TED Talks category classification, and recommendation system

Report on Final Project, Machine Learning Applications.

Daniel Kwapien, Olga Bonachera del Pozo, Alejandro Sánchez Díez, Eduardo González Agüero.

Abstract

This final project focuses on the application of machine learning methods on a dataset comprising some TED Talks episodes. Leveraging Natural Language Processing techniques, the project aims to develop category classification and recommendation systems based on the processing of the transcripts of these talks.

Task 1: Text Preprocessing and vectorization

THE DATASET

During this project we considered different datasets or ways to create our own. Since we could collect information from any internet site, such as blogs. We decided to go with TED Talks. These are small talks of around 5 to 20 minutes where a speaker talks about a topic.

TED have their own webpage where they upload every talk and from it, using the json data that the web application receives we can collect, mainly, the following fields:

- Title
- Description
- Speaker
- Topics

In order to collect this data we used a popular scraping software called Scrapy. Using it and based on two GitHub repositories (Corral, n.d.) (Gordon, n.d.). We created our own spyder which is able to collect this data as of 7th of May of 2024.

Once we crawled the website, we collected over 6.000 talks with their corresponding fields, which we stored in a csv file. The reader can check the spyder at the submitted code with this report.

PREPROCESSING

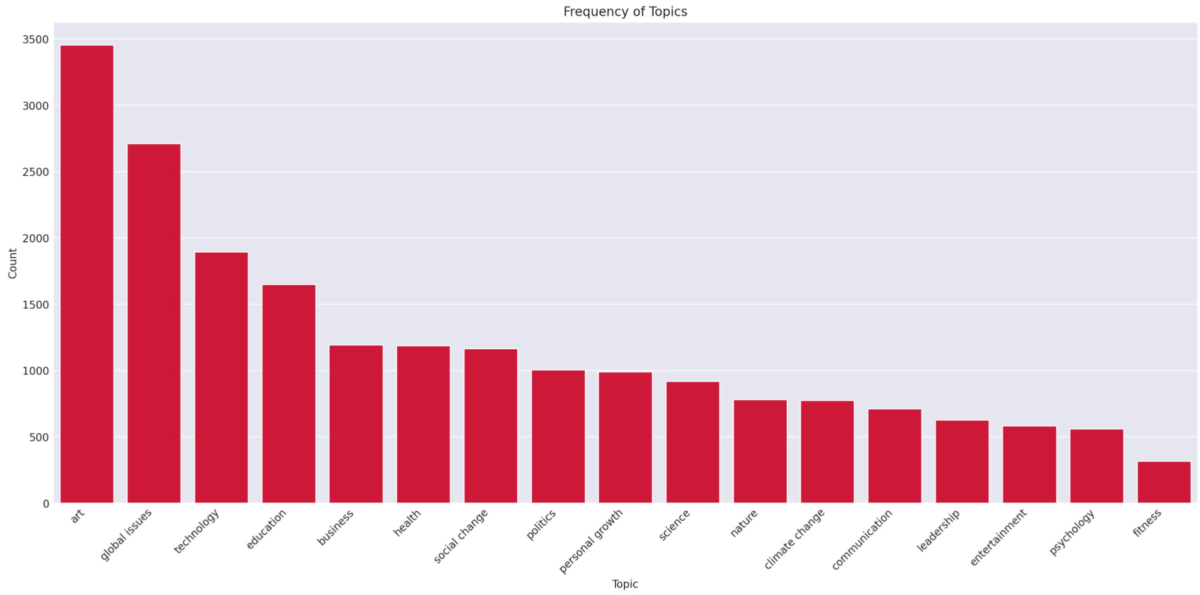
Before starting to work in Natural Language Processing and text vectorization, data cleaning was performed, such as the dropping of NAs, except for the *files* column was completely composed by NAs, probably due to an error while scraping, so it was removed.

After removing the *files* column 18 NAs were left, against a dataset of more than 6000, so those rows with missing values were dropped since such information loss is affordable given the size of the dataset, leaving 6351 rows and 10 columns.

Given the aim of the project, the column *topic* is highly relevant. Looking into it, it can be easily noticed that each document, a TEDtalk video, can, and often does, belong to several topics. We noticed that there were over 350 topics, where many of them just figure in a few videos. Thankfully the TED webpage has a filter where these topics are grouped by more broader ones. Even though we could not scrape them, we discovered that if we selected the broader the topic, the more specific were also selected and this was reflected in the URL. So we just selected every broad topic, 17 topics in total, and passed the URL through a regular expression to extract the topics. This process is also showcased in the submitted notebook.

Once we have our preprocessed dataset, we can check what we will predict. We want to predict the category of a video using the title and description provided.

First, here are our topics and how many times they appear.



We can observe how the most common topic is art and the least common is fitness. Also check that the classes are imbalanced, this is a problem we will treat in the classification part.

TEXT PROCESSING

Although the dataset was very clean since it was directly extracted from a JSON and not scraped through the HTML, we still wanted to ensure that there were no HTML tags or URL's, so we created a `wrangle_text` function that cleans the text and returns lowercase text.

Then, we moved into constructing our **SpaCy pipeline**, in our case we used the well know `en_core_web_md` model. This pipeline contains the following attributes: *tok2vec, tagger, parser, attribute ruler, lemmatizer and NER (Named Entity Recognition)*. We also added after the *NER* component a customized pipe called `normalize_doc_component`, which, as its name suggests, normalizes each doc and returns the final tokens we will use in our text vectorization. This function is basically the lemma of a given token if it has a vector, is alphanumeric, is not a punctuation mark, is not a stopword and is not an entity, such as a person, a place or a date.

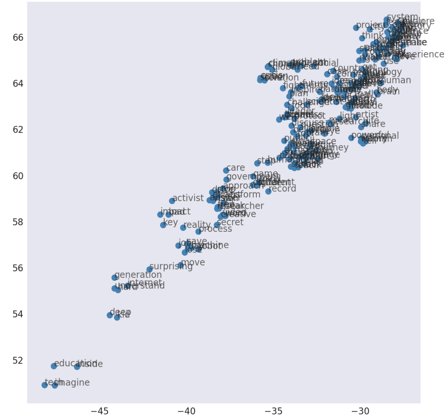
The pipeline ended up having the attributes and then, it was run on the documents. It took approximately one minute to apply the pipeline to all the descriptions in the dataset, using the method `nlp.pipe` with a batch size of 100 to speed up the computation.

TEXT VECTORIZATION

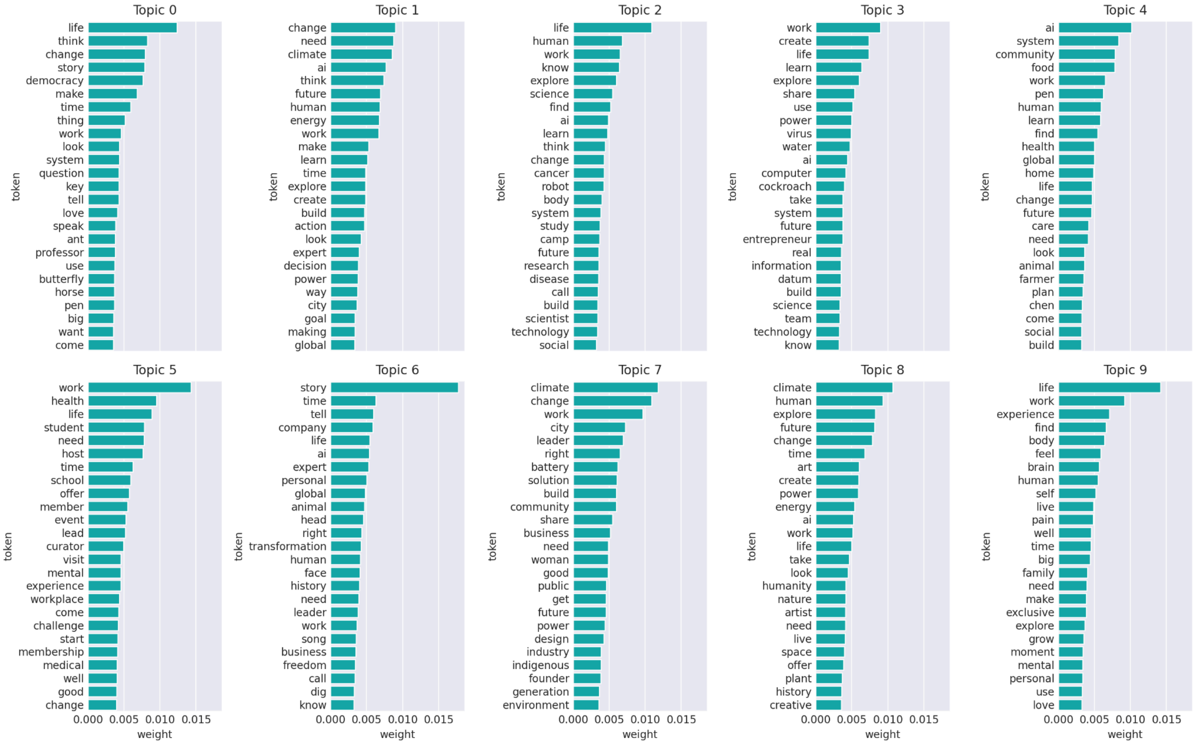
After our text preprocessing we moved into text vectorization. Once we have our tokenized sentences, we created a dictionary using Gensim (Řehůřek 2022), initially with a size of 14211 words. Then we removed words that appeared less than 2 times since they are irrelevant. We finally ended with a dictionary of 8466 terms.

Then, the Bag-of-Words corpus was obtained with the tokenized sentences input in the Bag-of-Words algorithm provided by Gensim, and lastly, this corpus was used to create the TF-IDF model. Finally, we saved the sparse representations as a SciPy Compressed Sparse Column matrix.

After that, we perform Word2Vec vectorization, also using Gensim. We decided to set the vector size to 100, with a window size of 5 words and minimum count of 1. This seems to be the best configuration, taking into account that the maximum size for a description is 71 tokens. Although we expect that the Word2Vec results will not be good since the size of our dictionary is relatively small and the descriptions follow the same semantic structure. As we can see this the representation in a 2-dimensional space of the word embeddings using TSNE.



This allowed us to use the Latent Dirichlet Allocation, a main point of interest in this project as it will allow us to study the topics, and see how good the vectorization of the documents has been.



1. Topic 0: Life, stories, change, democracy, and making an impact. An example of a talk on this topic would be "The Power of Vulnerability" by Brené Brown, which delves into the power of embracing vulnerability and the importance of human connection, addressing themes of change and personal growth.
2. Topic 1: Centers around climate change, AI, future technologies, and energy. "The Case for Optimism on Climate Change" by Al Gore, discussing the latest developments in renewable energy, technological innovations, and the potential for addressing climate change through global cooperation.
3. Topic 2: Life, human experiences, exploration, and learning."Your Elusive Creative Genius" by Elizabeth Gilbert, exploring the concept of creative inspiration and the human experience of creativity, drawing on personal anecdotes and historical examples.
4. Topic 3: Work, creativity, learning, and power. "How to Manage for Collective Creativity" by Linda Hill, discussing strategies for fostering creativity and innovation in the workplace, addressing power dynamics and organizational culture.
5. Topic 4: AI, community, food, health, and human learning. "How AI Can Enhance Our Memory, Work, and Social Lives", by Tom Gruber, explores the potential of artificial intelligence to augment human cognition and improve various aspects of life, including healthcare and social interaction.
6. Topic 5: Work, health, education."Every Kid Needs a Champion" by Rita Pierson, advocating for the importance of educators as champions for students, emphasizing the role of community support and engagement in education.
7. Topic 6: Storytelling, personal experiences, AI, and global issues. "The Danger of a Single Story" by Chimamanda Ngozi Adichie discusses the impact of storytelling on perceptions and stereotypes, highlighting the importance of diverse narratives in understanding global issues.
8. Topic 7: Climate change, urban development, leadership, and community solutions. "How teachers can help students navigate trauma", by Lisa Godwin, explores the complexities of identity and belonging, addressing issues of diversity, inclusion, and community resilience.
9. Topic 8: Climate change, human creativity, exploration, and energy solutions. "Averting the Climate Crisis" by Christiana Figueres - Christiana Figueres discusses the global efforts to address climate change, focusing on the role of innovation, collaboration, and renewable energy solutions.
10. Topic 9: Focuses on life experiences, self-awareness, human body, and emotions. Again, "The Power of Vulnerability" by Brené Brown. See topic 0.

Links to all videos can be found in the Reference section at the end of this document.

2. Task II – Machine Learning model

2.1 Task – Classification

In this section we will determine the performance with which a model can be trained to classify any document into the 17 different categories they were originally divided into.
It is important to note that any document may belong to two or more classes and the majority of documents have two or more labels assigned.

This might be a difficult classification problem. We first started by fitting a Random Forest and a OneVsRest SVM, but they had a bad performance since the classes were not balanced, so for the topics that appeared a lot the classifier did a fair job, but for the ones that appeared the least it had a poor performance.

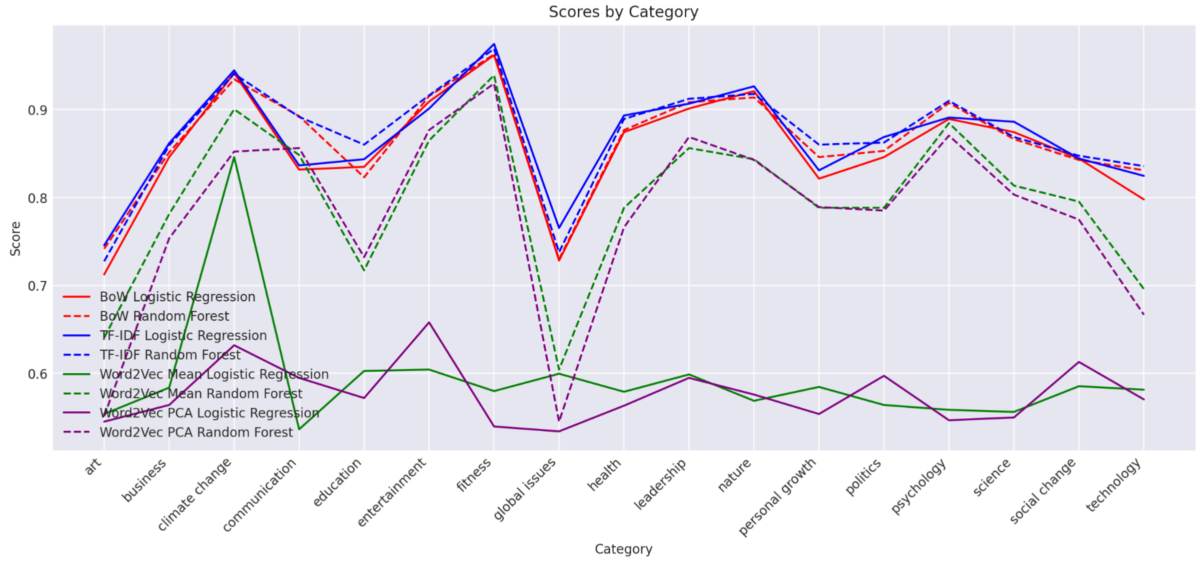
So we decided to binarize the different categories and try to classify them using a Logistic Regression and a Random Forest. We first fitted the models using Grid Search in order to find the best parameters and Cross-validation. We also oversampled the minority class using SMOTE so the classes would be balanced. These are the results using the best parameters.

Category	Bag-Of-Words		TF-IDF	
	Logistic Regression	Random Forest	Logistic Regression	Random Forest
Art	71.24%	74.40%	74.56%	72.74%
Business	84.59%	86.25%	86.17%	85.94%
Climate change	94.15%	92.81%	94.47%	94.07%
Communication	83.17%	87.44%	83.65%	89.17%
Education	83.80%	82.46%	84.36%	86.01%
Entertainment	90.91%	90.91%	90.12%	91.63%
Fitness	96.21%	94.94%	97.47%	96.91%
Global issues	72.82%	72.03%	76.54%	73.77%
Health	87.44%	87.36%	89.33%	88.94%
Leadership	90.12%	90.52%	90.67%	91.23%
Nature	92.10%	92.73%	92.65%	91.78%
Personal growth	82.14%	92.73%	83.09%	86.01%
Politics	84.59%	88.62%	86.88%	86.25%
Psychology	88.94%	92.33%	89.09%	90.99%
Science	87.44%	87.44%	88.62%	86.88%
Social change	84.43%	83.41%	84.44%	84.75%
Technology	79.77%	81.75%	82.46%	83.57%
Mean	**85.50%**	**86.45%**	**86.74%**	**87.10%**

Then, we used the Word2Vec embedding as the input to our classifiers. The first approach consisted in computing the mean of the embedding. This way we will have a single vector of size 100 as the input for our models.

The second approach consisted in applying PCA over the vectors. Since PCA requires a matrix of fixed length, we padded the vectors of a sentence to a fixed length and used the result matrix 100 rows and, in our case, 71 columns, as the input for the PCA. Since we want our vector to have size 100, we are interested just in the first component. We repeat this for each document and obtain an input for our models. These are the results.

Category	**Mean**		**PCA**	
	Logistic Regression	**Random Forest**	**Logistic Regression**	**Random Forest**
Art	55.37%	64.06%	54.50%	55.13%
Business	58.37%	78.12%	56.39%	75.35%
Climate change	84.59%	90.04%	63.19%	85.22%
Communication	53.63%	84.48%	59.47%	85.62%
Education	60.26%	71.72%	57.18%	73.22%
Entertainment	60.42%	86.49%	65.79%	87.67%
Fitness	57.97%	93.91%	53.94%	92.96%
Global issues	59.95%	60.42%	53.39%	54.50%
Health	57.89%	78.83%	56.31%	76.61%
Leadership	59.87%	85.62%	59.47%	86.88%
Nature	56.87%	84.36%	57.58%	84.28%
Personal Growth	58.45%	78.83%	55.37%	78.90%
Politics	56.39%	78.83%	59.71%	78.51%
Psychology	55.84%	88.46%	54.66%	87.04%
Science	55.60%	81.35%	54.97%	80.33%
Social change	58.53%	79.54%	61.29%	77.48%
Technology	58.13%	69.58%	57.03%	66.66%
Mean	**59.30%**	**79.70%**	**57.66%**	**78.02%**



As we can see by the plot and the tables, the models performed a lot better with the Bag-of-Words and TF-IDF representations, being TF-IDF the representation that gives the best results. This might be because descriptions are a type of text that is always made with the same structure, it is not like a summary, but like an abstract, so usually the redactors will write it the same way and use similar words for similar topics.

On the other hand, the Word2Vec representations perform poorly, mainly because of the small size of the dictionary and the type of text we are working with, so the embeddings does not capture semantics well.

Random Forest tends to perform slightly better than Logistic Regression in the frequency based representations and outperforms completely in the Word2Vec representations. This might indicate that working an ensemble of decision trees has better performance in this type of problem than generalized linear models.

Finally note that there are topics that always get a better or worse accuracy. This might be because the topics that have low accuracy are topics that appear more and are used more broadly, englobing a huge variety of topics with different descriptions. On the other hand topics that appear less have higher accuracy because they are more specific and the descriptions they have use a more specific set of words.

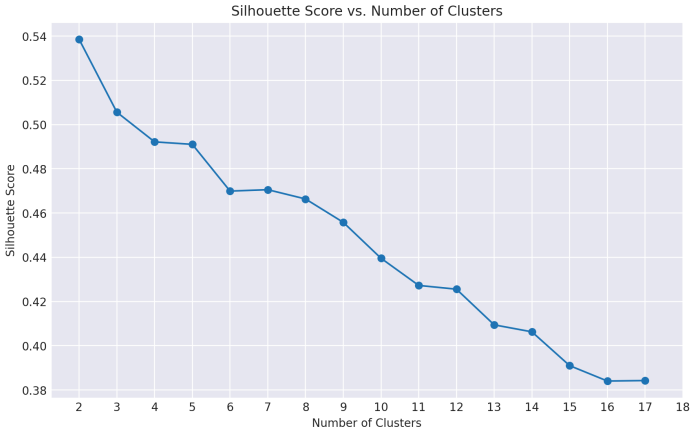
2.2 Task – Clustering

The amount of labels in the data, and the broadness of the fields they represent, meant that most talks belong to several of these topics. An example of this we have seen above, in the topic modeling, when a video could be considered to belong to topics 0 and 9. So, could an unsupervised Machine Learning model come up with more radical labeling that would minimize the number of overlapping clusters?

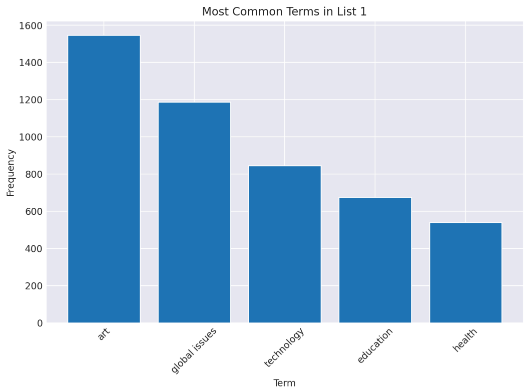
To test the model we used the silhouette score, which provides a number between -1 and 1, where a number close to -1 indicates bad clustering, and 1 optimal clustering, with everything in between representing a greater or lesser degree of overlap or error.

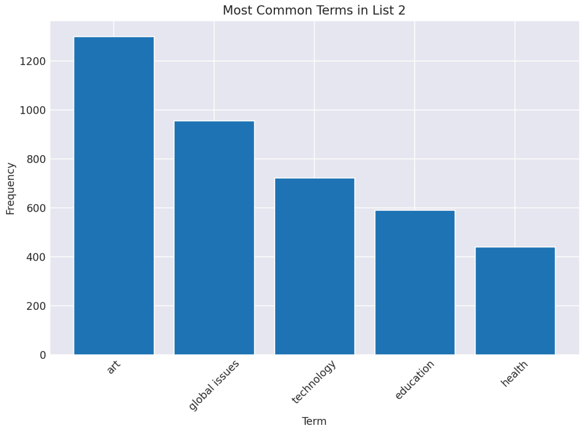
An original clustering algorithm was applied to PCA-embedded observations, from the last classification task carried out. We obtained an optimal *k* of 2, but when looking at the data found that only 0.7% of data points were being classified to one of the clusters. We decided to do away with this representation.

We used Word2Vec next. Results were, on face value, more coherent. The optimal value of *k* was obtained using silhouette score. The following plot represents the scores.

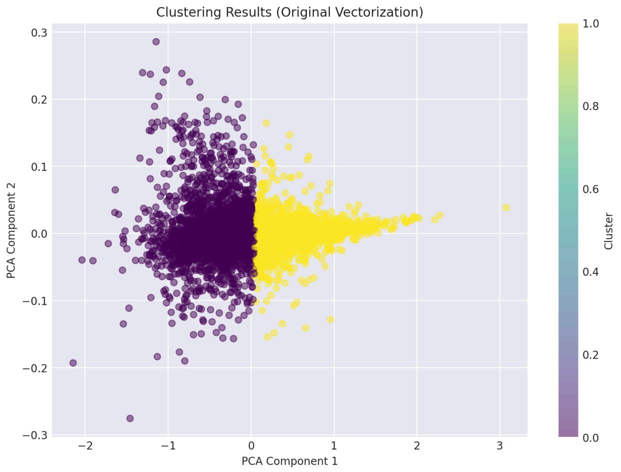


It can be seen that the optimal value still is 2, but that the range of values is quite reduced (a minimum of 0.38). These seemed to be encouraging results, and we could at this point come up with an interpretation— a distinction between talks on science and talks on the arts. This however, was getting too much ahead. We devices a way of obtaining the most common labels per each cluster after running a 2-Means algorithm. This was the result:





As can be deduced, this is a poor result. The algorithm 'missed' (it has no way of knowing) the point that we were trying to put forth, and essentially split the data in half, such that each had the same balance of labels. To observe this better, a PCA representation of the vectors was obtained, to be able to reduce the data to 2 dimensions, and plot the clusters:

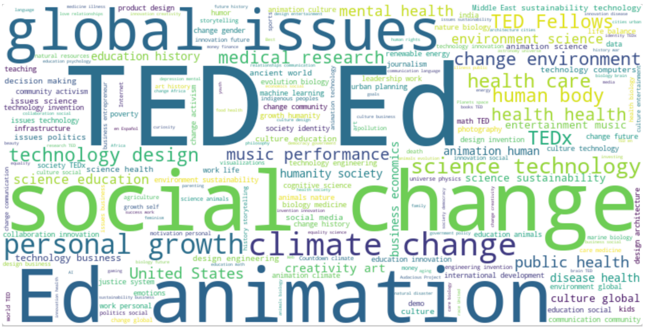


It was probably foolish to think that such a basic tool would be able to capture the nuances in the vector representation, straightforward as it is. As we can see from the PCA (which was helpful in improving the performance of the Word2Vec vectorization in the last classification model) there are no discernible clusters, and while the association of terms between labels and clusters is apt conceptually, it is not the case when performing Machine Learning models.

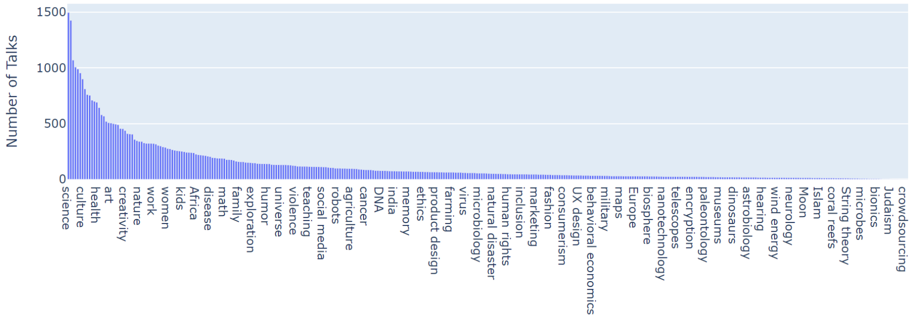
3 Task - Dash

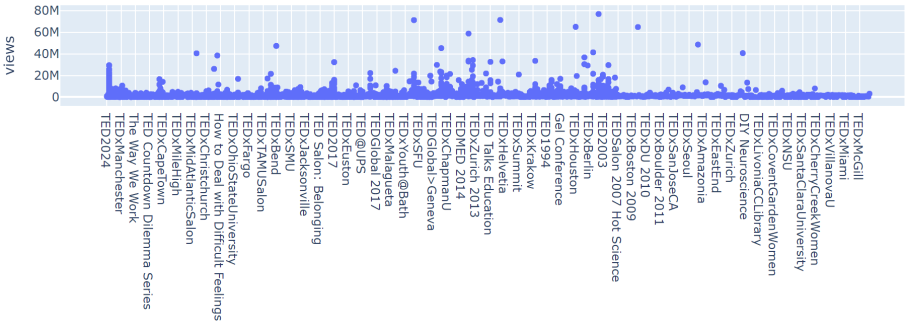
In order to provide more insights into the distribution of topics covered in TED talks, a Dashboard has been implemented.

Firstly we can observe a visual and intuitive cloud of the most spoken and repeated topics in the descriptions of the talks and conclude some of the most typical topics like global issues, social change, climate change, etc.

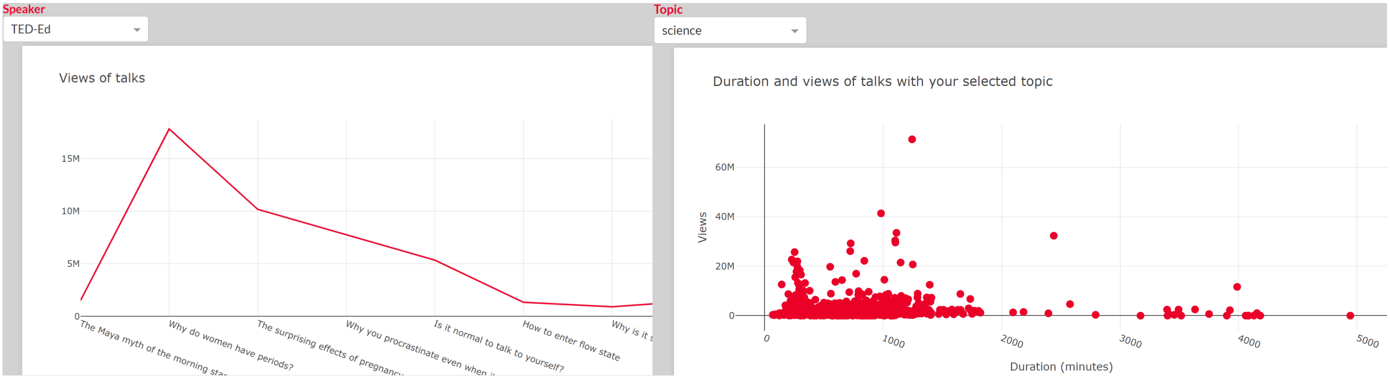


To further understand some of the most important features, we plotted the frequency of talks about each topic - concluding the most relevant ones- and the events with the amount of views - getting insights about the most popular events.





In the following graphs, a user can interact with the plott by entering a desired speaker (then a graph with the talks he/she has given with the corresponding number of views is displayed) or a topic (and a graph is shown with the views and the duration of the talk).



Speaker: TED-Ed

Topic: science

Views of talks
15M
10M
5M
0
The Maya myth of the morning sta
Why do women have periods?
The surprising effects of pregnanc
Why you procrastinate even when i
Is it normal to talk to yourself?
How to enter flow state
Why is it s
Duration and views of talks with your selected topic
Views
60M
40M
20M
0
0
1000
2000
3000
4000
5000
Duration (minutes)

REFERENCES

Corral, Miguel. n.d. “corralm/ted-scraper: 🎙️ TED Talks web scraper.” GitHub. Accessed May 7, 2024. https://github.com/corralm/ted-scraper.

Gordon, Matthew A. n.d. “Ted-scraper.” Github. Accessed May 7, 2024. https://github.com/matthew-a-gordon/ted-scraper/blob/main/ted-scraper/spiders/tedscraper.py.

Brown, B. (n.d.). *The power of vulnerability*. Brené Brown: The power of vulnerability | TED Talk. https://www.ted.com/talks/brene_brown_the_power_of_vulnerability

Gore, A. (n.d.). *The case for optimism on climate change*. Al Gore: The case for optimism on climate change | TED Talk. https://www.ted.com/talks/al_gore_the_case_for_optimism_on_climate_change

Gilbert, E. (n.d.). *Your elusive creative genius*. Elizabeth Gilbert: Your elusive creative genius | TED Talk. https://www.ted.com/talks/elizabeth_gilbert_your_elusive_creative_genius

Hill, L. (n.d.). *How to manage for collective creativity*. Linda Hill: How to manage for collective creativity | TED Talk. https://www.ted.com/talks/linda_hill_how_to_manage_for_collective_creativity

Gruber, T. (n.d.). *How AI can enhance our memory, work and Social Lives*. Tom Gruber: How AI can enhance our memory, work and social lives | TED Talk. https://www.ted.com/talks/tom_gruber_how_ai_can_enhance_our_memory_work_and_social_lives

Pierson, R. (n.d.). *Every kid needs a Champion*. Rita Pierson: Every kid needs a champion | TED Talk. https://www.ted.com/talks/rita_pierson_every_kid_needs_a_champion

Adichie, C. N. (n.d.). *The danger of a single story*. Chimamanda Ngozi Adichie: The danger of a single story | TED Talk. https://www.ted.com/talks/chimamanda_ngozi_adichie_the_danger_of_a_single_story

Godwin, L. (n.d.). *How teachers can help students navigate trauma*. Lisa Godwin: How teachers can help students navigate trauma | TED Talk. https://www.ted.com/talks/lisa_godwin_how_teachers_can_help_students_navigate_trauma

Mistry, P. (n.d.). *The thrilling potential of SixthSense technology*. Pranav Mistry: The thrilling potential of SixthSense technology | TED Talk. https://www.ted.com/talks/pranav_mistry_the_thrilling_potential_of_sixthsense_technology