TED Talks Analysis

Antti Hemilä
Erik Husgafvel
Maria Pandele

Contents

1. Introduction

1.1 Main Goal

The main goal of our project is to provide users with meaningful tools to be able to discover insights about TED Talks series, to understand how talks are perceived by the public and how they have been evolving throughout the years. We want users to be able to explore this vast set of data on different levels starting from general features, such as where and when these talks are being given, and ending to deeper analysis like how different characteristics correlate and which talks form biggest clusters by being similar with their subjects.

1.2 Data

We decided to analyze a data set from Kaggle. The data set consists of 2550 TED-talks from the official TED.com website, and contains the data until September 21st, 2017.

1.2.1 Input data

Our input data is spread in 3 files:

- **ted_main.csv** - information about each talk from Kaggle
- **transcripts.csv** - transcript of the talk with audience reactions (Applause, Applause continues, Music, Music ends, Laughter) from Kaggle
- **cities_coordinates.csv** - GPS locations about the cities where the talks happened. First, we looked for both country and city of each event in our data set by running a scraper on the TED website. However, in some cases we had to manually fill in the data if it could not be gathered from the website. Second, we gathered the latitude and longitude coordinates of these locations. This was done by using MapQuest API on GPS Visualizer online tool.

1.2.2 Output data

For each visualization, we first processed the data and then loaded it on the website. Each data frame was tailored to the needs of each graph. Thus, in the end we had data frames oriented towards speakers, events, events locations, correlations and relations between the talks. The latter was computed by using a cosine similarity between transcripts.

2. General path and process

2.1 Milestone 1

We started the project by getting together, setting the overall individual goals, and started searching for a possible dataset. Our search culminated with three options: Wildfires in Australia, avocado usage in USA and TED Talks. We first decided to animate wildfires in Australia, but after a couple of hours we discarded the idea as we could not find any supporting data, nor a number of species, nor historical data that would have been easily accessible. After balancing a while between avocado data set and TED Talks, we chose to proceed with TED Talks as it seemed to have more gripping surface for visualization.

Since the dataset was published on Kaggle, we were able to explore all the notebooks on the website and understand how others had been analyzing this data. Researching more about TED talks, we noticed that these works were mostly basic data analysis and lacked meaningful and interactive visualizations. To the best of our knowledge, only one of the earlier projects had done a data visualization, but this one lacked depth in our opinion. Therefore we saw an opportunity to make a significant contribution.

In the first Milestone, we did an exploratory analysis, in which we assessed the quality of the dataset and the links between the variables. We also started planning possible visualizations for the webpage.

Our initial plan was to provide insights on what makes a good and highly performing talk. However, the dataset was rather descriptive meta-data (title, speaker, duration, transcript, ...) about each talk, and it lacked meaningful information on how people perceive them. For example we had quantitative ratings given by people to these talks but did not know how these ratings were given.

2.2 Milestone 2

In the second milestone we continued from milestone 1 by planning the visualizations, and after the main contents were planned, we started by implementing the functional prototype with the following visualizations: a parallel coordinate graph to illustrate general relationships between variables, force directed graph to show how the talks are clustered and bubble graph for ratings. Our website was built on a W3 template, and the page was just a long landing page.

At this point, we also planned the implementation of a speakers graph with gender and

occupation information, and an animation of events on a map and a time series. To have gender and occupation information, we planned web scraping as well as manual searching and clustering. Soon, we understood that clustering occupation data would require a lot of manual work, so we discarded the idea. However, as the idea of gender data was lucrative, so we incorporated it to our plans about the graphs.

Since our data had a lot of dimensions we could have expanded on, we wanted to structure it in a narrative way that even users unfamiliar with TED Talks would be able to understand: starting from a general view, and chart by chart providing a more complex view on the talks. Thus we decided that firstly, we should give an overview by presenting the TED events, how they developed over time and space. Secondly, we wanted to dive deeper in who the speakers are. Lastly, we decided to present the correlations between the talks as well as relationships in a network.

2.3 Milestone 3

At the beginning, we put a lot of effort to finish our main visualizations. However, soon we decided that the bubble chart about the ratings is not feasible due to the low number of ratings and lack of development options. This led us the idea to present the TED speakers in a bubble chart instead.

Quite soon to the milestone 3 we realized that the W3 template does not suit our needs, so we divided the visualizations to multiple pages and generally enhanced HTML and CSS to our particular needs. At the same time, we divided the main Javascript to individual files, one for each visualization.

In the end, we implemented most of the ideas we mentioned in milestone 2. However, we decided not to synchronize the event-timeseries and the events map and their common animation, since we found it clearer to keep the two visualizations on their separate pages in order not to overwhelm the user.

During milestone 3 we used python package nltk to perform sentiment analysis for the transcriptions. However, the used training sets were movie reviews and twitter messages, so they did not perform very well in providing a continuous degree of positivity/negativity for a long transcription. Rather, for most of the talks the sentiment analysis indicated 100% confidence.

At about two weeks into milestone 3, we had a meeting about the future of the visualizations and color. For a long time we had talked about web scraping gender information for the speakers, but at that time we decided to focus our resources on other parts of the website, as gender information would only have been a nice additional feature.

After two weeks, many of the final visualizations were working, after which we followed iterative development: we provided each other feedback how to further improve the graphs and the overall layout.

By accident, a couple of days before the deadline we found that in milestone 1 we had missed 10 rows that had earlier published date than film date. So we adjusted the data accordingly.

3. Visualizations and implementation

3.1 Overall implementation

The project is divided in several pages. Each visualization has its own HTML and Javascript file and each page uses CSS flex layout. For navigation on the website, we decided that just by clicking a "Next" button it is not enough so we added a topbar where you can directly go to each visualization, since one of our project's main goal is to allow the user to explore the dataset.

As a colorscheme, we decided to stick with the **red**-**black** theme of TED to which we added **white** and shades of **purple**.

3.2 Event timeseries

The graph per se is very simple. The bubble's size was chosen to represent the number of talks, as the number of talks indeed is the variable that defines how large a particular event has been.

The largest work was related to coloring of the different types of events: separating from the webpage common color theme was necessary as nine colors were needed. Further, nearby groups had to have different colors and preferably colors that are related to the events themselves.

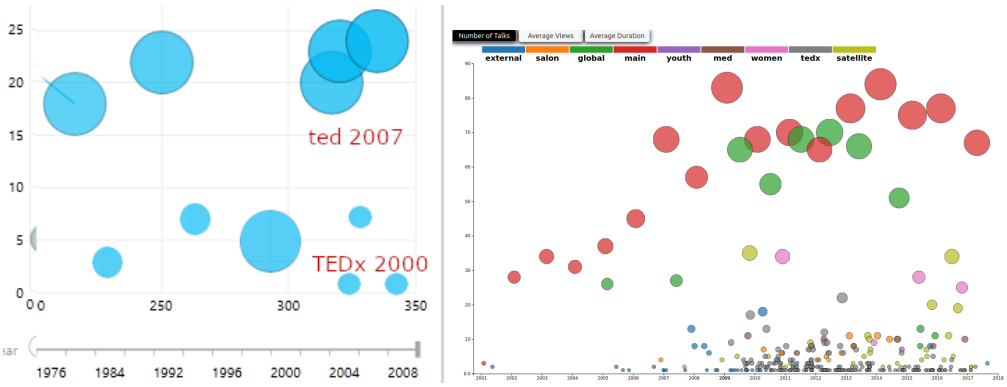

Figure 3.1: From sketch to final timeseries

3.3 Events map

Initially, we wanted to display the events by using the same colors as in the timeseries graph, by type of event, but realized that would just clutter the visualization. Moreover, we found out that in one location there could be different types of events. So we decided to keep it simple and use 2 constrasting colors from our colorscheme. To draw the map we downloaded a topology json file from World-Atlas, from which we cut out Antarctica. Like we mentioned earlier, the map is not animated together with the timeseries.

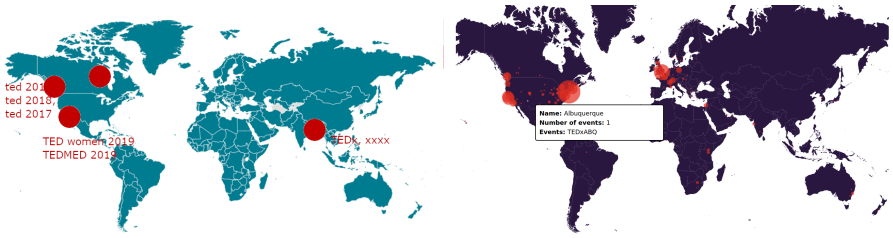

Figure 3.2: From sketch to final map

3.4 Speakers

The initial idea of the graph was to present information about how people had been rating TED Talks. According to that plan, the graph would have first counted how many times each different rating has been given over all videos in the data set, and then formed bubbles with size relative to the count value. That plan was initially executed and the end result is visible in the provided pictures below in the figure 3.3. The first graph on the left presents the end result of the plan.

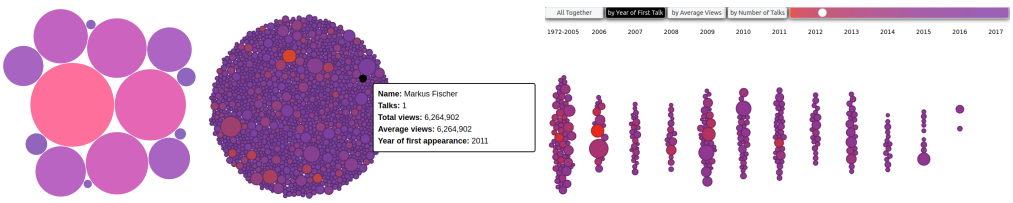

Figure 3.3: From bubble graph of ratings to speakers graph

However, during the process we started to understand that the ratings data was not the most convenient set of data to be presented in our project. First of all, it did not suit well to our story. After examining locations of the events and the development of different event types, the connection to examine data of how viewers had been rating the videos felt particularly vague. In addition, the ratings data did not suit well to be presented as a bubble graph. Our vision of the graph contained thoughts about being able to divide the graph in different sections based on different variables and thus discover new information about the data set. We realized that the ratings data was not a suitable set of data for our purposes.

After presenting different event types of TED Talks, we felt much more convenient to present the data of speakers. Examining the speakers after learning about different events felt like a natural continuation to our story. In addition, speakers data set fit better to our thoughts about being able to examine the data based on a single variable. Now it was possible to divide the bubbles on the graph based on certain condition, thus helping to discover new information about the speakers in an interesting manner. We also visioned about a possibility to learn about interesting speakers with a circular dendogram. The idea was that the user could dive into a

bubble by double clicking it and thus open a new view with a dendogram telling more in-depth information about the chosen speaker. Unfortunately, we could not achieve this vision in the given time frame.

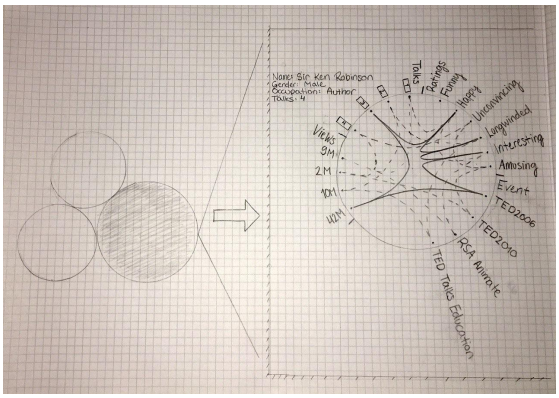

Figure 3.4: Visioning how the bubble graph would've turned into circular dendogram

At this point it became clear that the previous solution of drawing a bubble graph using d3.hierarchy and d3.pack -functions was not sufficient anymore, as we wanted the graph to react on user events and to move the same bubbles on the graph to new locations with a smooth transition. Simple bubble graph started to turn into force directed bubble graph. Now it was possible to follow certain interesting bubble move to new locations on the graph based on whether the user wanted to examine the division of bubbles by year of first appearance of the speaker, by average views count or by number of held talks.

With the colour of the bubbles we were able to highlight those speakers, who had been most successful by the number of talks they had held so far. With the size of the bubble we chose to present the total count of views for each speaker, as it brought an interesting insight out of the graph that not necessarily the speakers with the largest number of talks had the biggest number of viewing times. As people tend to underestimate the relative differences between the sizes of bubbles, the area is used to display the total number of views: total number of views has a logarithmic tendency which makes it suitable to map it to size. Finally, we decided to add a range selector for user to be able to filter out noise from the graph made by the speakers that had hold only one speech, thus making it possible for the user to concentrate on the speakers that were more valuable in terms of number of held talks and number of viewing times.

3.5 Correlations

We planned this chart to be a tool for the user to explore differences between the dimensions and find interesting talks. Thus the main challenge was the vast amount of data and dimensions. First, we took all possible dimensions after which we dropped the worst ones out one-by-one until only dimensions that had relevant interactions were left. Second, the large number of talks was still problematic as the lines overlapped. This was solved by diminishing the thickness as well as the opacity of the line depending on the total number of lines that are being displayed. Brushes for each dimension were also added to support subsetting the data interactively.

After the data was filtered, the challenge was to choose a color scheme. Milestone 2 sketch had a unique color scheme that was intuitive to understand based on traffic lights. However, during milestone 3, in order to have a coherent coloring with all of the graphs, background color was turned white and the color scale was changed to black-purple-red. As the purple and red were very close to one another, the main way to ensure clarity was to create a spectrum saturation

and brightness, though even this had problematic as the used "TED-red" is already quite dark. For coloring the event type variable, same scheme was used as with timeseries.

Finishing touch to the graph was given by changing the color of the path that was hovered and by including a clicking event that would open the TED page of the talk for the user. Lastly, a search bar was included to give user an option to check the favorite talk statistics in 2017 terms.

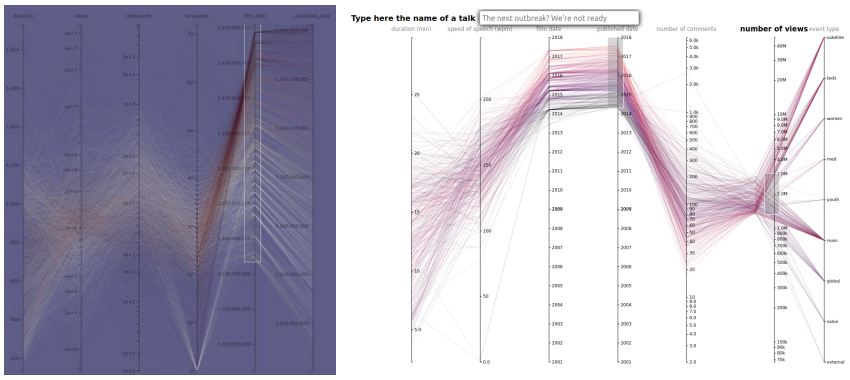

Figure 3.5: From sketch to final correlations

3.6 Related talks network

The network of talks was first built with a force directed layout by using the column "related talks" ("watch next" in the TED website). The network looked like one connected component - a giant hairball. And without knowledge about the meaning of the edges, rendering it hard to tweak. So we decided that it's better to compute our own edges.

We experimented with two metrics: tags (which also appear on the website) as well as cosine similarity between transcripts. After visualizing the two of them with the networkx library in python and looking at the number of edges and connected components we decided that the cosine similarity performed best. To make the skeleton for the visualization we used the Gephi software to generate an SVG. We used the OpenOrd layout since its aim is to better distinguish clusters and then run an algorithm to solve overlapping nodes. We then animated the exported SVG to better highlight a specific talk and added a search bar.

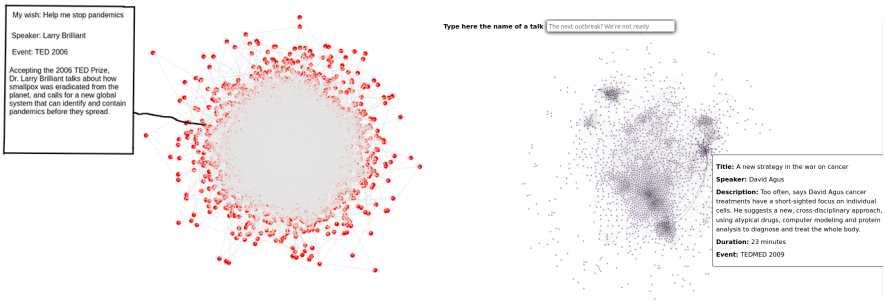

Figure 3.6: From network sketch to final

4. Peer assessment

4.1 Common work

We all worked on the overall design and colors for our website. We decided together what data processing we needed, and we wrote each milestone and the processbook together.

In developing the website, we divided tasks amongst ourselves but we also helped each other with ideas and technical problems we encountered.

4.2 Antti Hemilä

- Events timeseries
- Parallel coordinates
- Implementing main layout and navigation
- Manually filled in locations for ted talks

4.3 Erik Husgafvel

- Speakers graph
- Screencasting
- Tooltip

4.4 Maria Pandele

- Exploratory data-analysis
- Majority of data processing and scraping webpage
- Talk network
- Events map