Mon, May 2

PFCH INFO

STUDENT

Priyanka Gangwal

Analyzing TED Talks Data

Table of Contents

Dataset

The dataset was downloaded from Kaggle. It contains various information from the first online published Ted Talk till 2017. The dataset has information on the url, tags used, comments, views, transcripts, ratings, etc.

Area of interest

Analyzing qualitative data–by doing sentiment analysis along with learning Python to create data visualizations motivated me to work towards this dataset.

Deliverables

Github repository and a research report.

II Research Areas

- **Sentiment Analysis**
  - Sentiment scores and magnitude of titles, descriptions and transcripts
  - Analyze text per min sentiments
- **Visualizations**
  - Top 10 and least 10 viewed videos
  - All sentiment scores and top 10 views, bottom 10 views.
  - All magnitude and views - top 10 views, bottom 10 views.
  - Sentiments per minute - top 5 and bottom 5 viewed videos
- **Other text analysis**
  - Views received by every tag
  - Word clouds of transcripts - top 20 and bottom 20 views
- **Webscraping**
  - Author nationality details from Wikipedia
- **Correlations & Predictions**
  - Does more "laughter" in audience reaction mean more views?
  - Funniness (laugh count / mins), and length > can we predict the number of views?

- **Sentiment Analysis**
  - Sentiment scores and magnitude of titles, descriptions and transcripts
  - Analyze text per min sentiments
- **Visualizations**
  - Top 10 and least 10 viewed videos
  - All sentiment scores and top 10 views, bottom 10 views.
  - All magnitude and views - top 10 views, bottom 10 views.
  - Sentiments per minute - top 5 and bottom 5 viewed videos
- **Other text analysis**
  - Views received by every tag
  - Word clouds of transcripts - top 20 and bottom 20 views
- **Webscraping**
  - Author nationality details from Wikipedia
- **Correlations & Predictions**
  - Does more "laughter" in audience reaction mean more views?
  - Funniness (laugh count / mins), and length > can we predict the number of views?

Sentiment Analysis

Approach

- Used Google Cloud Natural Language API to conduct sentiment analysis on the title, description and the transcripts.
- It gave scores and magnitude
- Overall data and did sentiment analysis on talks per minute on videos with normalized views - divided the views by the number of years the video has been online.

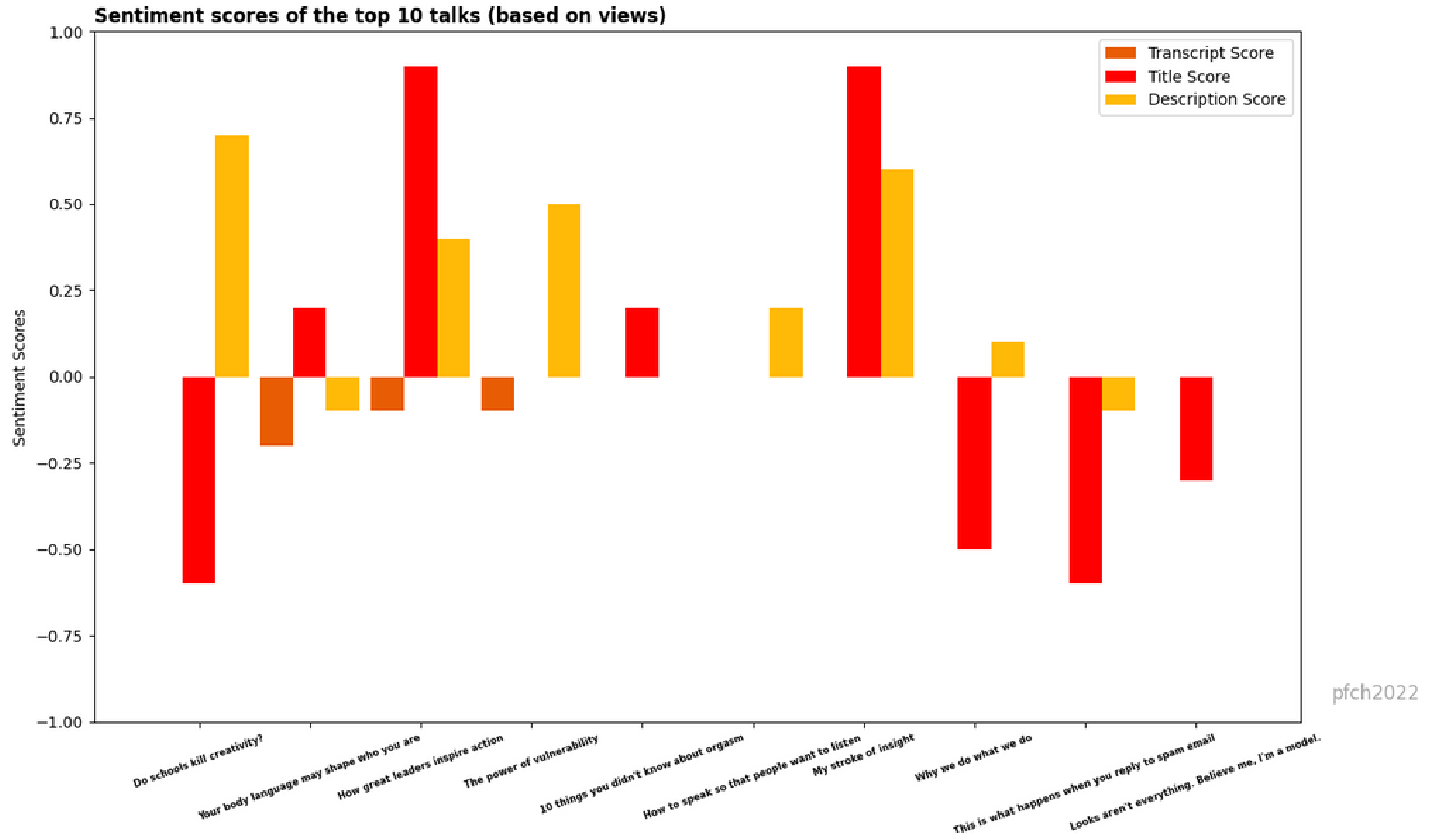
	transcript_x	minute	score	magnitude
s_to_the_future_of_medicine	Thank you. It's really an honor and a privilege to be here spending my last day as a teenager. Today I want to talk to you about the future, but first I'm going to tell you a bit about the past. My story starts way before I was bor	1	0.0	3.5
s_to_the_future_of_medicine	They were living in Hungary, and my mother was born. And when my mother was two years old, the Hungarian revolution was raging, and they decided to escape Hungary. They got on a boat, and yet another divergence — t	2	-0.1000000014901160	2.799999952316280
s_to_the_future_of_medicine	I found this passion not far from here, actually, when I was nine years old. My family was on a road trip and we were in the Grand Canyon. And I had never been a reader when I was young — my dad had tried me with the Ha	3	0.0	2.5999999046325700
s_to_the_future_of_medicine	I wanted to be like the explorers I'd read about in the book, who went into the jungles of Africa, went into the research labs and just tried to figure out what this deadly virus was. So from that moment on, I read every medical	4	-0.1000000014901160	3.5
s_to_the_future_of_medicine	At that time, I was really interested in neuroscience and wanted to do a research project in neurology — specifically looking at the effects of heavy metals on the developing nervous system. So I started that, and worked in hi	5	0.0	2.200000047683720
s_to_the_future_of_medicine	And being naive about the whole field, I kind of thought, "Oh, you have cell death in Alzheimer's which is causing the memory deficit, and then you have this compound — purine derivatives — that are promoting cell growth."	6	0.0	3.400000095367430
s_to_the_future_of_medicine	And this is really what I want to talk to you about today — about cancer. At first when I heard of cancer stem cells, I didn't really know how to put the two together. I'd heard of stem cells, and I'd heard of them as the panacea	7	0.0	2.5
s_to_the_future_of_medicine	The more I read, the more I looked at cancer differently and almost became less fearful of it. It seems that cancer is a direct result to injury. If you smoke, you damage your lung tissue, and then lung cancer arises. If you drink	8	-0.1000000014901160	6.800000190734860
s_to_the_future_of_medicine	And cancer is originating in the lung trying to repair — because you have this excessive proliferation of these remarkable cells that really have the potential to become lung tissue. But it's almost as if the body has originated t	9	-0.1000000014901160	3.700000047683720
s_to_the_future_of_medicine	Shouldn't we think about manipulation, rather than elimination? If somehow we can cause these cells to differentiate — to become bone tissue, lung tissue, liver tissue, whatever that cancer has been put there to do — it wou	10	-0.1000000014901160	2.299999952316280
s_to_the_future_of_medicine	So, I looked further into it, found as many articles as I could, and it was amazing — because it turned out that it was very rare. Some articles even went as far as to say that skeletal muscle tissue is resistant to cancer, and fur	11	0.0	3.900000095367430
s_to_the_future_of_medicine	So the fact that not only did cancer not seem to originate in skeletal muscles, but cancer didn't seem to go to skeletal muscle — there seemed to be something here. So these articles were saying, you know, "Skeletal — met	12	-0.3000000119209290	3.0
s_to_the_future_of_medicine	Meaning that you have muscle cells, but they're not dividing, so it doesn't seem like a good target for cancer to hijack. But then again, this fact that the metastases didn't go to skeletal muscle made that seem unlikely. And fu	13	-0.3000000119209290	2.700000047683720
s_to_the_future_of_medicine	Some of my hypotheses are that when you first think about skeletal muscle, there's a lot of blood vessels going to skeletal muscle. And the first thing that makes me think is that blood vessels are like highways for the tumor	14	-0.2000000029802320	2.400000095367430
s_to_the_future_of_medicine	And one article that really stood out to me when I was just reading about this, trying to figure out why cancer doesn't go to skeletal muscle, was that it had reported 16 percent of micro-metastases to skeletal muscle upon au	15	0.1000000014901160	2.900000095367430
s_to_the_future_of_medicine	It's grabbing its blood vessels for itself. Therefore, when a tumor comes into skeletal muscle tissue, it can't get a blood supply, and can't grow. So this suggests that maybe if there is an anti-angiogenic factor in skeletal musc	16	-0.1000000014901160	2.5
s_to_the_future_of_medicine	And, furthermore, when skeletal muscle is injured, that's what causes chemokines — these signals saying, "Cancer, you can come to me," the "go signs" for the tumors — it causes them to highly express these chemokines.	17	0.0	2.5
s_to_the_future_of_medicine	So, is it possible that the tumor cells are going to the skeletal muscle tissue, but once in contact inside the skeletal muscle tissue, MyoD acts upon these tumor cells and causes them to become skeletal muscle cells? Maybe	18	-0.2000000029802320	4.5
s_to_the_future_of_medicine	It's different when a bacteria comes into the body — that's a foreign object — we want that out. But when the body is actually initiating a process and we're calling it a disease, it doesn't seem as though elimination is the righ	19	0.1000000014901160	3.799999952316280
shares_organic_designs	My name is Lovegrove. I only know nine Lovegroves, two of which are my parents. They are first cousins, and you know what happens when, you know — So there's a terribly weird freaky side to me, which I'm fighting with a	1	0.0	6.699999809265140
shares_organic_designs	I could have been, perhaps, but I work in this world where I trust my instincts. So I am a 21st-century translator of technology into products that we use everyday and relate beautifully and naturally with. And we should be de	2	0.2000000029802320	4.599999904632570
shares_organic_designs	Here you see the machining, the milling of a block of acrylic. This is what I show to the client to say, "That's what I want to do." At that point, I don't know if that's possible at all. It's a seductor, but I just feel in my bones that	3	-0.1000000014901160	6.300000190734860
shares_organic_designs	Each bottle is different, meaning the water level will give you a different shape. It's mass individualism from a single product. It fits the hand. It fits arthritic hands. It fits children's hands. It makes the product strong, the tessel	4	0.2000000029802320	7.800000190734860
shares_organic_designs	But I just know that nature — nature improves with ever-greater purpose that which once existed, and that strangeness is a consequence of innovative thinking. When I look at these things, they look pretty normal to me. But	5	0.0	6.699999809265140
shares_organic_designs	And I had access to all the best girls. It was fabulous. All the guys in the rugby team couldn't understand. Anyway — this is a meringue. This is another sample I have. A meringue is made exactly the same way, in my estimati	6	0.1000000014901160	6.199999809265140
shares_organic_designs	It cost 1.7 million dollars to develop. It's called "Go," by Bernhardt, USA. It went into Time magazine in 2001 as the new language of the 21st century. Boy. For somebody growing up in Wales in a little village, that's enough. I	7	0.1000000014901160	6.599999904632570
shares_organic_designs	I set out to look at natural forms. If you took the idea of fractal technology further, take a membrane, shrinking it down constantly like nature does — that could be a seat for a chair. It could be a sole for a sports shoe. It could	8	0.2000000029802320	6.0
shares_organic_designs	That is the power of organic design. It contributes immensely to our — sense of being, our sense of relationships with things, our sensuality and, you know, the sort of — even the sort of socio-erotic side, which is very impor	9	0.1000000014901160	5.900000095367430

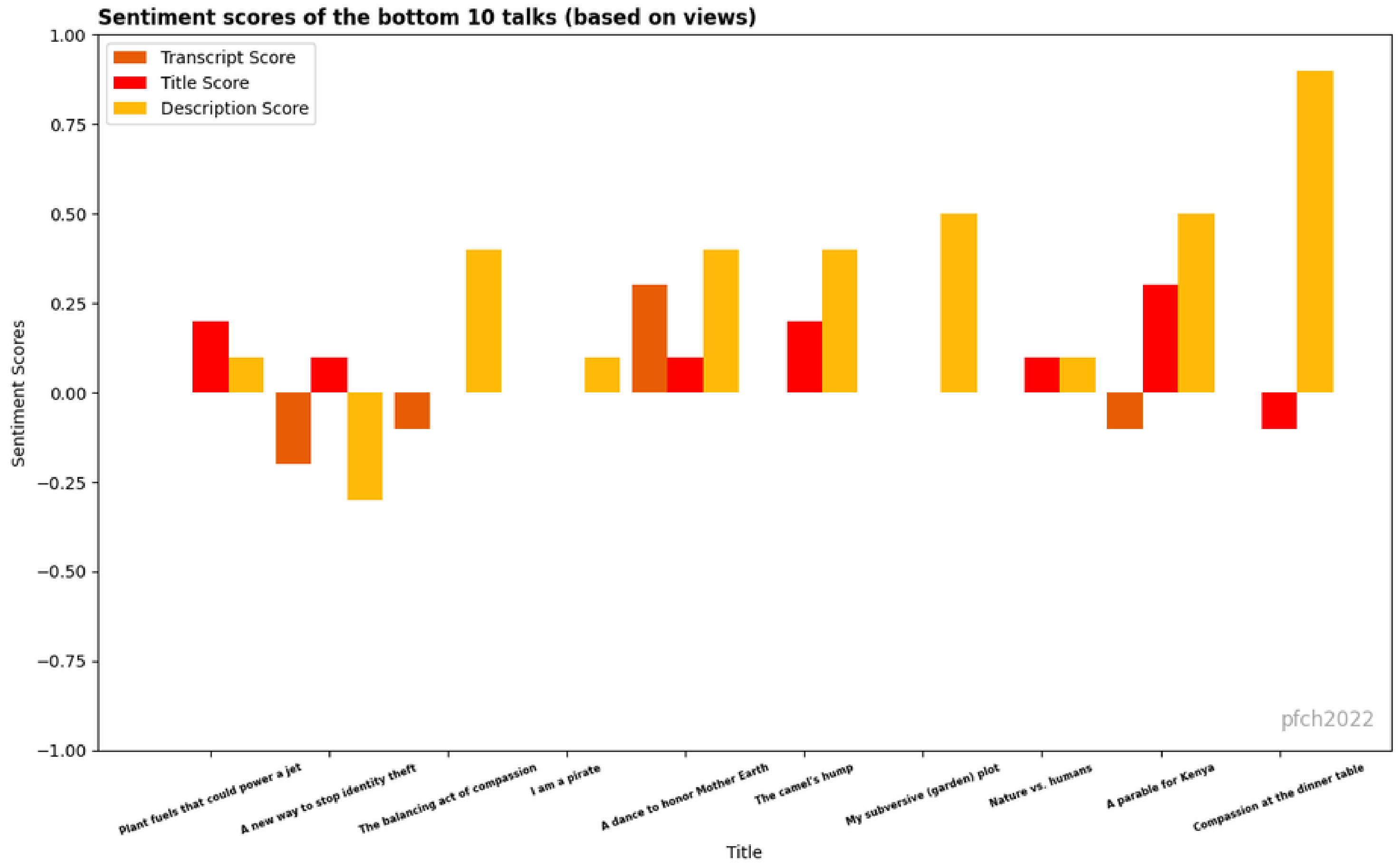
II Research Areas

- Sentiment Analysis
  - Sentiment scores and magnitude of titles, descriptions and transcripts
  - Analyze text per min sentiments
- Visualizations
  - Top 10 and least 10 viewed videos
  - All sentiment scores and top 10 views, bottom 10 views.
  - All magnitude and views - top 10 views, bottom 10 views.
  - Sentiments per minute - top 5 and bottom 5 viewed videos
- Other text analysis
  - Views received by every tag
  - Word clouds of transcripts - top 20 and bottom 20 views
- Webscraping
  - Author nationality details from Wikipedia
- Correlations & Predictions
  - Does more "laughter" in audience reaction mean more views?
  - Funniness (laugh count / mins), and length > can we predict the number of views?

Visualizations

Comparing sentiment scores of the top 10 most and least viewed videos

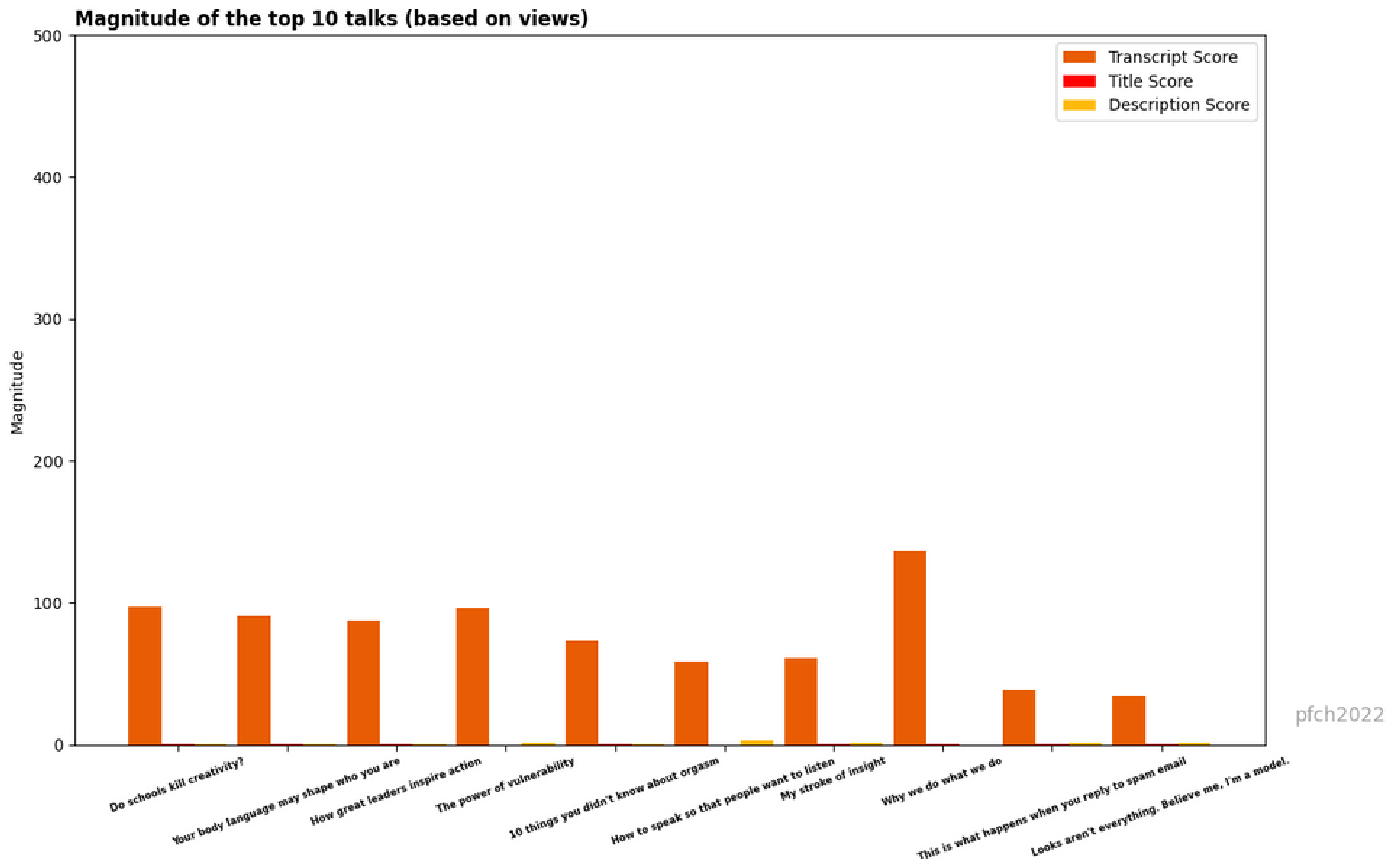


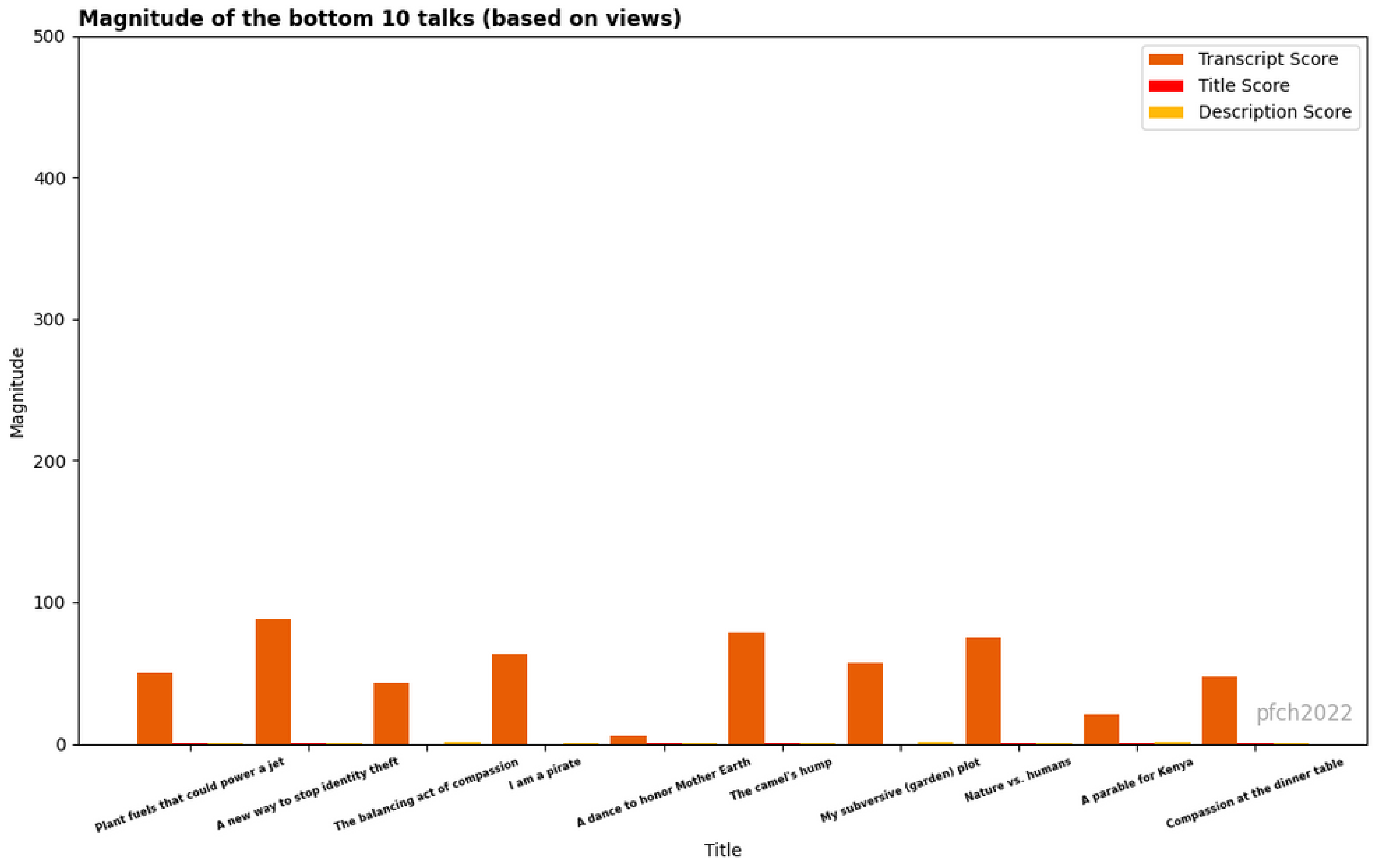


- Compares transcript, description and title sentiments
- Most videos have positive descriptions
- The top viewed videos scores were higher all across
- Transcripts overall seem to be neutral
- Recipe for a successful talk - The titles have to clearly pick a side

Visualizations

Comparing sentiment magnitude of the top 10 most and least viewed videos

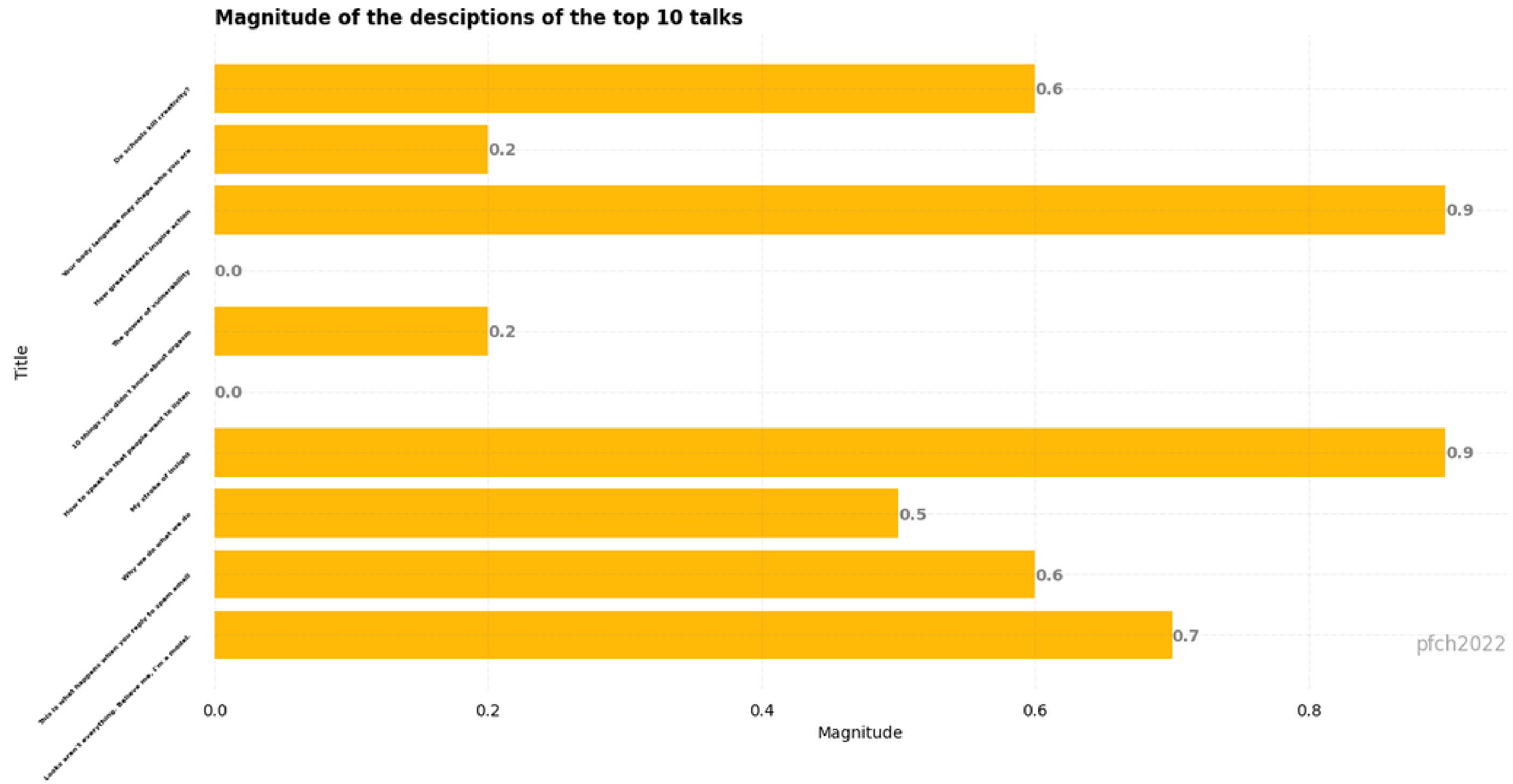


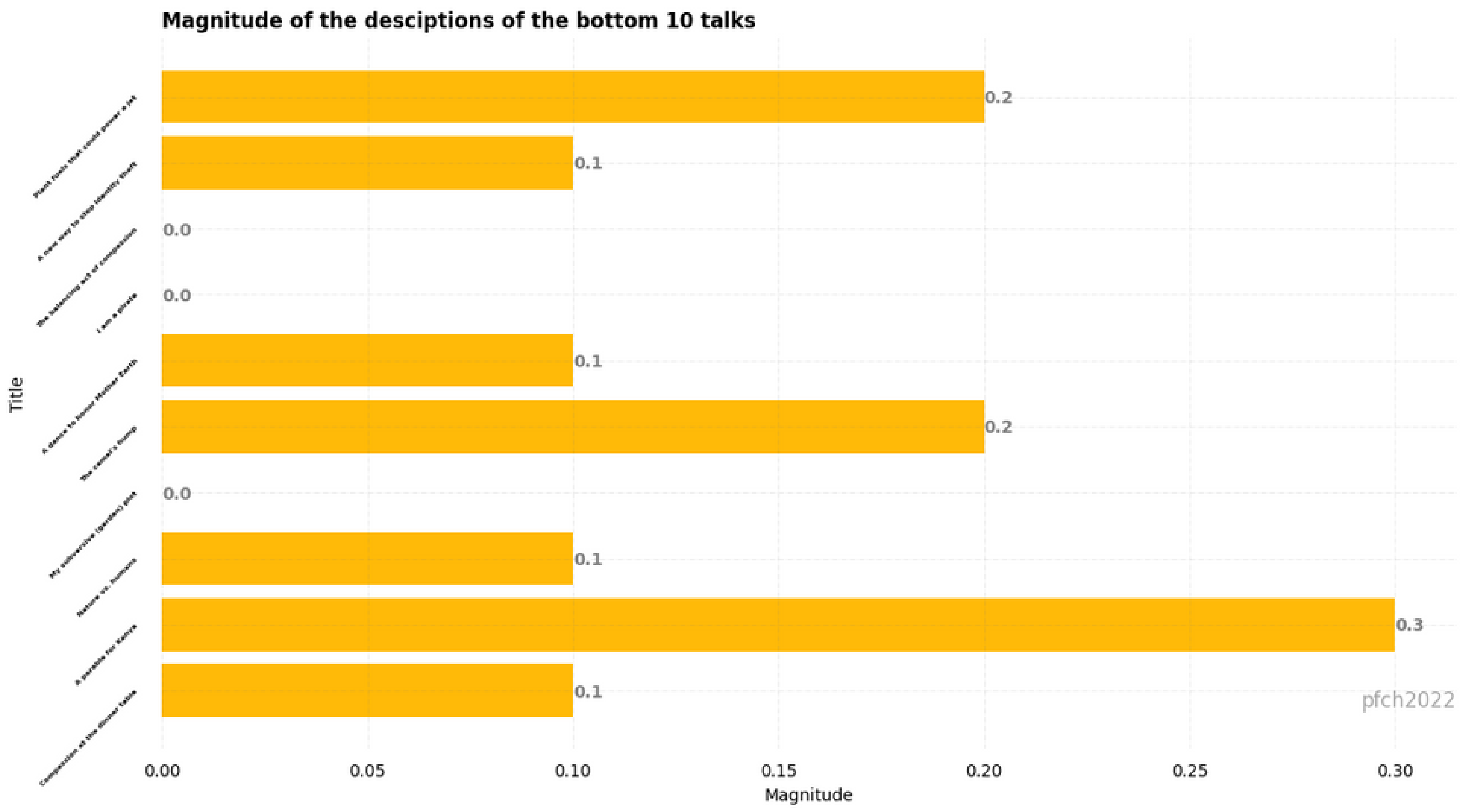


- Top viewed videos have higher magnitude in their transcripts

Visualizations

Comparing sentiment magnitude of the top 10 most and least viewed videos

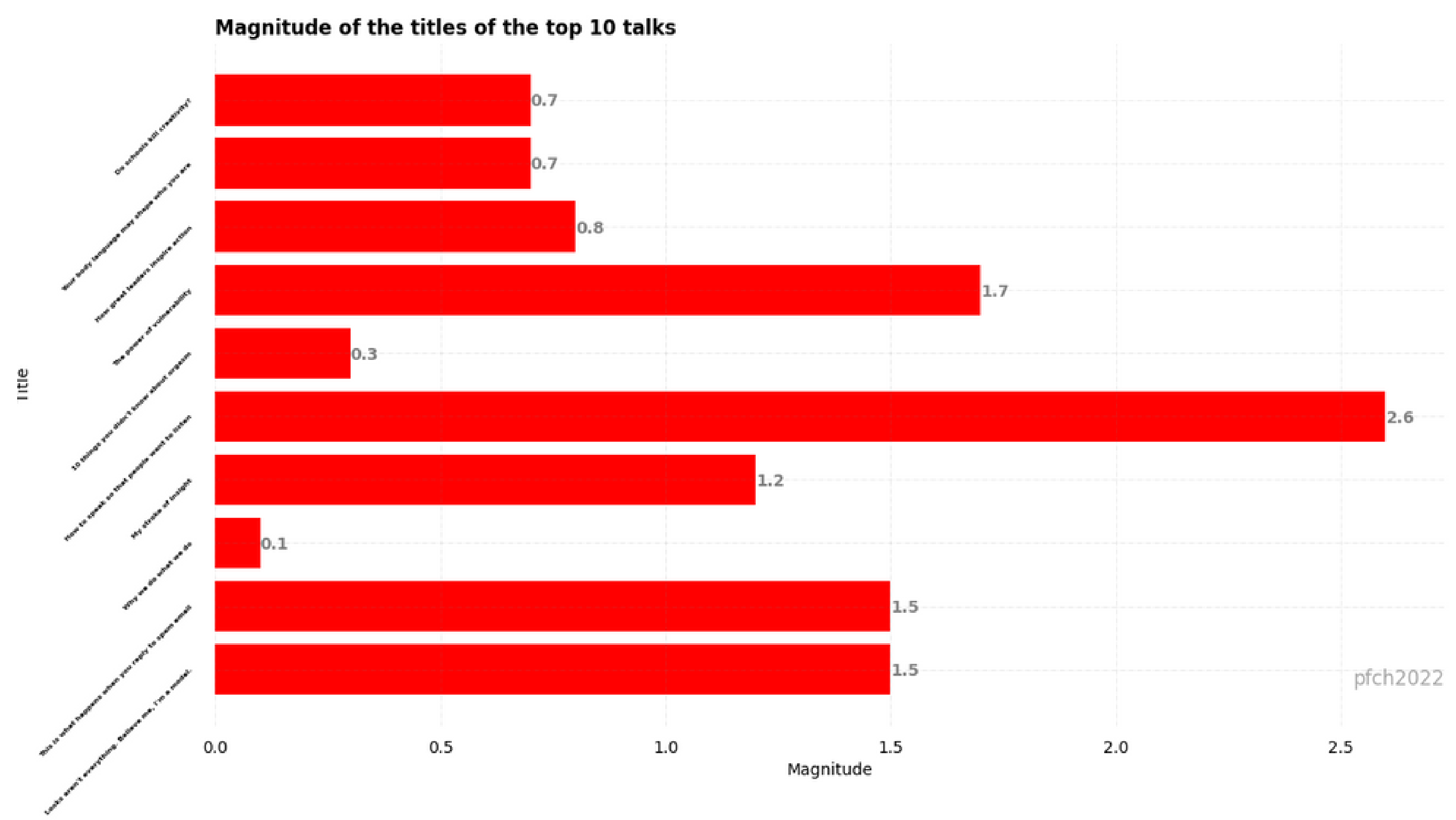


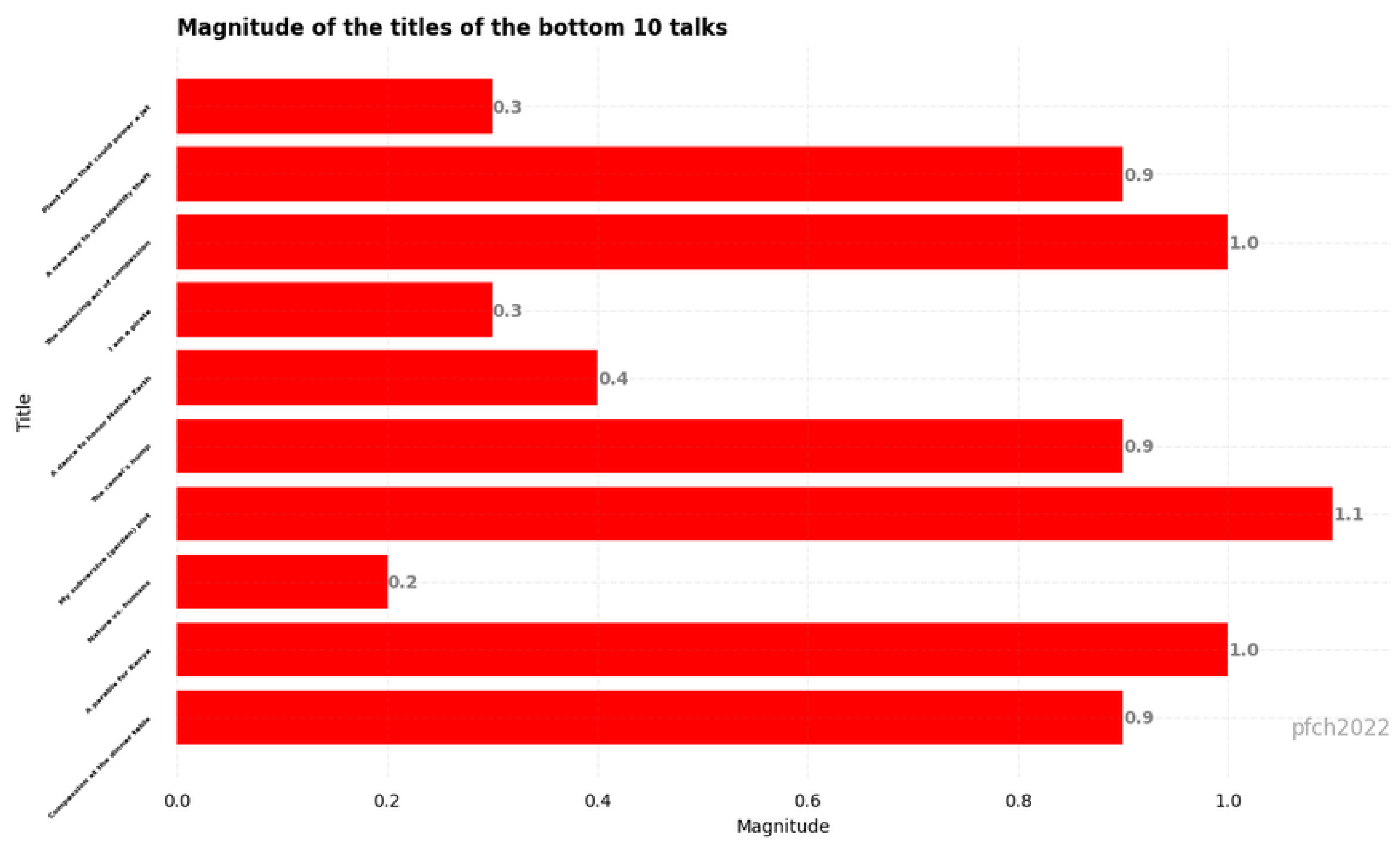


- Top viewed videos have higher magnitude in their transcripts
- Top viewed videos have lower magnitude in their descriptions

Visualizations

Comparing sentiment magnitude of the top 10 most and least viewed videos



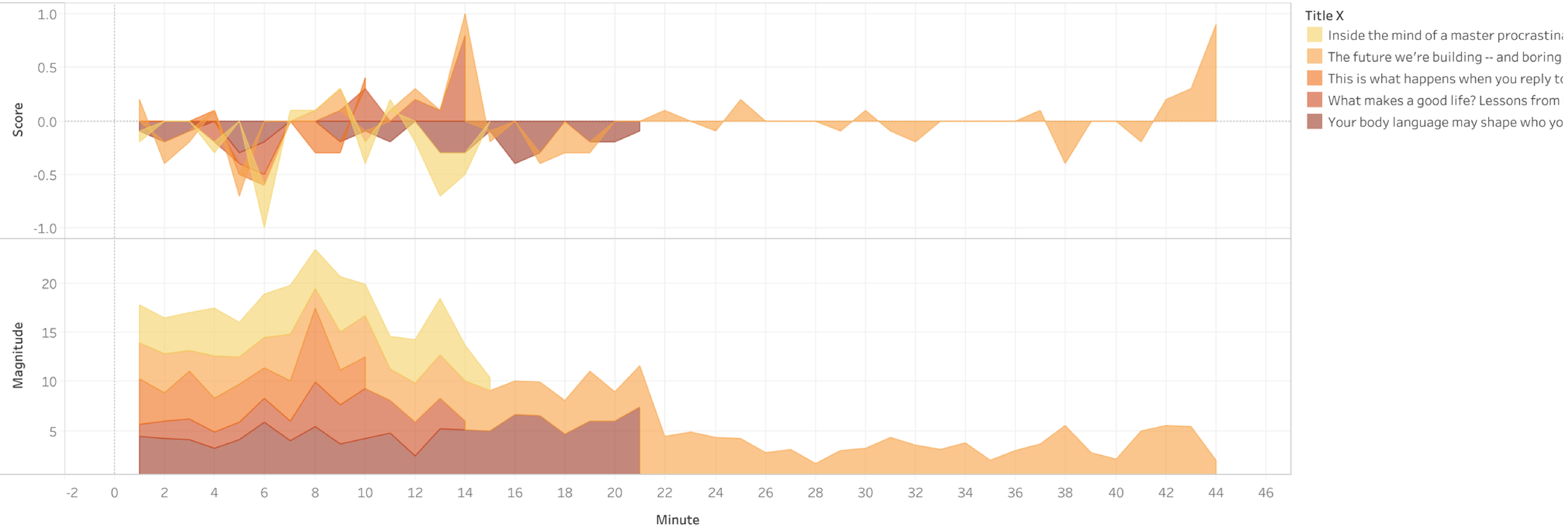


- Top viewed videos have higher magnitude in their transcripts
- Top viewed videos have lower magnitude in their descriptions
- Top viewed videos have higher magnitude in their titles

Visualizations

Comparing normalized views sentiment and magnitude of the top 5 most and least viewed videos

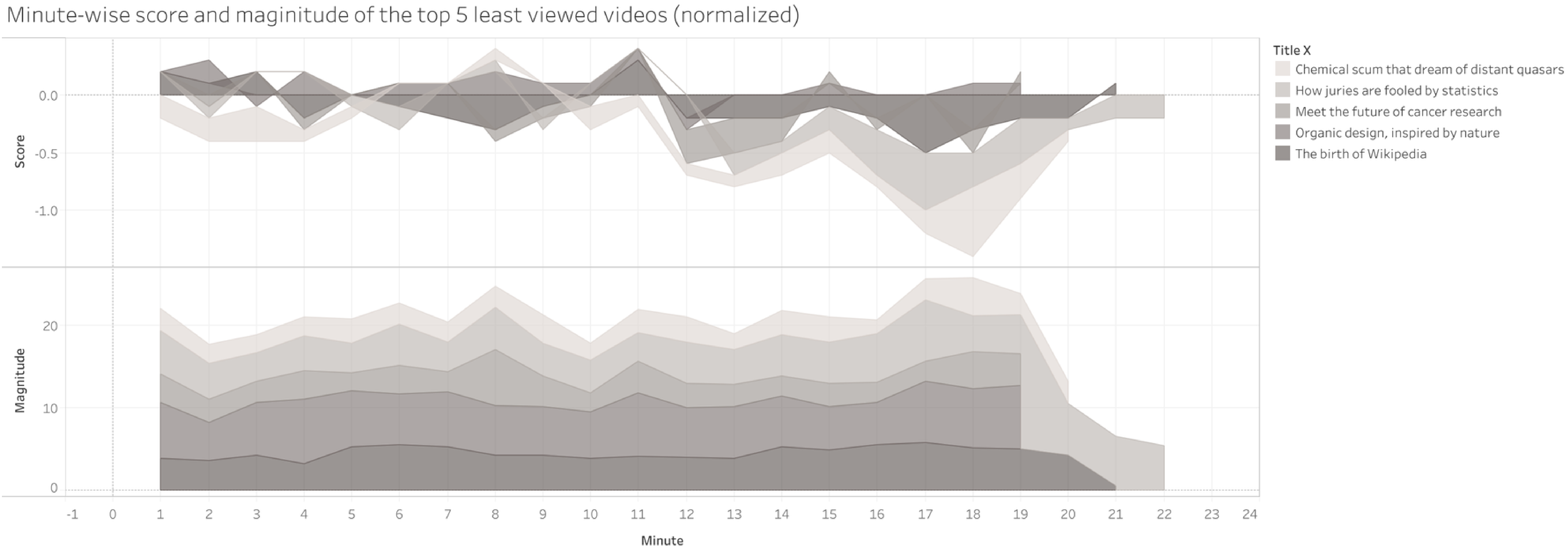
Minute-wise score and maginitude of the top 5 most viewed videos (normalized)



- Most talks have variations or emotions in their talks
- Intensity of the talks gradually peak and then come down

Visualizations

Comparing normalized views sentiment and magnitude of the top 5 most and least viewed videos



- Most talks don't have varying emotions in their talks
- Top viewed videos have higher magnitude in their content
- More unpredictable the talk the, the better it performs

- Sentiment Analysis
  - Sentiment scores and magnitude of titles, descriptions and transcripts
  - Analyze text per min sentiments
- Visualizations
  - Top 10 and least 10 viewed videos
  - All sentiment scores and top 10 views, bottom 10 views.
  - All magnitude and views - top 10 views, bottom 10 views.
  - Sentiments per minute - top 5 and bottom 5 viewed videos
- **Other text analysis**
  - Views received by every tag
  - Word clouds of transcripts - top 20 and bottom 20 views
- Webscraping
  - Authors nationality details from Wikipedia
- Correlations & Predictions
  - Does more "laughter" in audience reaction mean more views?
  - Funniness (laugh count / mins), and length > can we predict the number of views?

Other text analysis

- Question - Which tags have received the most views?
- Cleaning
- Normalized the views by dividing views by the length of the tags
- Added all the normalized views with each tag

speaker_occupation	tags
Author/educator	['children', 'creativity', 'culture', 'dance', 'education', 'parenting', 'teaching']
Climate advocate	['alternative energy', 'cars', 'climate change', 'culture', 'environment', 'global issues', 'science', 'sustainability', 'techn
Technology columnist	['computers', 'entertainment', 'interface design', 'media', 'music', 'performance', 'simplicity', 'software', 'technology']
Activist for environmental justice	['MacArthur grant', 'activism', 'business', 'cities', 'environment', 'green', 'inequality', 'politics', 'pollution']
Global health expert; data visionary	['Africa', 'Asia', 'Google', 'demo', 'economics', 'global development', 'global issues', 'health', 'math', 'statistics', 'visu
Life coach; expert in leadership psychology	['business', 'culture', 'entertainment', 'goal-setting', 'motivation', 'potential', 'psychology']
Actor, comedian, playwright	['Christianity', 'God', 'atheism', 'comedy', 'culture', 'humor', 'performance', 'religion', 'storytelling']
Architect	['architecture', 'collaboration', 'culture', 'design', 'library']
Philosopher, cognitive scientist	['God', 'TED Brain Trust', 'atheism', 'brain', 'cognitive science', 'consciousness', 'evolution', 'philosophy', 'religion']
Pastor, author	['Christianity', 'God', 'culture', 'happiness', 'leadership', 'motivation', 'philanthropy', 'religion']

Other text analysis

- Culture, ted-x, brain, art, science, techology, global issues have the most views
- Tags like blockchain, conservation, nuclear energy, blindness, bio-ethics have the least views
- Use of action words like make, want, say, look, etc.

Tags analysis

Bottom 20 views

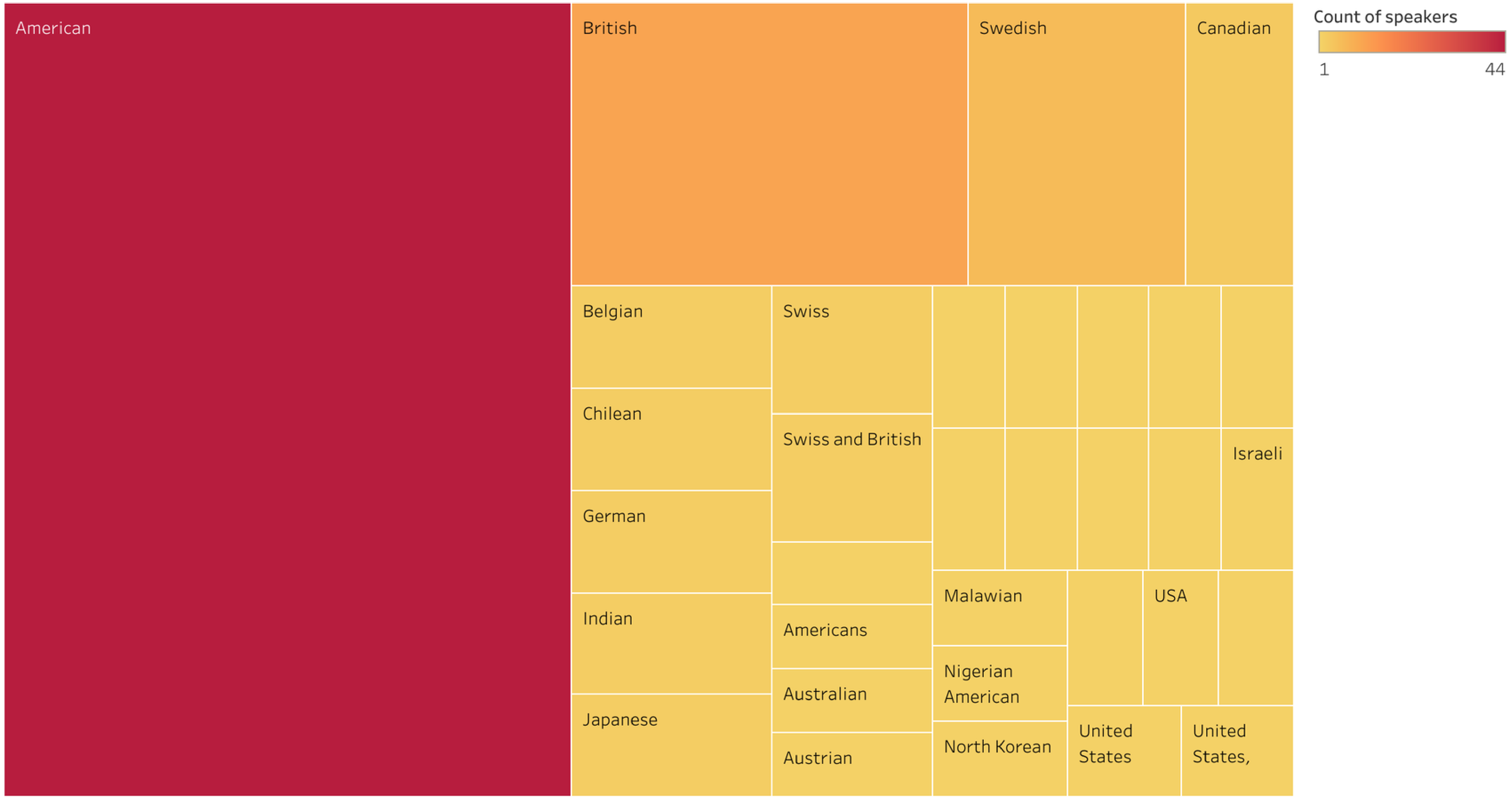
Top 20 views

- **Sentiment Analysis**
  - Sentiment scores and magnitude of titles, descriptions and transcripts
  - Analyze text per min sentiments
- **Visualizations**
  - Top 10 and least 10 viewed videos
  - All sentiment scores and top 10 views, bottom 10 views.
  - All magnitude and views - top 10 views, bottom 10 views.
  - Sentiments per minute - top 5 and bottom 5 viewed videos
- **Other text analysis**
  - Views received by every tag
  - Word clouds of transcripts - top 20 and bottom 20 views
- **Webscraping**
  - Author nationality details from Wikipedia
- **Correlations & Predictions**
  - Does more "laughter" in audience reaction mean more views?
  - Funniness (laugh count / mins), and length > can we predict the number of views?

Web Scraping

- Used Wikipedia API to extract speaker nationality of the 100 top most viewed videos
- Got a list of the speakers with their nationality
- Insight - Majority of the speakers are from the US, UK, Sweden and Canada

Nationality of the authors of the top 100 most viewed talks and with a Wikipedia profile



- Sentiment Analysis
  - Sentiment scores and magnitude of titles, descriptions and transcripts
  - Analyze text per min sentiments
- Visualizations
  - Top 10 and least 10 viewed videos
  - All sentiment scores and top 10 views, bottom 10 views.
  - All magnitude and views - top 10 views, bottom 10 views.
  - Sentiments per minute - top 5 and bottom 5 viewed videos
- Other text analysis
  - Views received by every tag
  - Word clouds of transcripts - top 20 and bottom 20 views
- Webscraping
  - Author nationality details from Wikipedia
- Correlations & Predictions
  - Does more "laughter" in audience reaction mean more views?
  - Funniness (laugh count / mins), and length > can we predict the number of views?

Correlation

- Is there a correlation between funniness and views?

```
0.2013428742460559
0.04815203089194869
      comments                                       description  duration                    event  ...  nfunny_count   age        nviews   ncomments
317         76  While we all agree that compassion is a great ...       946  TEDSalon 2009 Compassion  ...           0.0  9.18  1.919880e+04    8.278867
286         12  Keith Bellows gleefully outlines the engineeri...       966                  TED2002  ...           0.4  9.30  1.992204e+04    1.290323
192         43  Joseph Lekuton, a member of parliament in Keny...       326           TEDGlobal 2007  ...           0.2  9.86  2.035761e+04    4.361055
310         29  In 1998, aircraft designer Paul MacCready look...      1368                  TED1998  ...           0.3  9.20  2.142815e+04    3.152174
319         30  Join Rev. James Forbes at the dinner table of ...      1118    Chautauqua Institution  ...           0.1  9.18  2.226688e+04    3.267974
...        ...                                                ...       ...                      ...  ...           ...   ...           ...         ...
2022       527  What keeps us happy and healthy as we go throu...       766        TEDxBeaconStreet  ...           0.2  2.03  8.178289e+06  259.605911
2074       310  Tim Urban knows that procrastination doesn't m...       843                  TED2016  ...           1.9  1.80  8.191892e+06  172.222222
1265      2290  Body language affects how others see us, but i...      1262           TEDGlobal 2012  ...           0.4  5.25  8.220077e+06  436.190476
2352       250  Elon Musk discusses his new project digging tu...      2450                  TED2017  ...           0.6  0.67  8.456773e+06  373.134328
2027       150  Suspicious emails: unclaimed insurance bonds, ...       588         TEDGlobal>Geneva  ...           4.0  1.98  1.034140e+07   75.757576

[2461 rows x 25 columns]
```

- Measuring funniness - counting the reaction "laughter" from the transcript
- Normalized the outcome by dividing laughter count by duration of the video to get laughs per min
- Ran a correlation of laughs per views and normalized views
- Low correlation score - 0.04

Prediction

- Can we predict video views if we know the length of a video and number of laughs per minute?
- Used sklearn
- Coefficients: 0, 583 with an R squared value of -0.06

Thank you.