Cross Cultural Analysis Report

Xu Han (xh2379@columbia.edu)
Columbia University

Abstract

Our main goal is to distinguish cultural differences between different videos from different countries that focus on the same event. This is overall a challenging task. We take several attempts to approach our goal, including building models, extracting useful features, finding better ways to acquire materials, etc. In this report, we provide a brief discussion on the progress we make and difficulties we face during our explorations, including downloading videos, brainstorming new topics and furnishing techniques on automating the process to improve efficiency.

1. Introduction

In the previous semester, we had made some progress toward our goal to distinguish cultural differences between different videos focusing on the same event. By extracting nearly duplicate images from the videos (Figure 1) and texts from transcripts and descriptions, we analyzed the cultural aspects among different videos from multiple sources. We mainly focused on the event that Alphago won a Chinese Go player Ke Jie. This event attracted world attention, especially from China, and we used it to study the cultural differences between China and the US.

Figure 1. Nearly duplicate images from different videos

Much of the progress was done my manually inspecting the videos and look for cultural similarities and differences. This is useful during the beginning of the exploration but could be

very time consuming and tedious. Many videos are not useful and some videos can be very long. A problem that we faced when looking at Alphago videos is that many of these videos contain a very long time of the game itself rather than comments and interviews which is more useful for cultural analysis. So perhaps the event Alphago might not be a perfect choice for such purpose.

To overcome these problems, we came with a few new ideas, such as choosing events in a smarter way and making the process more algorithmic. Our current pipeline is much more fluent than before and requires significantly less manual inspection, though manual inspection still plays a very important role.

2. Improvements on Previous Work

2.1. Stabilizing the Download Pipeline

An important component of this research project is obtaining videos. In order to extract information from the videos and analyze, we first need to download the videos. Our method is to parse the search result page (Figure 2) to obtain the URLs of the videos, and then use a handy tool named "you-get" (https://github.com/soimort/you-get) to download the videos from the URLs.

However, previously this process was not stable and we were also not very familiar with you-get. So the process was not able to become automated. This semester, we organized the process in a better way and carefully engineered our code for scraping and for running you-get. Currently, the download pipeline is much more stable and easier to run.

2.2. New Source for Chinese Videos: Youku

Previously we used Youtube for the source of US videos. Youtube has a huge amount of videos and provides friendly APIs for downloading its videos, thus making it a perfect source for US videos. For Chinese videos, however, there does not exist a video source that has all these advantages. So we used multiple sources, such as Bilibili (https://www.bilibili.com), Iqiyi

(https://www.iqiyi.com) and Tencent Videos (https://v.qq.com). Bilibili videos are easy and fast to download, but the site has a relatively small number of videos. Iqiyi and Tencent Videos have a much larger size (still tiny comparing to YouTube), but their videos are either slow or difficult to download.

In order to obtain more videos and have better download speed, we developed a new scraper for the Chinese video website Youku (https://www.youku.com) (Figure 2). It has about the same amount of videos as Tencent Videos and it provides a decent speed. Thus currently we have a total of four Chinese video sources which can provide a satisfying amount of videos and download speed. If we need more, it should not be difficult to develop scrapers for other video websites as the pipeline for downloading videos are currently stable and under control.

Figure 2. The search result page of the video source Youku

3. New Events Other than Alphago

Previously we focused on a single event, Alphago, to perform our research. This has many drawbacks. With a single event, there may be results that actually only appear in this specific event rather than happening in general. For example, Alphago videos tend to have many nearly duplicate images showing the Go game. Thus we need to explore new events in order to achieve more general results. We would like events that are followed by both China and the US, easy to gather information and shows cultural differences. Also, we would like events from different aspects, such as technology, nature, social and so on. This will help improve generality of our results.

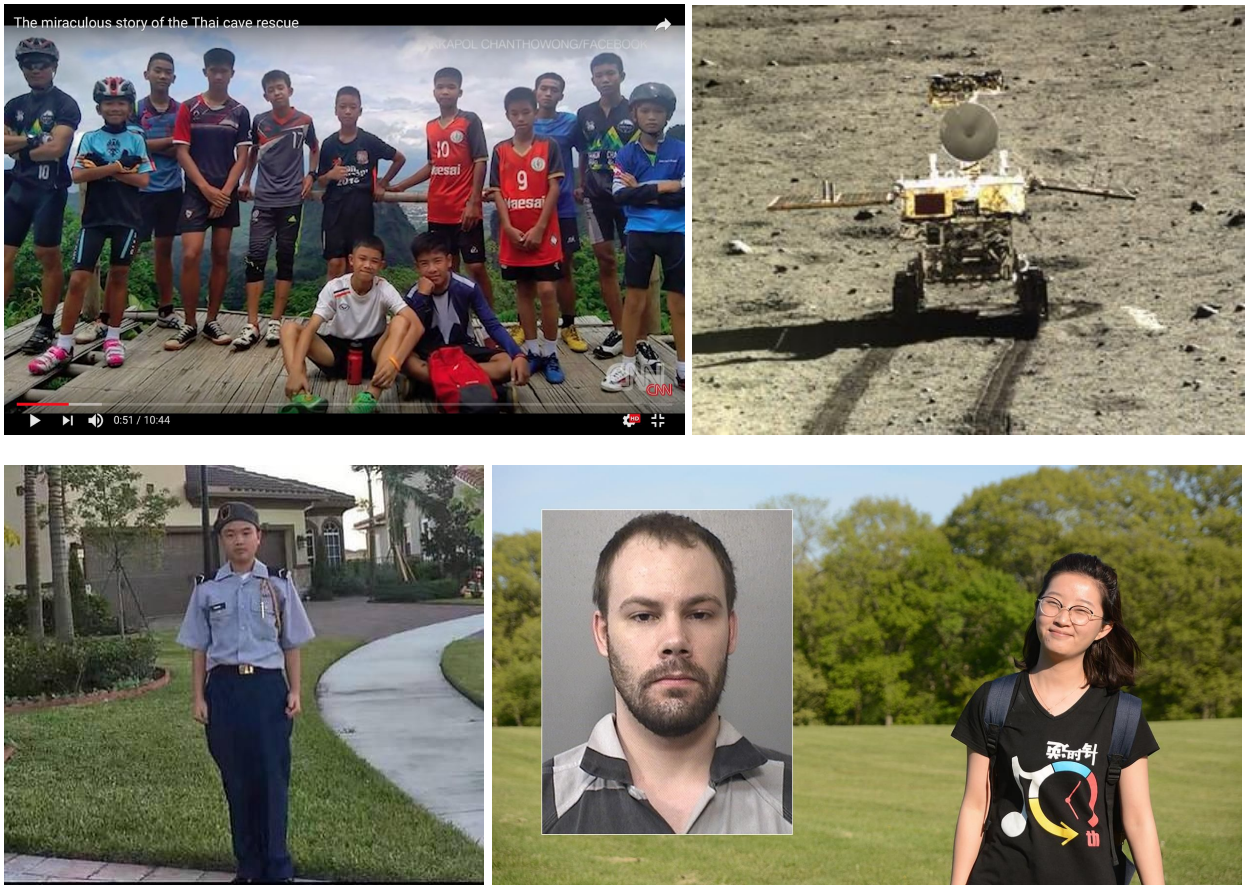

Figure 3. (1) Top-left: Members and the coach of the football team trapped in Tham Luang cave (2) Top-right: Chinese lunar rover, Yutu (3) Bottom-left: The heroic victim Peter Wang in Stoneman Douglas high school shooting (4) Bottom-right: The kidnapped Chinese visiting scholar Yingying Zhang and the suspect Brendt Christensen

3.1. Tham Luang Cave Rescue [1]

Twelve members of a football team of Thailand and together with their coach were trapped in a cave after they entered it for practice and heavy rain partially flooded the cave. They were finally rescued miraculously after being trapped for more than two weeks. This event received wide attention among the world.

3.2. Chinese Lunar Rover [2]

The Chinese lunar rover, Yutu, had landed on the moon and collected significant amount of useful research data. According to Wikipedia, "The mission marks the first soft landing on the Moon since 1976 and the first rover to operate there since the Soviet Lunokhod 2 ceased operations on 11 May 1973" [2]. This event attracts an extremely high amount of attention in China, since it marks a significant milestone of Chinese technology, and also in the US, for the US also plays a very important role in high technology and space exploration.

3.3. Stoneman Douglas High School Shooting [3]

On February 14, 2018, a gunman opened fire at Marjory Stoneman Douglas High School in Parkland, Florida, killing seventeen students and staff members and injuring seventeen others [3]. In this event, we focused on one specific victim, Peter Wang, who held the door open in order to create an escaping opportunity for the others and was shot dead in the incident. He is a Chinese American, which makes him very suitable for the research of cultural difference between China and the US.

3.4. Kidnapped Chinese Scholar [4]

Yingying Zhang was a visiting scholar in the United States from China, who has not been seen since she got into a car at a bus stop on the University of Illinois at Urbana–Champaign campus on June 9, 2017 [4]. Based on evidence uncovered during the investigation, law enforcement officials said they believed Zhang was no longer alive [4]. This is similar to the event of the

Florida high school shooting, which is another tragedy that received very much attention from both China and the US.

4. Attempts for Finding More Suitable Events

All of the events mentioned previously are worked through brainstorming and discussions. Just like automating the download pipeline for acquiring videos, we would also like the process of obtaining suitable events a less time consuming one. In order to do this, we made use of a useful resource, Google Trend, which gives very useful information for our purpose.

4.1. Google Trend

Google Trend is a website under Google that can visualize the popularity of certain search queries. We can specify the time interval, the region and other parameters such as topic and it will return a graph that shows the change of trend. For example, Figure 4 clearly shows the periodic feature of the popularity of the term “Olympics”. Note that the result does not show the actual number of searches. Instead, it scales the most popular data point to 100, and other data points accordingly. By carefully studying the results of Google Trend, we might be able to work out a decent way to obtain suitable events.

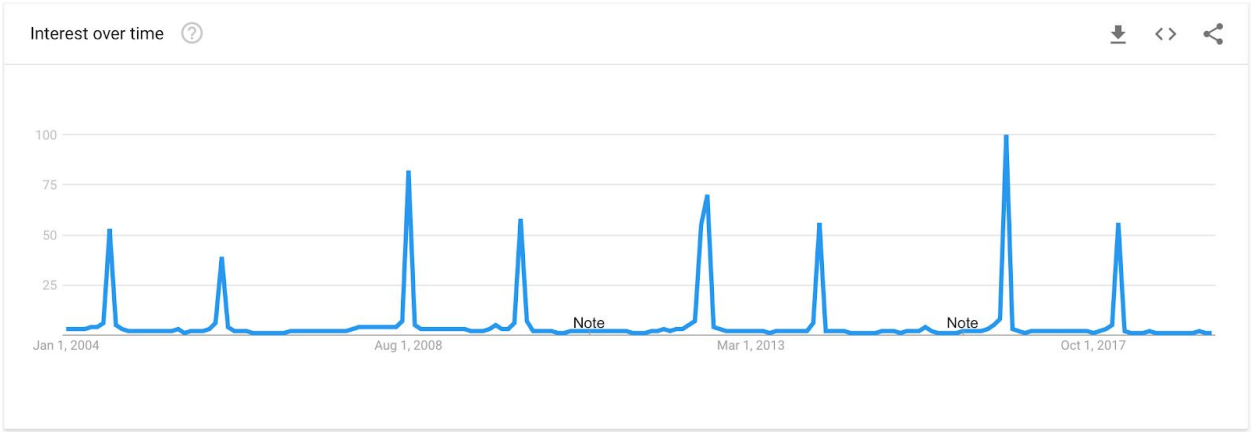

Figure 4. The result of Google Trend for the term “Olympics”

4.2. Capturing Data from Google Trend Using pytrends

Using the browser to type queries and then looking at the results is obviously not a good idea to automate the whole process. Therefore, we need a method to algorithmically obtain the data and perform analysis on it. In order to do so, we used a tool named pytrends (https://github.com/GeneralMills/pytrends). It is a non-official API that provides a handy way to acquire data from Google Trends in Python. Below is an example of such data.

	World Cup	Hurricane Florence	...	Chicago Bears	Milwaukee Brewers
2004-01-01	1	0	...	11	4
2004-02-01	1	0	...	6	4
2004-03-01	1	0	...	7	8
2004-04-01	0	0	...	11	13
2004-05-01	1	0	...	8	15
2004-06-01	1	0	...	9	15
2004-07-01	1	0	...	9	17
2004-08-01	1	0	...	21	8
2004-09-01	1	0	...	25	7
2004-10-01	1	0	...	17	2
2004-11-01	1	0	...	18	3
2004-12-01	0	0	...	14	3
2005-01-01	0	0	...	9	5
2005-02-01	1	0	...	7	6
2005-03-01	1	0	...	8	8
2005-04-01	1	0	...	10	17
2005-05-01	1	0	...	7	15
2005-06-01	1	0	...	7	14
2005-07-01	1	0	...	11	12
2005-08-01	1	0	...	30	11
2005-09-01	1	0	...	26	8
2005-10-01	1	0	...	21	2
2005-11-01	1	0	...	35	2
2005-12-01	2	0	...	36	3
2006-01-01	1	0	...	27	3
2006-02-01	1	0	...	8	8
2006-03-01	1	0	...	8	9
2006-04-01	2	0	...	11	22
2006-05-01	5	0	...	9	17
2006-06-01	52	0	...	8	15
...	...	...	...	...	...
2016-11-01	1	0	...	33	2
2016-12-01	1	0	...	29	3
2017-01-01	1	0	...	16	3
2017-02-01	0	0	...	14	4
2017-03-01	1	0	...	25	6
2017-04-01	1	0	...	24	21
2017-05-01	1	0	...	17	21
2017-06-01	2	0	...	13	25

2017-07-01	1	0	...	13	33
2017-08-01	1	0	...	36	27
2017-09-01	1	0	...	59	34
2017-10-01	3	0	...	48	4
2017-11-01	1	0	...	33	3
2017-12-01	1	0	...	34	4
2018-01-01	1	0	...	26	8
2018-02-01	1	0	...	16	10
2018-03-01	1	0	...	30	14
2018-04-01	1	0	...	30	31
2018-05-01	3	0	...	18	29
2018-06-01	89	0	...	14	34
2018-07-01	56	0	...	18	44
2018-08-01	1	0	...	42	37
2018-09-01	1	100	...	87	65
2018-10-01	1	2	...	60	100
2018-11-01	1	0	...	78	5
2018-12-01	1	0	...	99	7
2019-01-01	1	0	...	63	9
2019-02-01	1	0	...	17	9
2019-03-01	1	0	...	29	20
2019-04-01	1	0	...	27	46

[184 rows x 79 columns]

We can also make a plot of part of the data in Python to help better understanding (Figure 5).

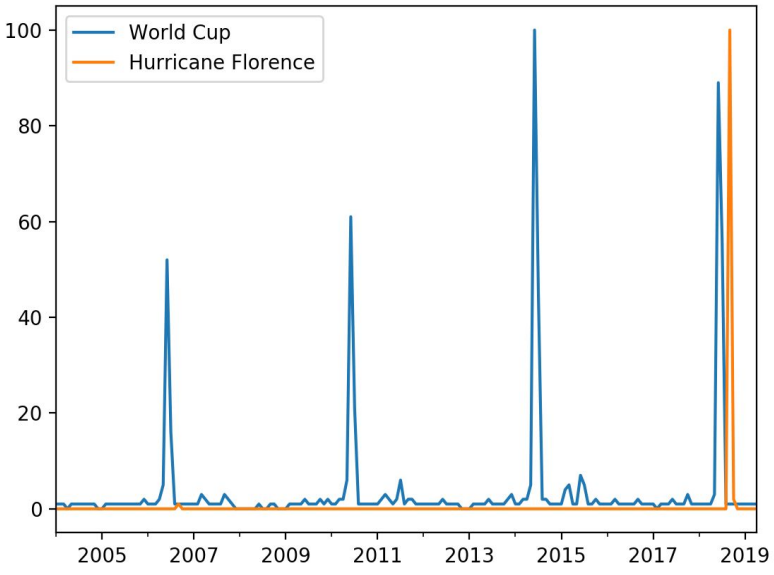

Figure 5. Python plot of Google Trend data

4.3. Observations of Different Events on Google Trend

From our observations of many different topics, we found that there are some interesting patterns that we might make use of. For instance, we can look at Figure 6 and see the difference between "Donald Trump" and other queries such as "Yingying Zhang", "Peter Wang" and "Thailand Cave Rescue". A major difference is that the popularity of "Donald Trump" stays relatively high in a very long time while all other three shows a sudden spike in its popularity over time. What coincides with this observation is that all three events are relatively easier to find nearly duplicate images from videos. Our guess is that if something remains popular for a very long time, there will be more information related and thus there will be more "noise". On the contrary, those events having a spike tend to have less information other than the event happened at the spike.

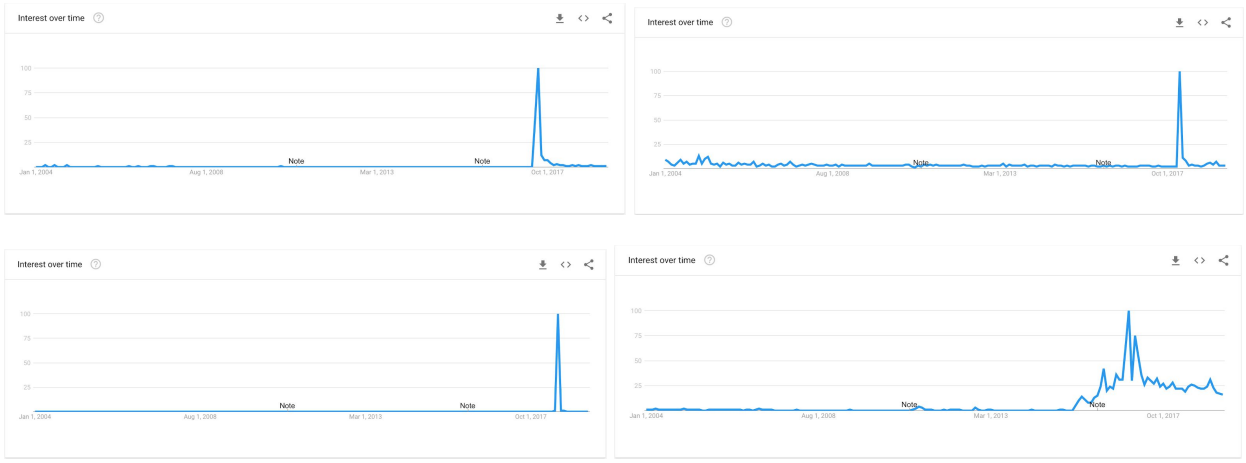

Figure 6. Results of Google Trend given queries: (1) Top-left: Yingying Zhang (2) Top-right: Peter Wang (3) Bottom-left: Thailand Cave Rescue (4) Bottom-right: Donald Trump

4.4. Compare Maximum to Average

Based on our observations and hypothesis that those events having a spike tend to provide more shared images, which is useful for our purpose, the next step is to find a way to identify such spikes. A straightforward way is to compare the maximum value with the average value over time. If the maximum value is significant comparing to the average overall popularity, then we identify it as a spike.

This will successfully identify the spikes in the previous examples, since they are dominating in the entire graph. However, this method fails on many cases including, for instance, the case where there are multiple spikes, such as “World Cup” in Figure 5. Also, it is very unstable since it only considers a single value, the maximum. Thus we need to make improvements on this method.

4.5. Compare Maximum to Moving Average and Count Spikes

As mentioned in the previous section, the simple approach of comparing the maximum value to the average faces many downsides. So we made the following improvements: Instead of using the overall, use the moving average. In addition, count the number of spikes instead of considering the maximum value only. A sample result is shown below. The score of a data point is calculated as the smoothed ratio of its value and the moving average over a sliding window of size 24 (namely the average of the past 24 months). A data point is identified as a spike if its score is greater than 3. In the results, the column “score” is the score of the maximum data point and the column “spikes” is the number of data points identified as spikes.

	score	spikes	time
Olympic Medal Count	7.76471	5	2016-08-01 00:00:00
Hurricane Florence	7.76471	1	2018-09-01 00:00:00
Hurricane Michael	7.74194	1	2018-10-01 00:00:00
Kavanaugh Confirmation	7.69679	2	2018-09-01 00:00:00
The Haunting of Hill House	7.67442	2	2018-10-01 00:00:00
Altered Carbon	7.54286	1	2018-02-01 00:00:00
Royal Wedding	7.54286	2	2011-04-01 00:00:00
MOLLIE TIBBETTS	7.5	1	2018-08-01 00:00:00
Dubrow diet	7.43662	2	2018-10-01 00:00:00
Naomi Osaka	7.41573	1	2018-09-01 00:00:00
Chloe Kim	7.35376	1	2018-02-01 00:00:00
AVICII	7.25275	1	2018-04-01 00:00:00
ARETHA FRANKLIN	7.19346	1	2018-08-01 00:00:00
Johnny Weir	7.19346	4	2018-02-01 00:00:00
Brett Kavanaugh	7.17391	3	2018-09-01 00:00:00
ANTHONY BOURDAIN	7.13514	1	2018-06-01 00:00:00
World Cup	7.13514	6	2014-06-01 00:00:00
Urban Meyer	7.1159	1	2018-08-01 00:00:00
Florida Shooting	7.1159	1	2018-02-01 00:00:00
Shaun White	7.09677	3	2018-02-01 00:00:00
MAC MILLER	7.04	1	2018-09-01 00:00:00
BURT REYNOLDS	6.73469	1	2018-09-01 00:00:00

The Shepherd's Diet	6.71756	2	2017-02-01 00:00:00
STAN LEE	6.6	1	2018-11-01 00:00:00
XXXTENTACION	6.51852	1	2018-06-01 00:00:00
Lost in Space	6.51852	1	2018-04-01 00:00:00
Insatiable	6.42336	1	2018-08-01 00:00:00
Election Results	6.42336	6	2016-11-01 00:00:00
Lindsey Vonn	6.22642	3	2018-02-01 00:00:00
Nick Foles	6.21176	2	2018-02-01 00:00:00
...	...	...	...
Eminem	3.95802	1	2018-09-01 00:00:00
Cardi B	3.59673	2	2018-04-01 00:00:00
On My Block	3.36735	4	2019-04-01 00:00:00
Milwaukee Brewers	3.29177	1	2018-10-01 00:00:00
American Idol	3.27543	1	2006-05-01 00:00:00
Chicago Bears	3.0662	1	2007-01-01 00:00:00
New York Yankees	3.03797	1	2017-10-01 00:00:00
Los Angeles Dodgers	3.00683	1	2017-10-01 00:00:00
Cleveland Browns	2.98981	0	2018-09-01 00:00:00
Boston Celtics	2.83262	1	2018-05-01 00:00:00
Mediterranean diet	2.82051	0	2019-01-01 00:00:00
Fasting diet	2.65327	0	2019-01-01 00:00:00
Castle Rock	2.59843	0	2018-08-01 00:00:00
Los Angeles Lakers	2.53116	0	2018-07-01 00:00:00
Ariana Grande	2.4696	0	2014-09-01 00:00:00
Philadelphia Eagles	2.39347	0	2018-02-01 00:00:00
Dr. Gundry diet	2.27194	0	2018-09-01 00:00:00
Cleveland Cavaliers	2.12732	0	2018-05-01 00:00:00
Keto diet	2.04651	0	2019-01-01 00:00:00
2000s fashion	2.01373	0	2018-10-01 00:00:00
80s mens fashion	1.72662	0	2018-09-01 00:00:00
Grunge style	1.53667	0	2017-10-01 00:00:00
Hipster style	1.47404	0	2012-12-01 00:00:00
Men's Fashion	1.35246	0	2016-12-01 00:00:00
Fodmap diet	1.30113	0	2019-02-01 00:00:00
1980s fashion	NaN	0	2004-05-01 00:00:00
1990s fashion	NaN	0	2004-05-01 00:00:00
Harajuku fashion	NaN	0	2005-02-01 00:00:00
Corbra Kai	NaN	3	2005-04-01 00:00:00
Boston Red Sox	NaN	2	2004-10-01 00:00:00

[79 rows x 3 columns]

Here is a visual demonstration of how the algorithm works (Figure 7).

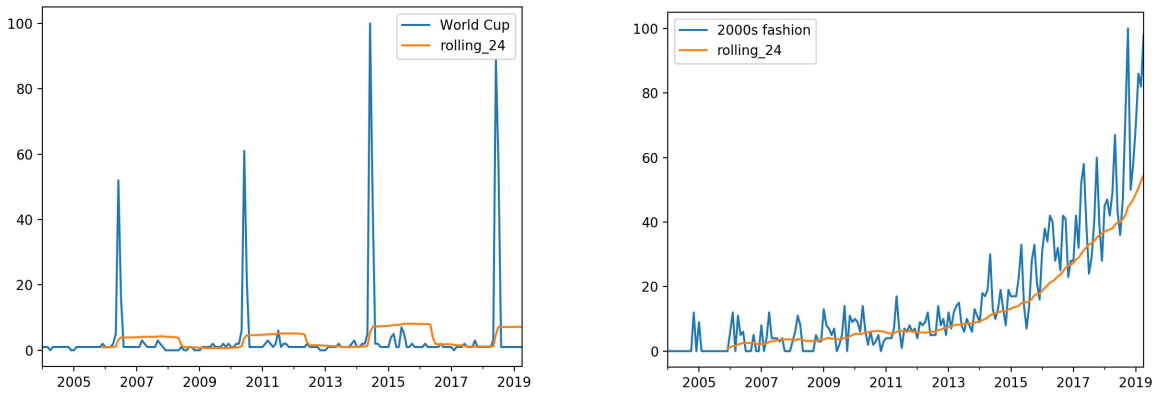

Figure 7. Demonstration of Moving Average

In the figures, the orange line is the moving average line. We can see that it is much smoother than the original line, leading to increased stability, and thus provides a good statistics of measuring how likely a data point is a spike. By counting the number of spikes, we overcome the downside that multiple spikes may lead to more noisy data. The fewer spikes, the more likely nearly duplicate images can be found.

4.6. Year in Search

To this point, the efforts in the above sections are focused on verifying if an event likely contains shared images. However, that still leaves a question: How do we get the events in the first place? This is where the "Year in Search" function of Google Trend (Figure 8) comes into account. It provides the topics that are the most popular in previous years. We can choose regions and we can get 10 topics from each category. There are many categories, including food, fashion, people, etc. This not only provides abundant events for our research, but also provides an opportunity to observe characteristics among different categories.

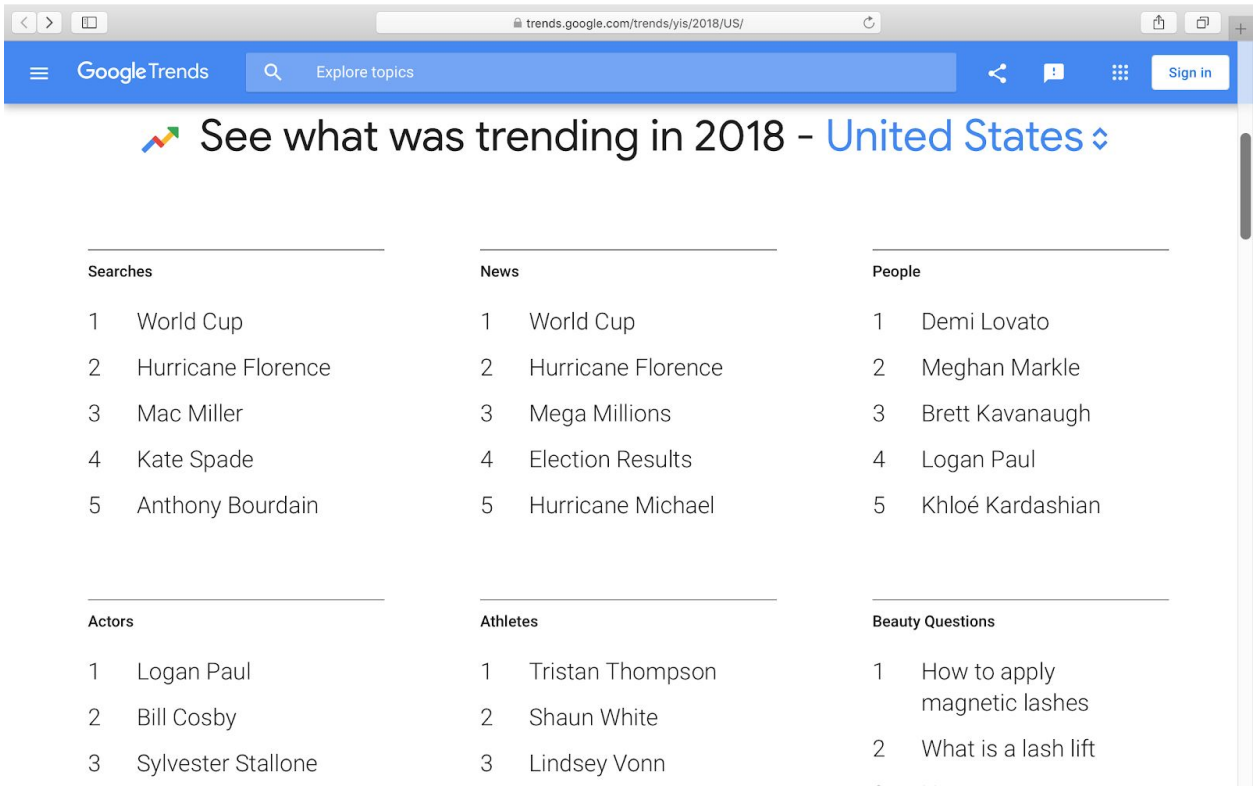

Figure 8. Year in Search 2018 in Google Trend

5. Results

5.1. Download

The download pipeline for this project is now stable. Although not perfect, significant amount of work should not be necessary to acquire videos needed. Currently we have over 600 videos for each language (Chinese and English) and for each event. Youtube is always a stable and efficient source for English videos. For Chinese sources, all four are stable. Bilibili and Youku are efficient, while Tencent Videos and Iqiyi sometimes face difficulties in terms of downloading speed. However, since what we need is usually only several hundred videos, this should not be a huge problem.

5.2. Ranking of Topics from Different Categories

Looking at the results from section 4.5, we see that news, athletes, deaths, together with a few TV shows usually have higher score. This might suggest that they will provide less noisy

information in videos. We dug through some of the events and looked at the relating videos. After some efforts, we found that some events, including natural disasters, deaths and events that happen specifically in some small area (such as September 11 attacks) usually give shared images, which partially verifies the observation that news and deaths have higher scores. On the other hand, topics such as sports events and athletic games, when giving the athlete's name only as the query, contain more noisy information, which contradicts the ranking. The reason for this might be that giving only the name returns many different games of the athlete. If the search of videos is restricted to some specific event of the athlete, the situation becomes better. Therefore, our method of ranking the topics obtained from Google Trend can serve as a filter and may provide us with suitable ones for our research.

5.3. The Effect of Deaths

During our exploration of finding events that provide easier-to-find nearly duplicate images, we found a special phenomenon: When somebody dies, it is very likely that a spike appears in Google Trend data and we could find nearly duplicate images in videos. We have already mentioned that events including some news and deaths have higher ranking but deaths give very good results compared to other categories that it needs special indication. Typical examples are the high school shooting event and kidnapping event whose data are already shown in Figure 6. Two other examples are shown below in Figure 9. All four has huge spikes. Another observation comes from the result shown in section 4.5. The names of the dead are capitalized. We can see that Their faces appear very frequently in many different videos. In addition, for people who are not famous, such as Peter Wang and Yingying Zhang, the portraits are almost shared in every video. This phenomenon is probably because the deceased could no longer produce any more photos, limiting the noise of the event. So videos can only use the photos in the past, which limits the total number of such photos. In the case that the person was not famous, such photos are rare, causing more sharing among different videos.

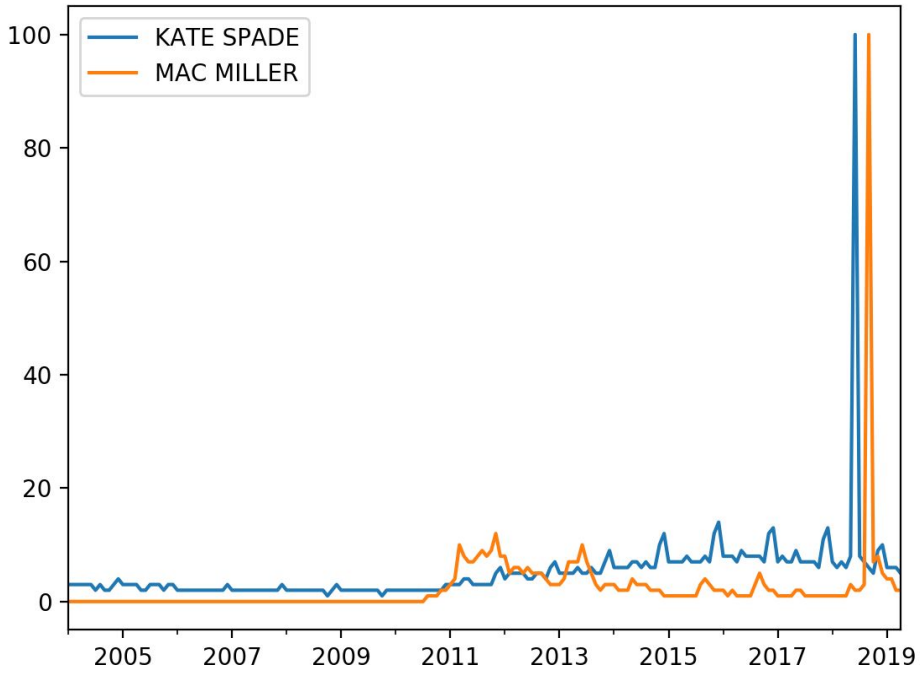

Figure 9. Google Trend data for Kate Spade and Mac Miller

6. Conclusion

This semester's work is based on the previous semester with some improvements and some new ideas, with the goal of understanding cultural differences from different videos remain unchanged. The download pipeline is completed in this semester. Now the downloading process is much more fluent and easier to handle. Previously we focused on a single event, Alphago. This semester, we introduced several new events and downloaded over 1000 videos for each event. Besides that, we also made attempts to find suitable topics by making use of the data from Google Trend, a handy tool provided by Google. We take the topics from the Year in Search function of Google Trend, and rank them according to our algorithm. Our observations suggest that some topics are usually ranked high, especially deaths (the "Loss" section in Year in Search), and are much more likely to give nearly duplicate images in different videos, which is very useful for our purpose.

7. Reference

1. Wikipedia contributors. (2019, May 7). Tham Luang cave rescue. In *Wikipedia, The Free Encyclopedia*. Retrieved 23:10, May 18, 2019, from https://en.wikipedia.org/w/index.php?title=Tham_Luang_cave_rescue&oldid=895947021
2. Wikipedia contributors. (2019, May 10). Yutu (rover). In *Wikipedia, The Free Encyclopedia*. Retrieved 00:43, May 19, 2019, from https://en.wikipedia.org/w/index.php?title=Yutu_(rover)&oldid=896484648
3. Wikipedia contributors. (2019, May 12). Stoneman Douglas High School shooting. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:29, May 19, 2019, from https://en.wikipedia.org/w/index.php?title=Stoneman_Douglas_High_School_shooting&oldid=896721266
4. Wikipedia contributors. (2019, April 21). Disappearance of Yingying Zhang. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:45, May 19, 2019, from https://en.wikipedia.org/w/index.php?title=Disappearance_of_Yingying_Zhang&oldid=893391433