"Do schools kill creativity?" Well, they help analyze popularity!

Anthony Coache & Florence Larose

Université du Québec à Montréal, Département de mathématiques

Objectives

- State a popularity measure;
- Categorize talks;
- Identify characteristics differentiating popular to unpopular talks.

Data

The dataset contains multiple variables collected on September 21st, 2017 about TED talks uploaded to the official TED website.

- We built the adjacency matrix from the list of dictionaries of related talks to create a new variable: number of videos linking to the talk.
- We grouped talks by event type, such as TEDGlobal and TEDx.

However, the published date on TED.com covers the period from 2006 to 2017.

- Recent videos necessarily had less visibility than old videos. With the URL addresses, we used web scraping to collect additional variables on each talk on May 10th, 2018. As expected, recent TED talks had a significant increase in visibility over the two collected dates, which is shown in Figure 1.

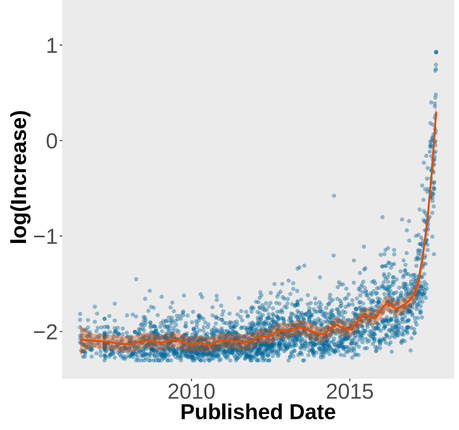

Figure 1: Logarithm of increase of ratings (in %) from Sept. 2017 to May 2018

Popularity Measure

Popularity: *the fact that something is liked, enjoyed or supported by many people.* [1]

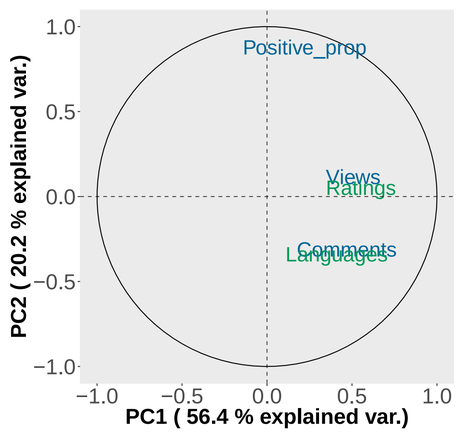

Figure 2: Contribution of variables to variability on first and second principal components

We want to favor both quantity and quality. Therefore, our measure of popularity is defined as the first principal component of principal component analysis (PCA) [2] on languages, proportion of positive ratings and the logarithm of number of comments, ratings and views in 2018. It increases as quantity or quality increases, as illustrated in Figure 2.

Latent Dirichlet Allocation

The goal of this topic modelling algorithm is to assign each talk to one of the k topics, where k is fixed. Let α and β be hyper-parameters and Dir_X be the Dirichlet distribution of order X. Consider D TED talks containing keywords, denoted as w_i, over a vocabulary of V keywords. Let also z_i be the topic index of word w_i.

- Talks are represented as random mixtures θ over latent topics, i.e.

$$\theta_i \sim \text{Dir}_k(\alpha), \quad \forall i = 1, \ldots, D.$$

- Topics are characterized by a distribution ϕ over all keywords, i.e.

$$\phi^j \sim \text{Dir}_V(\beta), \quad \forall j = 1, \ldots, k.$$

We need to compute the posterior distribution of latent topics given TED talks, which is

$$\frac{\mathrm{P}[\phi, \theta, \boldsymbol{z}, \boldsymbol{w} | \alpha, \beta]}{\mathrm{P}[\boldsymbol{w} | \alpha, \beta]}. \qquad (1)$$

However, we cannot compute (1) explicitly in general. Hence, we perform approximate inference techniques, in this case Markov chain Monte Carlo. Using the fact that $\boldsymbol{z}$ is a sufficient statistic for θ and ϕ, we can use a collapsed Gibbs sampler [3]. Thus, topic assignment for each talk is given by

$$\underset{x \in \{1,\ldots,k\}}{\arg\max} \frac{\alpha + n(i,x)}{k\alpha + \sum_{l=1}^{k} n(i,l)}, \ \forall i = 1, \ldots, D,$$

where $n_{i,j}$ is the number of keywords in talk i assigned to topic j.

Random Forest

First, we need to introduce regression trees, a machine learning tool used to divide the predictor space into subregions. Let Y be a vector of continuous observed responses and let $X = [X_1 \cdots X_p]$ be the design matrix. Prediction of a new observation is the mean response value of all the training data falling in the same subregion. At each step, X_j and a cutoff point s are chosen such that they minimize

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

where R_1 and R_2 are defined as

$$R_1(j,s) = \{X | X_j < s\}$$
$$R_2(j,s) = \{X | X_j \geq s\}.$$

One known drawback of decision trees is their high variance. To overcome this overfitting, we can build a random forest [4]:

- Create B bootstrapped samples of our dataset;
- Build a decision tree on each sample;
- Consider a random subset of m predictors at each split;
- Predict a new observation by averaging the predictions of each tree;
- Measure the importance of a predictor by calculating the increase in the out-of-bag prediction error when it is excluded from the model.

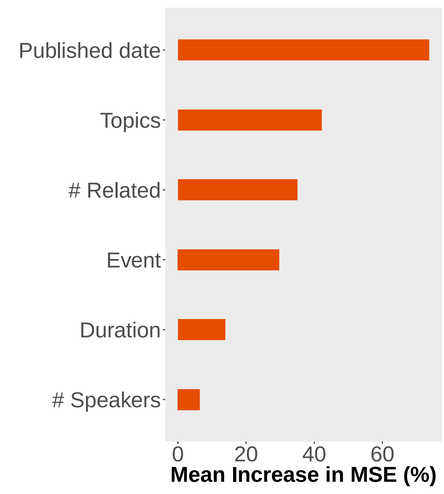

Figure 5: Variable importance from random forest (Out-of-Bag Mean Square Error)

Popularity and Characteristics

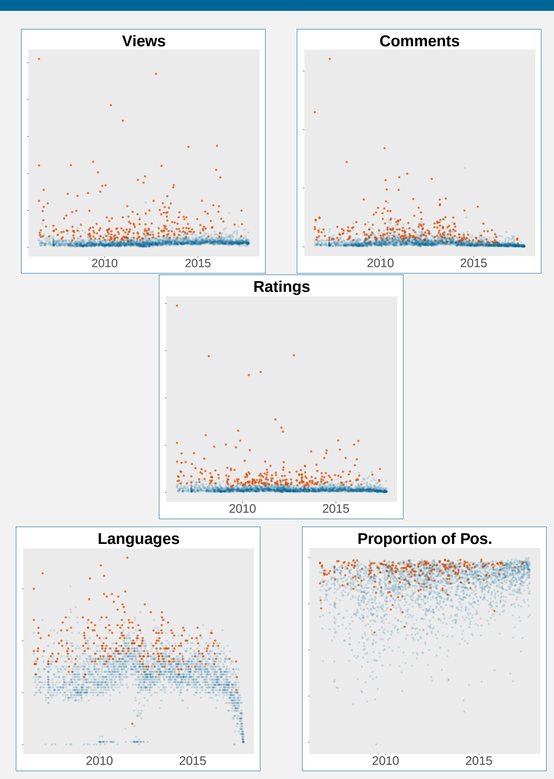

Figure 3: Characteristics of top 10% most popular talks (in orange) based on PC1

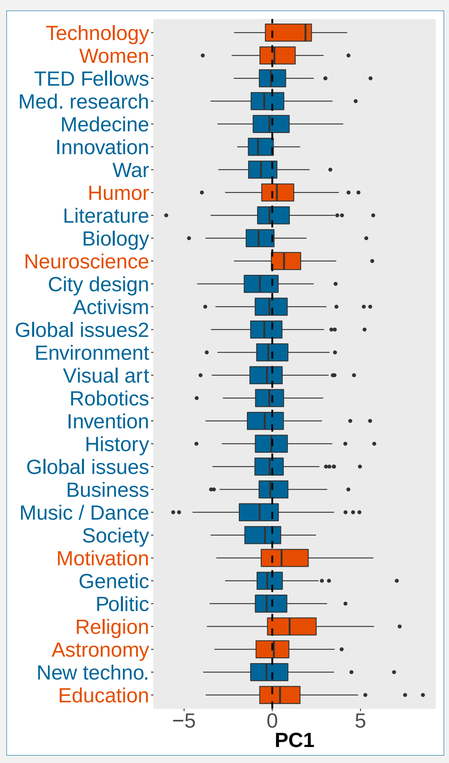

Figure 4: Popularity by topics (positive median in orange)

Results

- Latent Dirichlet allocation [5] with $k = 30$ categorized talks in topics such as

1. Medecine;
2. Religion;
3. Astronomy;
4. Global issues;
5. Business;
6. Education;
7. Environment;
8. Etc.

- Random forest model was built with $B = 500$ trees considering $m = 2$ predictors at each split. The stopping criterion for splitting was a minimum terminal node size of 5.
- Our model included the following predictors:

1. Duration;
2. Published date on TED.com;
3. Event where it was presented;
4. Number of speakers;
5. Number of videos linking to the talk;
6. Topic (as factor).

- Random forest performed prediction on our popularity measure defined above.
- Figure 5 shows the importance of each predictor in our model.

Conclusion

- Our popularity measure favors talks with highest number of views, comments, ratings, languages and positive proportion of ratings, as shown in Figure 3.
- Our model provides some insight on the contribution of each variable for predicting popularity.
- As expected, popularity increases when the talk is on the official TED website for a longer period of time.
- Some topics, such as religion and neuroscience, seem to be more popular than others, as illustrated by Figure 4.

Limitations

- It should be noted that the number k of clusters in latent Dirichlet allocation was chosen arbitrarily and that we assigned labels by ourselves to each cluster. One could argue that any other text mining tool or different parameters would have led to somewhat different topics.
- As dataset varies over time, it would have been interesting to collect those variables periodically. Therefore, we could consider the number of views of each talk as a Poisson point process and perform parametric estimation of the intensity function with an asymmetric function. A popularity measure could then be the mode of this intensity function.

References

[1] Popularity. *Cambridge Advanced Learner's Dictionary & Thesaurus*. Cambridge University Press, 2018.

[2] Jérôme Pagès. *Multiple Factor Analysis by Example Using R*. Chapman and Hall/CRC, 2014.

[3] George Casella and Edward I George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Acknowledgements

We are grateful to Jean-François Coeurjolly and Fabrice Larribe who provided insight and expertise. We also thank Olivier Binette for helpful comments and all members of ÉMoStA for their support.