

# **Analysing the effect of content of speech on viewership results**

Text Mining Course Project

**Miquel Torrens Dinarès**

*Barcelona Graduate School of Economics*

June 27, 2016

## **Abstract**

With the full expansion of the Internet worldwide and the appearance of massive social networks and social media, the amount of informational content quickly available to any human user has increased exponentially in the recent years. Because of that, in order to do a good selection of consumption of such content, users must dedicate their scarce time resources in a disciplined manner in order to maximise their utility given obvious constraints. In this project, we intend to analyse if the composition of the content is relevant for consumption, using content of the multimedia website TED Talks, by applying a few text mining techniques to extract information on what topics attract more or less viewership online. Results seem to point towards the relevance of the topic, emphasising this effect over the effect of less influential characteristics, such as duration or sentiment, among others.

*Keywords:* Speech popularity, Latent Dirichlet Allocation, TED Talks.

## **Framework**

During the 21st century, the amount of media content available to the average user has increased exponentially with the expansion of the Internet. Online media offer a varied, plural and sizeable range of content compared to traditional sources of general information. To that, we should pair the fact that the current consumer of such content has increased their education and expanded the limits of their interests. Nowadays, it is not strange to see considerable interest in content related to subjects on which the generic user may have superficial knowledge, such as science or technology. At the same time, the communication tools and skills of experts on these fields have evolved enough to make them attractive to the general public. It is not rare to see semi-technical content in

magazines, newspapers and online media of general interest, which may have not been equally common some years ago.

Simultaneously, the bloom of content has forced consumers to be more selective, since the time available for consumption has not increased at the path of the available content. With that, the importance of quality over quantity has become more relevant, and thus with a wide range of choice, the composition of the content stands out as an important factor of decision. More options imply being more restrictive in the consumption choice, and we will try to analyse how the topic of this content will affect the selection.

In this project, we intend to investigate towards this direction, by examining what features of the content of a speech are boosting the amount of attention that consumers pay to it. To do so, we analyse the content of the speeches published online at the popular multimedia website TED Talks, and we will measure how the composition of the talks is explaining the number of visits that they are receiving online. Since the website is free and open to anyone, and it contains content related to any sort of topic, we are especially interested in understand whether the topic of the content can have an effect on the final consumption aggregates.

This is a fundamental question in the media industry, since the atomization of content among all suppliers is forcing producers to put a strong emphasis on what content is actually published. Understanding which topics and what characteristics of it will partially determine the attractiveness of publications is an issue quite relevant from a sociologic, economic and entrepreneurial point of view.

## Data

The data used for this project consist of the transcripts of all talks posted on the popular website of TED Talks, at <http://www.ted.com>. The information retrieved includes the transcripts themselves, as well as some additional metadata on each talk. We summarise the main features extracted in the following list:

- Transcript of the talk, i.e. raw text of all statements said during the speech.
- Title of the talk
- Full name of the speaker performing
- Date in which the talk was posted online, in month format
- Date in which the talk was filmed, in month format
- Number of visualisations of the talk online at the moment of retrieval
- Duration of the talk, in seconds
- Set of qualitative tags assigned to the talk, which are adjectives that are suitable in each case, e.g. "inspiring", "ingenious", "beautiful"... These are voted by users from a set of adjectives made available by the website. Then, the website tags the speeches with the top voted options.
- Set of topics, e.g. "communications", "future", "parenting"... This set of topics is more diverse and is assigned by the website, without user interaction.

All data collected date as of June 1, 2016. The scrapper was launched at 9.00 AM GMT and the information retrieval lasted for about two hours. It was performed backward in time, which means that older talks were retrieved later, thus there might be a small difference to what would be considered a snapshot of the website. No other source was used as input for the project.

The data retrieved comprise a period of ten years of posted material, between June 2006 and May 2016. Nevertheless, the content posted may have been filmed prior to the uploading date, which means that a few talks were actually filmed in the early 2000s, or even before.

In the end, a total of 2,130 talks were successfully collected from the website. A small additional percentage of talks —between 1% and 2%— was not available in transcript mode and thus were not regarded. These are mainly concentrated within older talks. A handful of them were also disregarded because they were purely performance shows, such as music or dancing, which had no textual content to be analysed.

## Methodology

We will try to build a model that explains how the number of visualisations online of the speech is affected by the textual content of the talk. For that, we will exploit dictionary methods on the speeches, as well as a Latent Dirichlet Allocation (LDA) model that will provide probabilities of the talks belonging to a number of topics.

The entire material and the code for this project can be found in the following Github repository:

- <https://github.com/mtorrens/tm/tree/master/project>

We divide this section in the text mining and analysis part, and the viewership econometric model subsection.

### Text analysis

As a first step, cleaning the raw text requires some effort and discussion. The steps taken can be summarised as follows:

1. Tokenisation: this implies eliminating all punctuation, rebuilding the set of contractions and stripping the text down to the set of words involved. This way, the entire text of the speech is reduced to a bag of lower-cased words without diacritics or unrecognisable characters.
2. Cleaning the set of tokens: only those tokens that are not numbers or years are kept, i.e. a string such as "1950" is suppressed. In this context, it is helpful as this sort of tokens are not useful in topic detection, given that the plurality of topics is wide enough.
3. Removing stopwords: all tokens that include little meaning or that are too abundant are kept out of the model. These include pronouns, frequent connectors, modal

verbs, articles and highly-frequent verbs, e.g. "this", "which", "me", "not", "thing", "go", among many others. The full list of stopwords used for this corpus can be found in the aforementioned Github repository.

4. Removing rare words: words that once stemmed appear only once in the whole corpus are considered rare, and they are removed from the documents. In our varied set of documents, this is a relevant fraction of the unique words observed, although they are by definition a small fraction of the actual words used. This step has been quite helpful in the dimensionality reduction of the document term matrix, with its subsequent computations.
5. Stemming: to stem the words, the Porter stemmer has been employed.

After all these operations, the final set of tokens is reduced to 21,944 unique terms.

Once the set of tokens is established, the document term matrix (DTM) is built and we make use of dictionary methods to score all documents. For completeness, both the DTM and the TF-IDF scores were computed. The dictionary used for scoring the documents is the Harvard IV set. These scores will be used afterwards in the viewership model.

For the sentiment score, the AFINN-111 set of rated words has been employed. Both the cumulative and the relative sentiment scores have been computed for all documents, which will be another ingredient for the viewership model.

The main text mining technique used for the model, however, will be the exploration of topics of the talks, through an LDA on the documents. The Gibbs sampler is run for 3,000 iterations over the 2,130 documents and a size of the vocabulary of 21,944 unique terms. The total number of words is of 1,729,588 words. After trying a set of different number of topics, ranging from 5 to 20, the final choice has been  $K = 12$ . The reason for this value is the clarity of the resulting top words used for each topic, which appear to naturally cluster in a clear set of topics. Also, the visualisation of the most-likely topic allocation on the titles of the talks seem appropriate in the vast majority of manually checked cases. The choice of number of iterations run responds to the long stabilisation of the log-likelihood after each iteration of the Gibbs sampler is run.

The twelve resulting topics have as most used words the following list:

Topic 0: life, year, women, live, stori, day, time, peopl, school, love  
Topic 1: design, citi, build, work, creat, space, project, art, place, kind  
Topic 2: cell, diseas, patient, cancer, drug, health, year, actual, gene, doctor  
Topic 3: energi, year, earth, univers, planet, light, space, time, life, actual  
Topic 4: water, year, anim, food, ocean, speci, fish, plant, world, live  
Topic 5: peopl, war, countri, state, govern, polit, world, power, nation, american  
Topic 6: peopl, world, year, countri, percent, need, dollar, problem, work, money  
Topic 7: know, laughter, think, peopl, littl, time, work, come, actual, start  
Topic 8: brain, robot, bodi, move, human, control, time, show, video, differ  
Topic 9: think, peopl, differ, know, human, mean, question, time, kind, much  
Topic 10: music, play, laughter, applaus, sound, game, word, languag, hear, thank  
Topic 11: comput, data, technolog, actual, inform, peopl, internet, world, start, year

From now on, we will name the topics as follows:

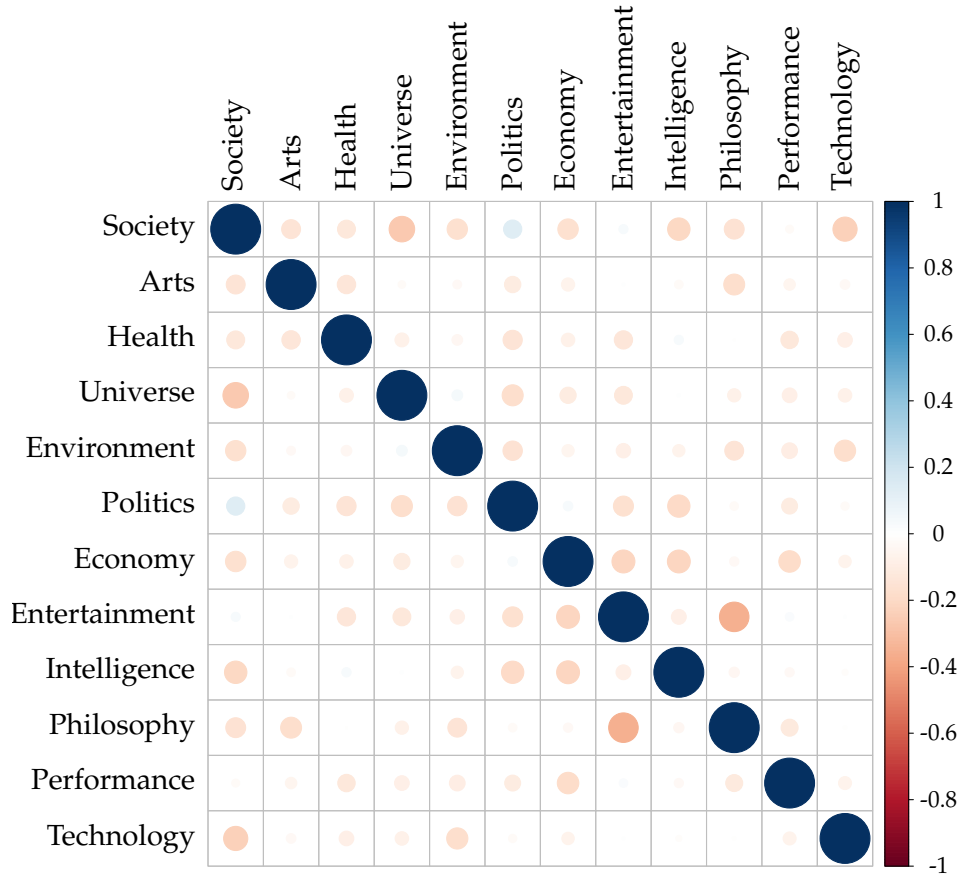
- Topic 0 (15.3%): Society
- Topic 1 (5.7%): Arts
- Topic 2 (5.8%): Health
- Topic 3 (5.4%): Universe
- Topic 4 (7.3%): Environment
- Topic 5 (5.3%): Politics
- Topic 6 (9.6%): Economy
- Topic 7 (15.9%): Entertainment
- Topic 8 (5.1%): Intelligence
- Topic 9 (16.7%): Philosophy
- Topic 10 (3.4%): Performance
- Topic 11 (4.7%): Technology

In parenthesis we have shown in what percentage of the documents, the topic is assigned with highest probability. We report what is the topic with maximum likelihood for the first ten documents, together with the title of the talk:

A smarter, more precise way to think about public health (top topic: 2)  
Why I bring theater to the military (top topic: 7)  
How barbershops can keep men healthy (top topic: 2)  
Drawings that show the beauty and fragility of Earth (top topic: 4)  
Your words may predict your future mental health (top topic: 9)  
The beauty of being a misfit (top topic: 0)  
How free is our freedom of the press? (top topic: 5)  
The laws that sex workers really want (top topic: 5)  
Our lonely society makes it hard to come home from war (top topic: 5)  
Good news in the fight against pancreatic cancer (top topic: 2)

We also report what is the correlation matrix between the resulting probabilities:

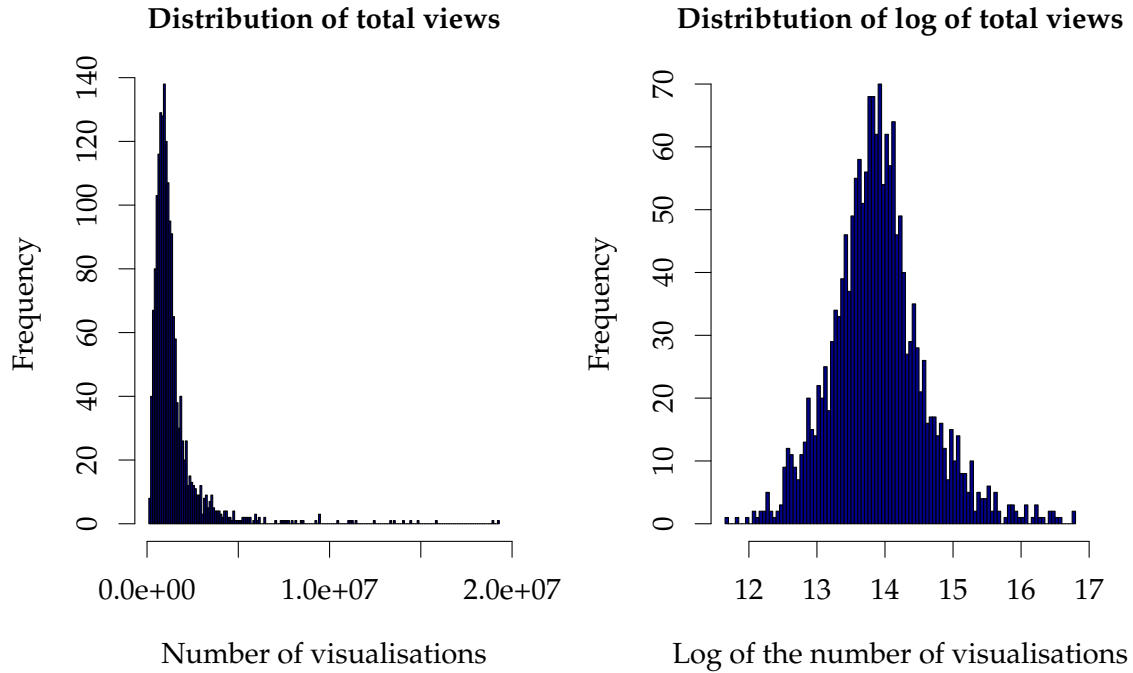
**Correlation plot of the probabilities of LDA Topics (document level)**



We can observe that most topic probabilities are moderately negatively correlated with the probabilities of being other topics, which is logical given that the probability space is represented by a simplex of degree  $K - 1$ . Therefore, two topics with stronger negative correlation suggest that their dissimilarity is also strong.

## Viewership model

To build the model to explain the number of visits a certain talk has received on the website, we will use several features of the text. The explained variable will be the natural logarithm of the number of visualisations. Given that the original variable is a count with high skewness, it seems appropriate to apply a logarithm to recover a more symmetric well-behaved empirical distribution. We can observe the shape of the observations prior and after the transformation:



After this, we will choose the dimensionality of the design matrix as follows. First of all, we make a selection of the observations that will be part of the model. We discard a set of them for different reasons:

- Documents that are too recent or too old. Those talks prior to January 2008 are discarded because we want to restrict the model to the period where the website had already sizeable market penetration. We also discard those documents from the month of May 2015, due to the fact that their visualisation counts were still immature at the date of retrieval. These filters combined suppress a total of 204 talks.
- Documents with extreme values on the number of visualisations. These are three outliers with an exceptionally high amount of views.
- Talks that last for less than three minutes, or over 20 minutes. The content of talks that are too short may have small significance on the number of views, as well as talks that are too long, which can affect the number of views negatively. Together, we disregard 167 talks with this filter.
- Talks that contain less than 100 words, as their textual content may have no effect on the number of visualisations, and which probably contain other determinant activities besides the speech. We eliminate 7 observations with this filter.

Once we have applied this set of filters, the total number of documents that remain to be part of the model is 1,749.

With respect to the predictors, the model works with features exclusively extracted from the text analysis, with three exceptions that act as control variables. These are, first, the natural logarithm of the amount of days during which the talk has been posted online, second, the natural logarithm of the duration of the the video in seconds and, third, a dummy variable indicating if the speaker is a recurrent speaker (1) or a new one (0). In addition to these three, the rest of variables included in the model are:

- The natural logarithm of the DTM score.
- The natural logarithm of the TF-IDF score.
- The natural logarithm of the number of words that the speech consists of. A variation was introduced using the number of unique words, which yielded identical results, as the correlation between the two is virtually 1. The case is the same with the number of characters.
- The cumulative sentiment score computed with the AFINN-111 sentiment dictionary.
- The topic probabilities obtained through the LDA analysis. Given that row-wise they sum to one, it is necessary to hold one of the features out. The reference topic of choice is "universe".
- Dummy variables that express if the document has been tagged as each of the qualitative tags explained in the data section. The set of tags is of length nine, which implies eight dummy variables, as row-wise these variables sum to two —each talk is tagged with the top-two tags—. Thus, the reference category will be the tag "informative".
- Dummy variables that express if the document has been tagged as each of the topics explained in the data section. Given that this set is large, we consider only those topics assigned to at least 5% of the talks, also to make the estimations more consistent. The table of results includes the model with significant topics, since many of the topics are far from significance even if present in a non-negligible fraction of talks.

With these, the full design matrix for the model consists of 1,749 speeches with 37 explanatory features each.

## Results

We build two OLS models. The first one is a preliminary model that regresses the outcome of the LDA probabilities on the logarithm of the number of views. The second one includes the control variables, as well as the rest of text features aforementioned.

The results can be examined in the following Table 1:

As we can see, the general explanatory power of the models is satisfactory. LDA probabilities alone already achieve and  $R^2 = 0.106$ , which is boosted up to 0.248 when all the features are taken into account.



Table 1: OLS Regression results

|                                | <i>Dependent variable:</i>                |                           |
|--------------------------------|---|---------------------------|
|                                | Logarithm of the number of visualisations |                           |
|                                | (1)                                       | (2)                       |
| Log of time posted (days)      |   | −0.208*** (0.019)         |
| Log duration of talk (seconds) |   | 0.124 (0.080)             |
| Recurrent speaker              |   | 0.106*** (0.035)          |
| Log of DTM score               |   | −0.005 (0.020)            |
| Log of TF-IDF score            |   | 0.016 (0.022)             |
| Log of number of words         |   | −0.105 (0.070)            |
| Cumulative sentiment score     |   | −0.001* (0.0004)          |
| LDA topic prob.: Society       | 0.510*** (0.151)                          | 0.188 (0.168)             |
| LDA topic prob.: Arts          | −0.716*** (0.200)                         | −0.524** (0.214)          |
| LDA topic prob.: Health        | −0.546*** (0.185)                         | −0.607*** (0.178)         |
| LDA topic prob.: Environment   | −0.656*** (0.187)                         | −0.699*** (0.178)         |
| LDA topic prob.: Politics      | −1.043*** (0.192)                         | −0.807*** (0.225)         |
| LDA topic prob.: Economy       | −0.349** (0.165)                          | −0.318* (0.185)           |
| LDA topic prob.: Entertainment | 0.672*** (0.172)                          | 0.450** (0.182)           |
| LDA topic prob.: Intelligence  | 0.348* (0.191)                            | 0.107 (0.186)             |
| LDA topic prob.: Philosophy    | 0.609*** (0.169)                          | 0.405** (0.172)           |
| LDA topic prob.: Performance   | 0.178 (0.215)                             | −0.031 (0.260)            |
| LDA topic prob.: Technology    | −0.159 (0.198)                            | −0.176 (0.198)            |
| Tagged as: fascinating         |   | 0.150*** (0.042)          |
| Tagged as: funny               |   | 0.253*** (0.058)          |
| Tagged as: inspiring           |   | 0.113*** (0.038)          |
| Tagged as: jaw-dropping        |   | 0.594*** (0.078)          |
| Tagged as topic: AI            |   | −0.088** (0.043)          |
| Tagged as topic: art           |   | −0.175*** (0.047)         |
| Tagged as topic: brain         |   | 0.264*** (0.074)          |
| Tagged as topic: business      |   | 0.120** (0.047)           |
| Tagged as topic: change        |   | −0.205*** (0.060)         |
| Tagged as topic: conference    |   | 0.113*** (0.035)          |
| Tagged as topic: culture       |   | 0.194*** (0.039)          |
| Tagged as topic: design        |   | −0.112** (0.046)          |
| Tagged as topic: global issues |   | −0.127*** (0.044)         |
| Tagged as topic: music         |   | −0.158* (0.093)           |
| Tagged as topic: politics      |   | −0.145** (0.063)          |
| Tagged as topic: science       |   | −0.090** (0.044)          |
| Tagged as topic: technology    |   | −0.119*** (0.040)         |
| Tagged as topic: war           |   | −0.137** (0.070)          |
| Constant                       | 13.795*** (0.124)                         | 15.192*** (0.327)         |
| Observations                   | 1,749                                     | 1,749                     |
| R <sup>2</sup>                 | 0.112                                     | 0.263                     |
| Adjusted R <sup>2</sup>        | 0.106                                     | 0.248                     |
| Residual Std. Error            | 0.653 (df = 1737)                         | 0.599 (df = 1712)         |
| F Statistic                    | 19.850*** (df = 11; 1737)                 | 16.995*** (df = 36; 1712) |

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

## Conclusions

Paragraph 1

Paragraph 2

## References

1. Blei, D.M.; Ng, A.Y. and Jordan, M.I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3 (4-5): pp. 993-1022.
2. Hansen, S.E. (2016). *Text Mining for Economics and Finance*. Course lectures. Barcelona Graduate School of Economics, Barcelona. Spring 2016.
3. *TED Talks: Ideas worth spreading*. Website: <http://www.ted.com/>.