CS 6630: Project Process Book

TED talks topic trend visualization

Hsuan Lee and Chien-Wei Sun

School of Computing, University of Utah

December 2, 2017

Contents

1 Overview **3**

1.1 Overview and Motivation 3

1.2 Related Work 3

1.3 Questions 4

2 Data **5**

2.1 Dataset 5

2.2 Exploratory Data Analysis 6

3 Design Evolution **9**

3.1 Prototype 9

3.2 Evolution 9

3.2.1 Network Chart 9

3.2.2 Word Cloud Chart 13

3.2.3 Line Chart 17

3.2.4 Button of Line Chart 17

3.2.5 Video List Table 19

3.2.6 Radar Chart 19

4 Implementation **21**

4.1 Overview 21

4.2 Interaction 22

4.2.1 Network Chart 22

4.2.2 Word Cloud 22

4.2.3 Line Chart 22

4.2.4 Buttons 22

4.2.5 Video List Table 22
4.3 Flow-Chart 22

5 Evaluation **23**
5.1 Solution in our question 23

Bibliography **24**

Chapter 1

Overview

1.1 Overview and Motivation

TED is a leading organization which provides influential and understandable talk to the world. These talks cover a lot of fields, from anthropology to machine learning, and also from biology to sociology. We are interested in the relationship between technology and world market, and we want to know if TED somehow shows the trend of popular technology or it provides a platform for topics which do not get much attention in the world.

The relevance of different categories is also what we want to discover. For example, several years ago, it was popular that researchers tried to innovate theory according to the behavior of insects, like ants and bees. There are many theories developped based on the cooperation pattern of those animals. In the past, people did not consider that there is a strong relevence between insects and learning theory. We also wonder if we can find situation which is similar to the example.

1.2 Related Work

When we are searching useful data for this project, we found the TED talks dataset and also a visulization by Sean Miller[**?**]. In this visualization, it shows statistics of the dataset and also allow user to search video by one tag. However, this visualization does not answer the questions we mention in the above. That's why we decide to build our own visualization of the dataset.

1.3 Questions

Here are questions we expect to answer at the end of this project:

- What are the trend of category tags appeared on TED talks?
- Is there any relationship between the TED talks and the big events happened in the world?
- Is there a strong relevence between two topics that in general people will not think they are related?
- Can we learn the trend of research on a specific field by analyzing the popularity of keywords? Or it shows the topics which people do not put attention on for now but will become important in the future?

Chapter 2

Data

2.1 Dataset

We find the dataset from Dataset Distribution Portal[**?**]. This dataset include the video recording from the TED website from 1972 to 2017. For each video, its data contains the following attributes:

id	speaker	URL	URL	description	transcript_URL	montl	year_film	event	duration	date_publis	tags
1	Al Gore	Averting the climate crisis	http://www.ted.co	With the same humor	http://www.ted.c	2	2006	TED2006	0:16:17	6/27/06	cars,alternative energy,culture,politics,science,climate change,environment,su
2	Amy Smith	Simple designs to save a life	http://www.ted.co	Fumes from indoor co	http://www.ted.c	2	2006	TED2006	0:15:06	8/15/06	MacArthur grant,simplicity,industrial design,alternative energy,invention,engine
3	Ashraf Ghani	How to rebuild a broken stat	http://www.ted.co	Ashraf Ghani's passio	http://www.ted.c	7	2005	TEDGlobal 2	0:18:45	10/18/06	corruption,poverty,economics,investment,military,culture,politics,policy,global
4	Burt Rutan	The real future of space expl	http://www.ted.co	In this passionate talk	http://www.ted.c	2	2006	TED2006	0:19:37	10/25/06	aircraft,flight,industrial design,NASA,rocket science,invention,engineering,entr
5	Chris Bangle	Great cars are great art	http://www.ted.co	American designer Ch	http://www.ted.c	2	2002	TED2002	0:20:04	2004/05/07	cars,industrial design,transportation,invention,design,technology,business,art
6	Craig Venter	Sampling the ocean's DNA	http://www.ted.co	Genomics pioneer Cra	http://www.ted.c	7	2005	TEDGlobal 2	0:16:51	2004/05/07	biotech,invention,oceans,genetics,DNA,biology,science,entrepreneur,biodivers
7	David Pogue	Simplicity sells	http://www.ted.co	New York Times colur	http://www.ted.c	2	2006	TED2006	0:21:26	6/27/06	simplicity,computers,software,interface design,music,media,entertainment,per
8	David Rockwell	A memorial at Ground Zero	http://www.ted.co	In this emotionally cha	http://www.ted.c	2	2002	TED2002	0:24:37	2006/12/07	New York,memory,interview,death,culture,architecture,disaster relief,cities,urba
9	Dean Kamen	To invent is to give	http://www.ted.co	Inventor Dean Kamen	http://www.ted.c	2	2002	TED2002	0:20:07	2004/05/07	robots,cars,industrial design,transportation,invention,education,innovation,soc
10	Dean Ornish	The killer American diet that'	http://www.ted.co	Forget the latest disea	http://www.ted.c	2	2006	TED2006	0:03:18	12/14/06	obesity,disease,health,health care,culture,food,science,global issues
11	Jane Goodall	What separates us from chin	http://www.ted.co	Jane Goodall hasn't fo	http://www.ted.c	2	2003	TED2002	0:27:25	2004/05/07	primates,Africa,culture,science,environment,animals,nature,global issues
12	Eva Vertes	Meet the future of cancer res	http://www.ted.co	Eva Vertes -- only 19	http://www.ted.c	2	2005	TED2005	0:18:49	10/02/06	wunderkind,cancer,disease,health,science,technology
13	Frank Gehry	A master architect asks, Nov	http://www.ted.co	In a wildly entertaining	http://www.ted.c	2	2002	TED2002	0:22:00	1/17/08	invention,interview,culture,architecture,design,creativity,business
14	Golan Levin	Software (as) art	http://www.ted.co	Engineer and artist Go	http://www.ted.c	2	2004	TED2004	0:14:53	2004/05/07	invention,software,music,entertainment,performance,technology,art
16	Helen Fisher	Why we love, why we cheat	http://www.ted.co	Anthropologist Helen	http://www.ted.c	2	2006	TED2006	0:23:27	2009/06/06	love,gender,relationships,cognitive science,psychology,evolution,culture,scien
18	Janine Benyus	Biomimicry's surprising less	http://www.ted.co	In this inspiring talk ab	http://www.ted.c	2	2005	TED2005	0:23:19	2004/05/07	biomimicry,DNA,evolution,biology,fish,science,environment,animals,design,tec
19	Kevin Kelly	How technology evolves	http://www.ted.co	Tech enthusiast Kevin	http://www.ted.c	2	2005	TED2005	0:20:00	11/14/06	philosophy,evolution,culture,choice,history,science,future,technology
20	Malcolm Gladwel	Choice, happiness and spag	http://www.ted.co	"Tipping Point" autho	http://www.ted.c	2	2004	TED2004	0:17:30	9/19/06	consumerism,marketing,economics,culture,media,food,choice,storytelling,bus
21	Mena Trott	Meet the founder of the blog	http://www.ted.co	The founding mother	http://www.ted.c	2	2006	TED2006	0:16:46	8/25/06	software,culture,design,entertainment,storytelling,business,communication,co
22	Michael Shermer	Why people believe weird th	http://www.ted.co	Why do people see th	http://www.ted.c	2	2006	TED2006	0:13:25	11/08/06	faith,illusion,culture,religion,science,entertainment
23	Peter Gabriel	Fight injustice with raw videc	http://www.ted.co	Musician and activist	http://www.ted.c	2	2006	TED2006	0:14:08	12/06/06	TED Brain Trust,film,culture,music,activism,social change,storytelling,global iss
24	Pilobolus	A dance of "Symbiosis"	http://www.ted.co	Two Pilobolus dancers	http://www.ted.c	2	2005	TED2005	0:13:45	2002/09/07	dance,science and art,science,nature,entertainment,performance
25	Richard Baraniuk	The birth of the open-source	http://www.ted.co	In 2006, open-learning	http://www.ted.c	2	2006	TED2006	0:18:34	8/21/06	open-source,library,education,culture,global issues,technology,business,collab
26	Rives	If I controlled the Internet	http://www.ted.co	How many poets coul	http://www.ted.c	11	2006	TEDSalon 2	0:04:07	12/14/06	love,poetry,philosophy,culture,entertainment,performance
27	Ross Lovegrove	Organic design, inspired by	http://www.ted.co	Designer Ross Lovegr	http://www.ted.c	2	2005	TED2005	0:19:30	8/15/06	industrial design,invention,product design,science and art,DNA,biology,nature
28	Seth Godin	How to get your ideas to spr	http://www.ted.co	In a world of too many	http://www.ted.c	2	2003	TED2003	0:17:01	2004/05/07	TED Brain Trust,marketing,culture,choice,storytelling,business
29	Steven Levitt	The freakonomics of crack d	http://www.ted.co	"Freakonomics" autho	http://www.ted.c	2	2004	TED2004	0:21:15	9/19/06	narcotics race economics culture cities business

Figure 2.1.1: Data get from idiap.ch

id	month filmed
Speaker	year filmed
headline	event
URL	duration
description	date published
transcript URL	tags

To better understand the impact of TED videos, we develop web crawlers to collect attributes like **rates**(how do people feel after watching a video), **views**(how many time a video has been played), and some potentially valuable data like datetime, redirected urls, and transcripts. We use **Scrapy**[**?**] as our crawler. Figure 2.1.2 displays the rating options on TED website.

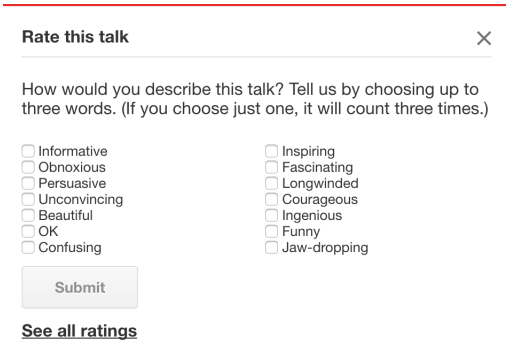

Figure 2.1.2: How people rate one video in TED website

Furthermore, in order to load data easily, we transfer our data from csv file to json form. We found this preprocessing can be accomplished painlessly by using **Pandas**[**?**] toolkit. Figure 2.1.3 shows what kind of data one video contains.

We plan to visualize the data according to the tags/keywords of the video. It is not efficient to search all the data to find which videos are related with one specific tag on javascript. For practical implementation, we will preprocess the dataset based on tags, which means to use tag as key to create input data.

2.2 Exploratory Data Analysis

In our design, the main chart user interact with is the network chart, which present the co-occurrence of tags. Hence, after we finish the job of collecting data, we move forward to build the co-occurrence matrix of tags. During this procedure, we observe that some tags

```
"id": 7,
"speaker": "David Pogue",
"headline": "Simplicity sells",
"URL": "http://www.ted.com/talks/view/id/7",
"description": "New York Times columnist David Pogue takes aim at technologyâ€™s worst
"transcript_URL": "http://www.ted.com/talks/view/id/7/transcript?language=en",
"month_filmed": 2,
"year_filmed": 2006,
"event": "TED2006",
"duration": "0:21:26",
"date_published": "6/27/06",
"tags": "simplicity,computers,software,interface design,music,media,entertainment,perf
"newURL": "https://www.ted.com/talks/david_pogue_says_simplicity_sells",
"date": "2006-06-27",
"views": "1646773",
"rates": [
  {
    "id": 7,
    "name": "Funny",
    "count": 968
  },
  {
    "id": 3,
    "name": "Courageous",
    "count": 46
  },
  {
    "id": 9,
    "name": "Ingenious",
    "count": 186
  },
  {
    "id": 1,
    "name": "Beautiful",
    "count": 60
  },
  {
    "id": 21,
    "name": "Unconvincing",
    "count": 104
  },
  {
    "id": 11,
    "name": "Longwinded",
    "count": 78
  },
```


Figure 2.1.3: Data of one video in JSON

appear in too many videos so that their existences are not meaningful to the matrix. These tags are ‘science’, ‘technology’, ‘global issue’. Since they show up in most of talks, we remove them from the matrix so that the network chart will look clear.

To create groups of tags, we apply k-means to divide them into 11 clusters, and one of them restore the outliers. Figure 2.2.1 are the results of two groups. One is the group whose center is tag ‘computers’, the other is the group whose center is ‘universe’. Color is used to distinguish the group in our design.

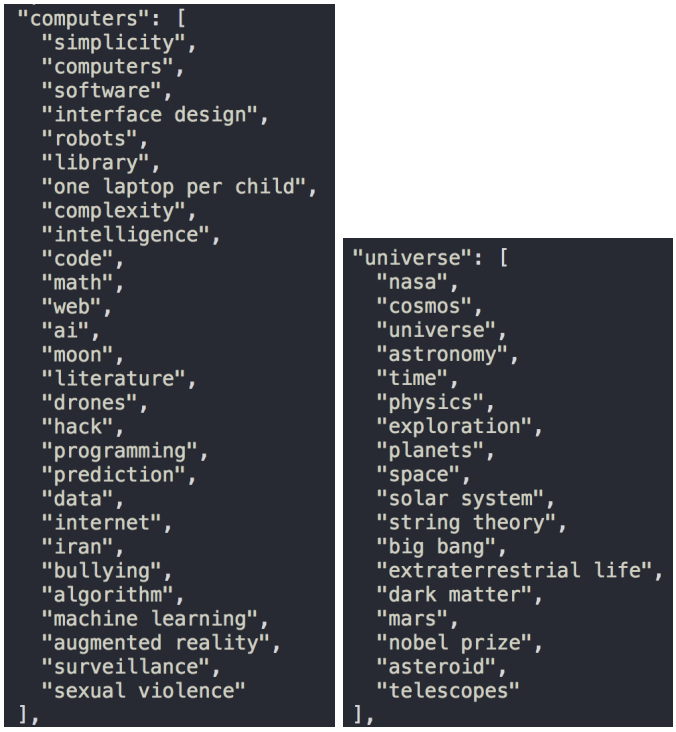

Figure 2.2.1: Clustering result

Chapter 3

Design Evolution

3.1 Prototype

Our design is based on the network layout, as shown in Figure 3.1.1. This network is composed of tags, and user can choose several tags they are interested in to through interaction with the network node. Next, the line chart in the middle of Figure 3.1.1 will display the tendency of chosen tags versus time/year. The last part help user to search for TED talks including these tags. User can decide the result is sorted by views or popularity.

We also want to compare the statistic of the tags between years, so we design a bar chart as shown in Figure 3.1.2. By making use of the sliding bar on the top, the statistics of two years is displayed. Figure 3.1.3 helps us to figure out what attributes are needed in each chart. It also shows the relationship of charts.

3.2 Evolution

3.2.1 Network Chart

First, we generate the network chart accoring the co-occurence matrix. Each node represent a tag, and the thickness of one link is decided by the co-occurence value between two tags. However, there are 403 tags and 19488 links on this chart, which make the network look crazy and take a lot of time to draw these lines, as shown in Figure 3.2.1.

We discuss how to fix this issue and propose two solution for that. One is to draw chord layout in the beginning. Chord layout help people understand the relationship between two groups. We can let user to click ribbon to then show the network layout of tags in these two

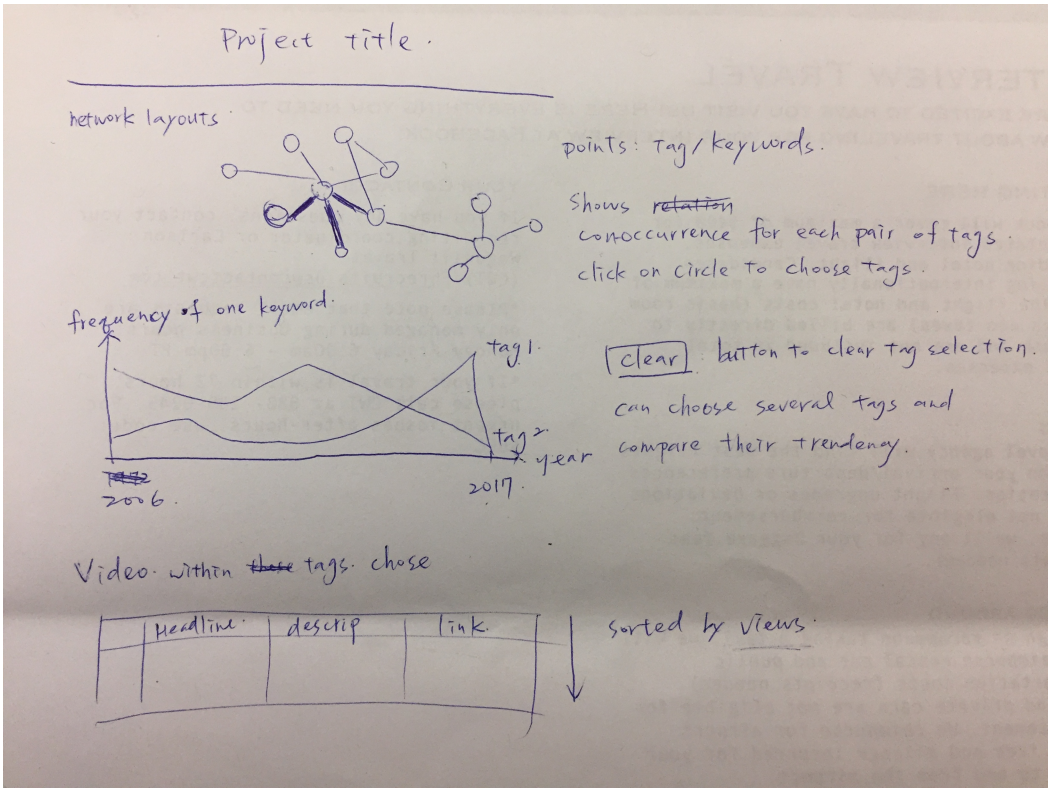

Figure 3.1.1: Design draft

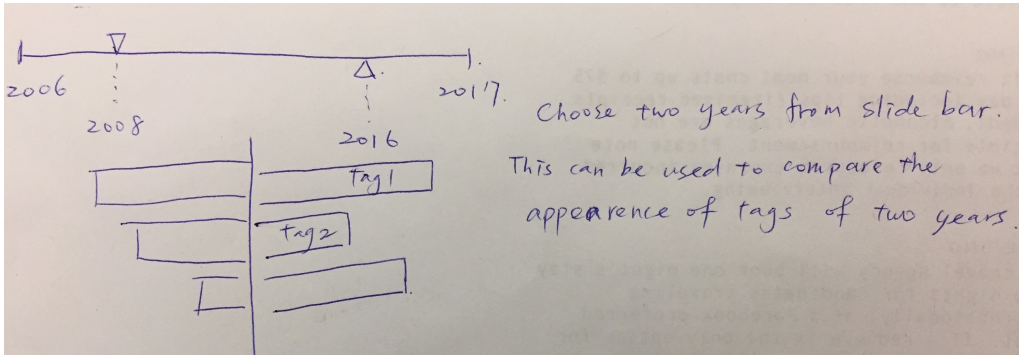

Figure 3.1.2: Design of Optional features

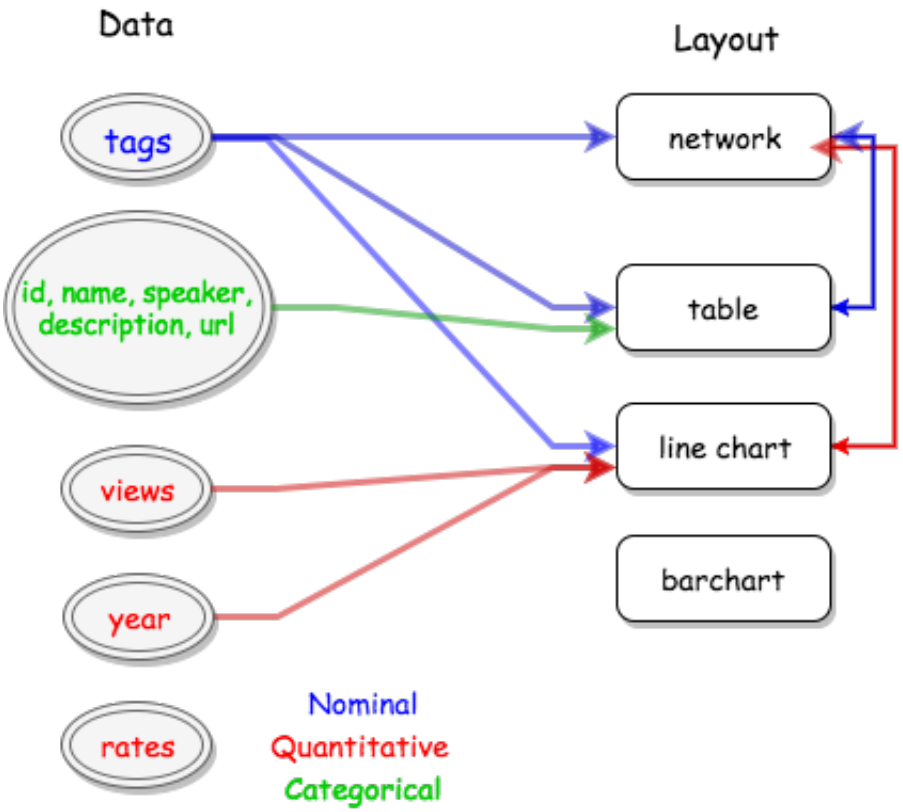

Figure 3.1.3: Category of data, and it relationship with the layout

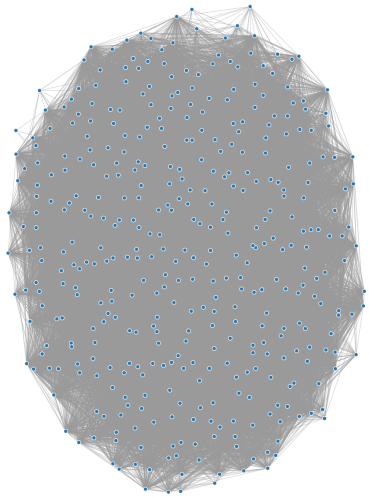

Figure 3.2.1: Network chart with Over 19,000 links

groups. However, this design does not allow us to observe all the related tags of one tag we choose. The other solution is to reduce the amount of links and nodes. We can provide an overview of network chart with nodes and links whose frequencies and value of co-occurence are bigger than threshold. Then, to zoom in on this chart, user can double-click on the tag they interested in to find all the other tag which is related to the chosen one. After applying the second method, our network chart looks better, as you can find in Figure 3.2.2.

Next, we want to observe the relationship between a chosen tag and other co-occurrence tags, so we implement a function that the network chart shows all the links whose edges include a specific tag after double-clicking on one the corresponding node. The image was not easy to understand and hard to find the most relative tag, as shown in Figure 3.2.3.

Figure 3.2.4 is the modified design to solve the above issue. This time, we grouped all the related tags by their categories. Instead of connecting the center with relative node, we connect it to an invisible group center and then link node with the center. This design help user to learn which category has strong co-occurrence within the specific, and also it is obvious to observe the most related tag. We call it **flower chart**, and this term is used to describe network chart within selected center in the following content.

To show the tag name of each node, we decide to apply tooltip instead of adding text to them. Also, to avoid the situation that the tooltip goes beyond the border, the direction of d3-tip is set to southwest. The information we provided in tooltip depend on whether it is flower chart or not. If network chart is not shown as flower chart, we will paste the tag name

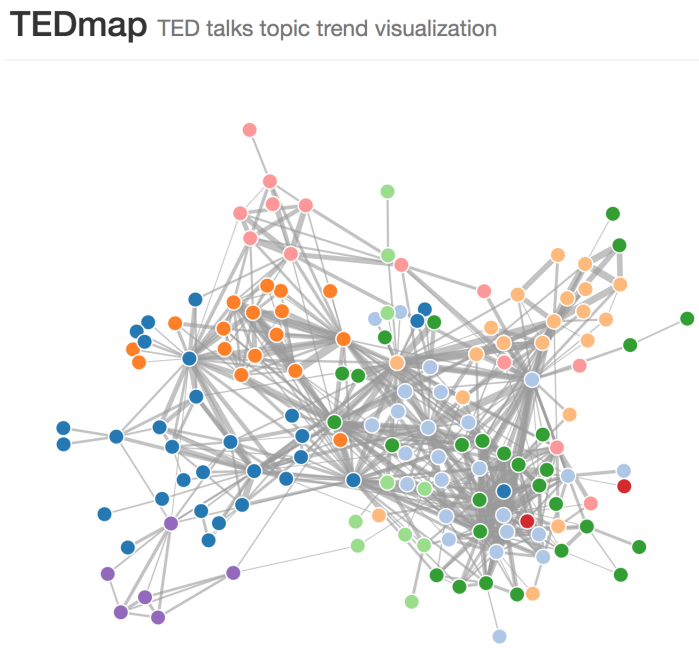

Figure 3.2.2: Network chart with link vale bigger than 15

and the top five strong related tags and co-occurrence with the node we hover on. Figure 3.2.5 is an example when the mouse hover on node which present 'biology'. If the chart becomes flower chart, then tooltip shows the tag name and co-occurence while we hover on that node, as you can see in Figure 3.2.6. Stroke width and color are changed to highlight which node we are watching.

Now this question comes to our mind: what if we want to know the co-occurrence between two tags in one year? To answer this this question, we add an button on the upper-right corner to let users choose which year they want to observe, as displayed in Figure 3.2.7. Example in Figure 3.2.8 demonstrate the comparation between 2003 and 2012.

3.2.2 Word Cloud Chart

After we draw the network chart and color each node according to category, we suddenly find that we did not explain the category and which tags are classified to. Therefore, we decide to add a word cloud chart on the top of page, which provides the information about the members each category includes, and the color of text follow the ordinal color scale we define in network chart.

In the beginning, the word cloud use a list of buttons for user to select the category, as

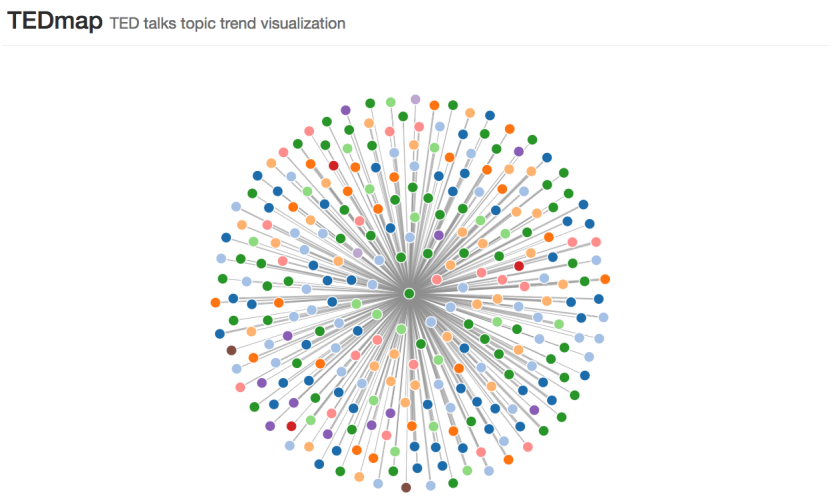

Figure 3.2.3: Zoom in for one tag

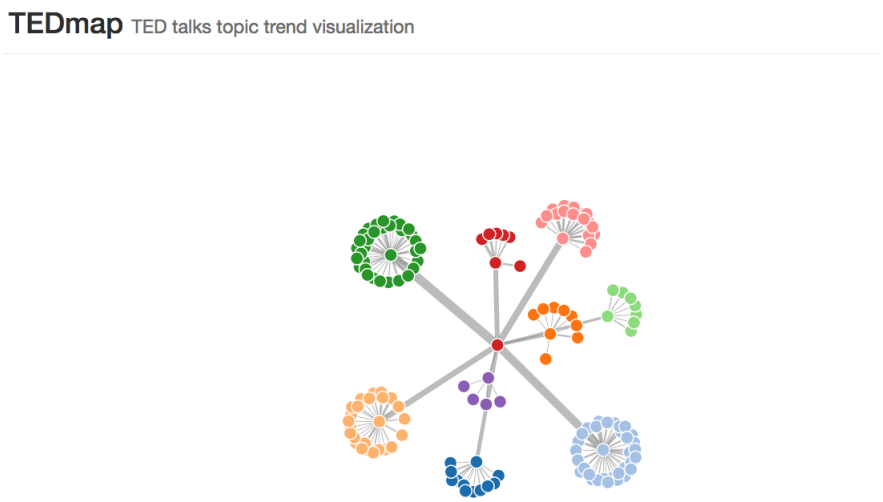

Figure 3.2.4: Zoom in for one tag with grouping

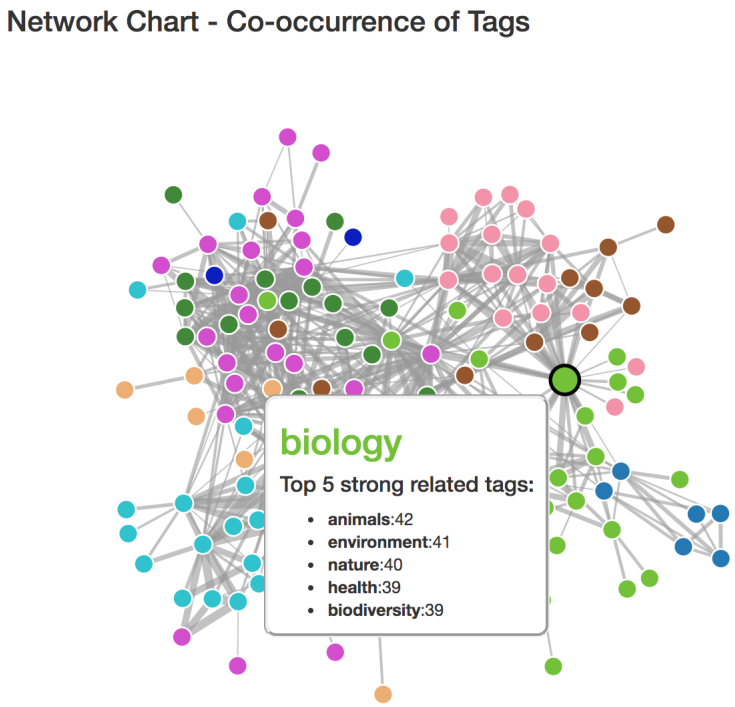

Figure 3.2.5: Tooltip design when no tag are focused on

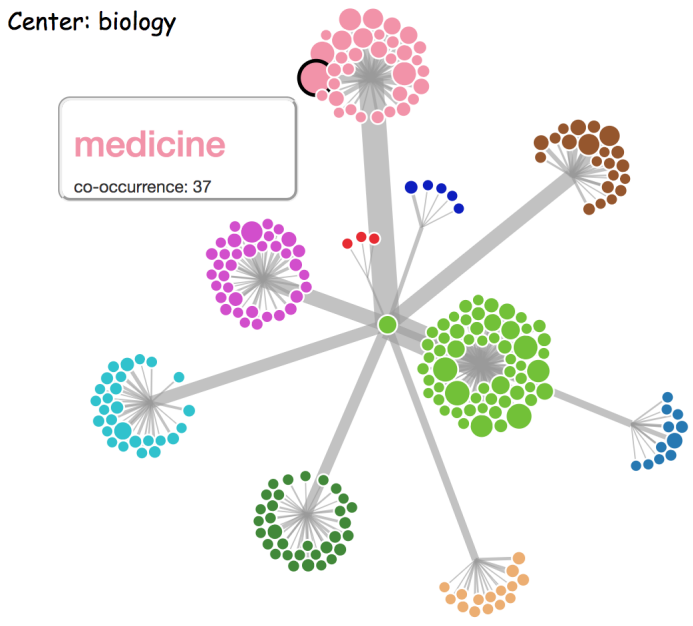

Figure 3.2.6: Tooltip design when we zoom in with center - biology

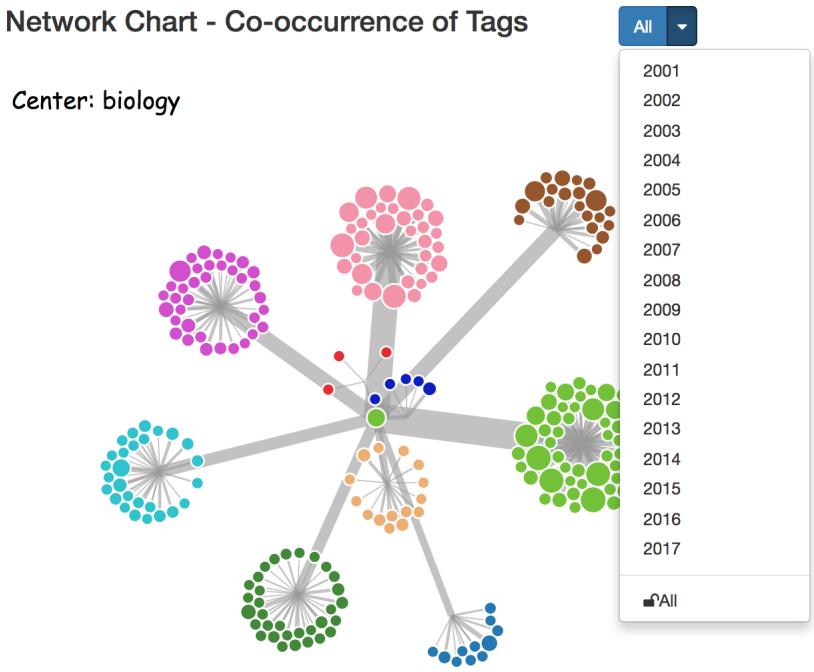

Figure 3.2.7: Dropdown list design for choosing year

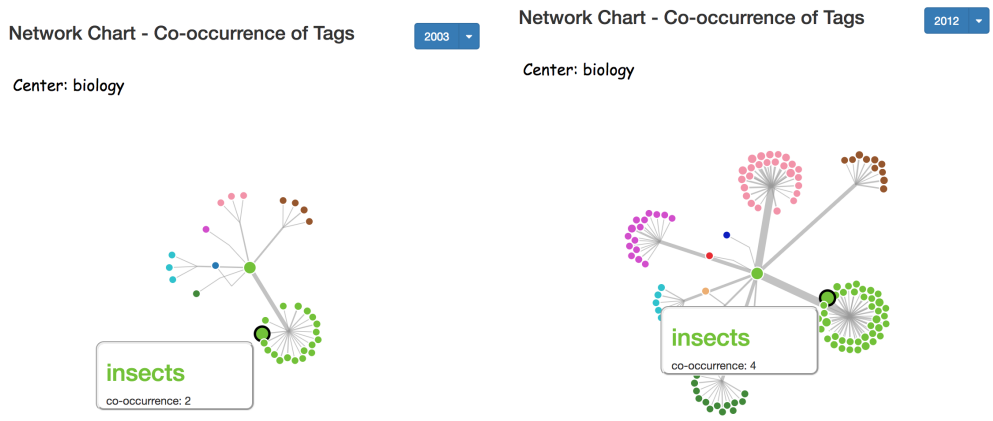

Figure 3.2.8: Comparation of flower chart in 2003 and 2012

Figure 3.2.9: Word Cloud button design, version 1

Figure 3.2.9

3.2.3 Line Chart

We need to present the video numbers of tags from 2012 to 2017 on our Line Chart design. We originally choose to use color as our channel to discern data, but soon we realize that colors are not enough for our hundreds of tags even if we use a gray scale on each hue. Therefore, we decide to add symbols[1] in the d3.js. on our Line Chart. Since symbols and the line on the Line Chart are both made by path element, it is more convenient for our implementation. The final design is show in Figure 3.2.12

3.2.4 Button of Line Chart

To better manage the interaction between these chart, we present Buttons Object. When users click on the Network or the Text Cloud, the clicked tag will be append into the set object behind the button. Once users click the button, we remove the tag in the set object. We easily and elegantly solve the problem of letting users have to many interface to interact with our components. We show our design topology in Figure 3.2.13

To increase connection of the Line Chart and the Table, we also need to let the Buttons have some hover event that can connected the Line Chart and the rows in the Table.

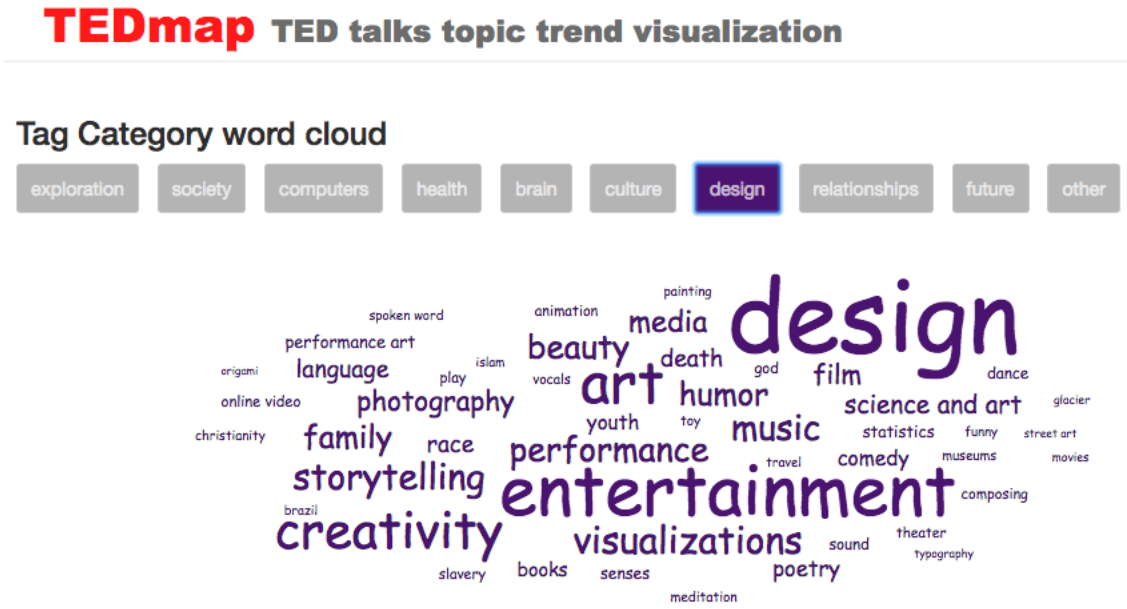

Figure 3.2.10: Word Cloud button design, version 2

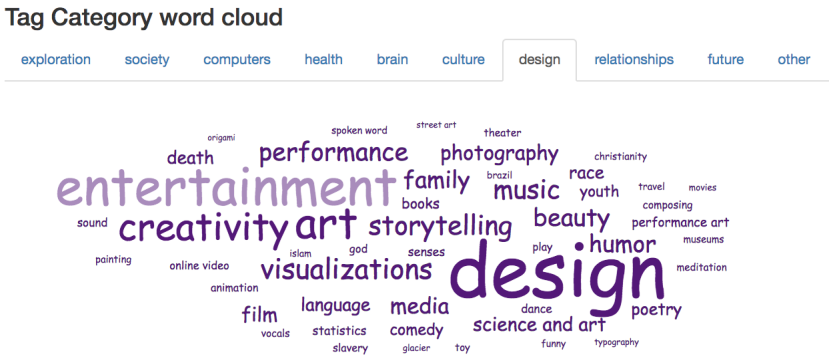

Figure 3.2.11: Word Cloud button design, version 3

Figure 3.2.12: Line Chart.

"topo".jpg

Figure 3.2.13: Topology of our design.

Figure 3.2.14: Our original Table Design.

Figure 3.2.15: Our new Table Design.

3.2.5 Video List Table

The requirement in the Table is to show the rest information that we haven't show in those components above. Table is the best way for our case.

We originally try to draw a svg table for the transition purpose. However, it need to much design and it is too hard to put all the information on the drawed svg table. Figure 3.2.14 and Figure 3.2.15 is our different two versions of design in Video List Table.

3.2.6 Radar Chart

After finish our must-have components, we decide to add an interesting Radar Chart on the tooltips of the Table. Comparing to the line chart on the official of TED(Figure 3.2.16), we believe that our Radar Chart is more likely to catch users eyes and let user understand the meaning of the rates of those videos.

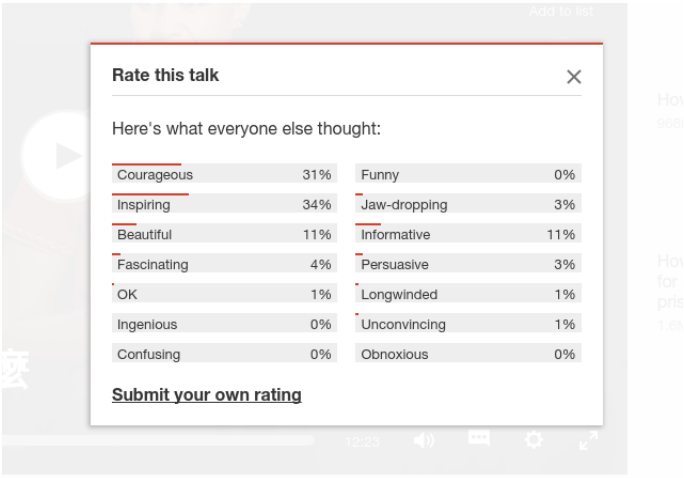

Figure 3.2.16: The Rate chart show on TED official website.

Figure 3.2.17: Radar Chart.

Chapter 4

Implementation

4.1 Overview

Figure 3.2.13 show the topology of our visualization web page. User can run through Word Cloud, Network Chart, Line Chart, Video List Table to either finding their next interested video, or understanding the trends of popular topic in TED in the recent decade. All of the components are interactive and the usage are easy to come up with.

When users connects to our website, they will intuitively see the Word Cloud and The Network Chart. In the Word Cloud, the texts are scattered and sized by their popularity. The bigger text, the more popular topics in TED video. When users cursor go over the text, tooltips will show on the Network Chat. Also, the cursor will turn to a pointer, mean the user can click. After clicking either on the Word Cloud or the Network Chart, the component below will react dependently on the tags they click so the user will see the change and lead to the next two component we have. Besides, the users can double click on the Word Cloud or the Network Chart to turn the Network Chart into Flower Chart. Flower Chart allow users to connect their interested tag with other tags it may have. This help users to find their next tags they may want to add on the Buttons. The number next to the a button is the number of video that has this tag.

Our Line Chart and Video List Table come to their eyes when the user scroll down in the page. As we mention above, all of the components in our design are able to interact intuitively. So the users' next step would be try to hover the things we have here. In the Line Chart, every components is able to interact by the hover events we built and both the Line Chart and Video List Table will react accordingly. This let the user to discover the video in TED

website from the topic they are interested in. When the user try to see into the detail on the Video List Table, they can discover the Radar Chart first with the sufficient information along with it. The radar chart can let the users have a first picture of how other people thinks about this video then the user can decide whether to view this video themselves.

4.2 Interaction

4.2.1 Network Chart

4.2.2 Word Cloud

4.2.3 Line Chart

4.2.4 Buttons

4.2.5 Video List Table

4.3 Flow-Chart

Chapter 5

Evaluation

5.1 Solution in our question

Bibliography

[1] D3.Symbol. https://github.com/d3/d3/blob/master/API.md#symbols.