform but also found to garner the highest number of appearances on the TED website.

Upon analysis of the highest average view counts -

- have appeared on TED Stage Talk (2.25 Million views) and the least approximate average count of 1 Million view count was noted on TED Institute Talks.
- are garnered by writers followed by educators and entrepreneurs. Some of the entrepreneurs, journalists, educators and designers have outperformed their peers and became popular with more view counts.
- may not generate a high number of comments as the comments have a correlation of 0.5 with view count.
- have tags in the order of Culture, Business followed by Science respectively.
- have no correlation with the event in which the Ted Talk is featured.
- approximates to around 3 Million views on the TED Talks related to Education and Community followed by topics on Quality Life and Compassion at 2.7 Million views.

Gensim pre-built LDA model is considered an ideal model for analyzing the TED Talk transcripts. Analysis of the TED Talk data suggests that the content is a prime driving factor rather than the public speaking abilities thereby making the view count less predictable.

Acknowledgement

The author would like to express his sincere thanks to Mr. Vijay Shankar, Ms. Manpreet Buddhiraja for their sincere efforts in helping and guiding him through the research paper. A special thanks to Dr Supraja P, Assistant Professor, SRMIST for her assistance with EDA and NLP in Python. Her comments have helped in greatly improving the manuscript.

REFERENCES

[1] David M. Blei, Andrew Y Ng, Michael I Jordan,"Latent Dirichlet Allocation" in Journal of Machine Learning Research 3 (2003) 993-1022
[2] Islam. Akef, Juan S. Munoz Arango in "Mallet vs Gensim: Topic modeling for 20 news groups report."
[3] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani in "Exploratory Data Analysis using Python"
[4] Diksha Khurana, Aditya Koli, Kiran Khatter and Sukhdev Singh in "Natural Language Processing: State of The Art, Current Trends and Challenges"
[5] Yoav Goldberg in "A Primer on Neural Network Models for Natural Language Processing".