**presentation**

**1**

Good afternoon, everyone! We are Group 7. It's my honor to speak here, and I am very glad to share our group project with you.

Our project is about Data Statistics and Visualization of social media, we grabbed some hotlist data from Weibo and Zhihu, and then analyzed the results.

**2**

The following speech consists of 4 parts: Introduction, Crawl hot list data, make wordcloud and the final discussion.

**1 introduction**

**Sample Selection Index**

Firstly, we hope to observe the current social attention through the visual word cloud. Therefore, we chose hot list data from two popular media (Weibo, Zhihu).

The word with more weight in the word cloud means that people pay more attention to.

And then we have two comparisons, in the vertical comparison, we look at how social concerns change through the time. And Horizontal comparison is to observe the similarities and differences of different popular websites.

by generating short-term hot topics and long-term hot topics

**Zhihu and weibo**

Zhihu and weibo are two mainstream media Websites in China.

They are similar but have some differences, and we will discuss later.

Hot search in Weibo reflects current affairs news and entertainment news.

And Zhihu is a question-and-answer platform. In addition to news discussions, Zhihu Hot List will also have some technical questions and answers, or some strange and popular questions.

**Steps**

There are 3 steps. First is Scraping hotlist data of Weibo and Zhihu, and then extract keywords, in this step, we use two different keyword extraction methods and finally compare them. the third step is generating word clouds, and compare the results horizontally and vertically. in two dimensions

**2 Crawl hot list data**

And then let’s talk about part two.

**Crawler tools**

There are some crawler tools, for example, BeautifSoup and Selenium. We choose Beautiful Soup to crawl Weibo and Zhihu's hotlist data. It is a Python library for pulling data out of HTML and XML files.

**Take with Cookies**

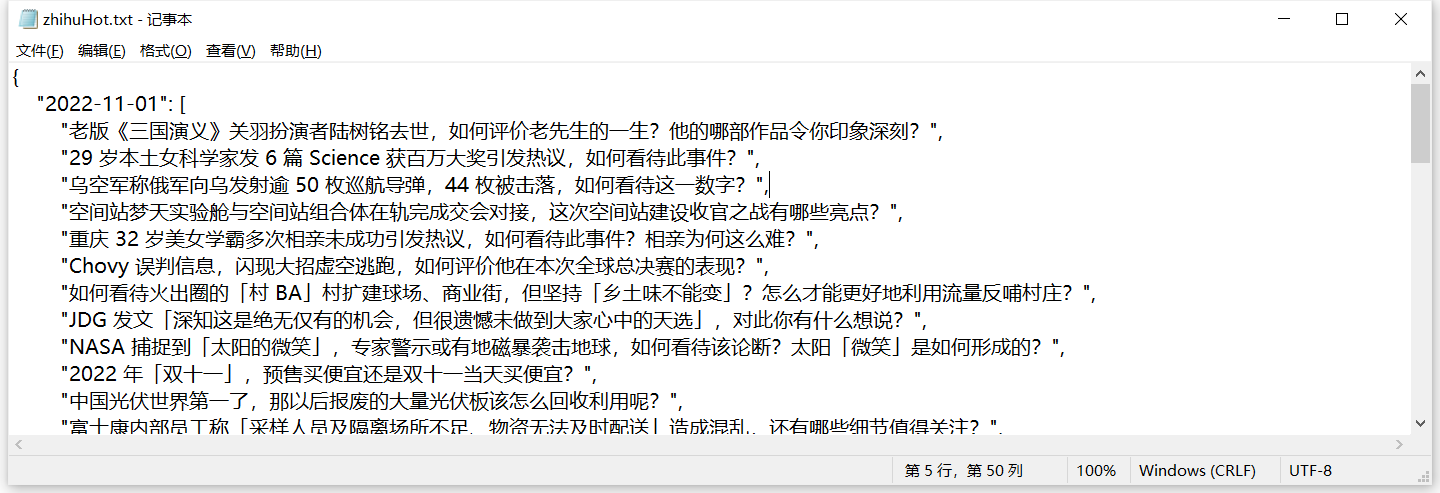
Weibo and Zhihu hotlist data requires user login verification. To bypass permission verification, we need to get the login status Cookie from the logged in user's page.

In more details, First create a crawler class Which has 4 attributes:

1. the login authentication cookie
2. the url of the site to be crawled
3. the request header
4. the final data to be obtained as properties of the crawler.

After connecting to the site, we request the site text and use beautifulsoup to clean returned data.

We get the data corresponding to the hotlist tag using API find and find\_all, and then After data return, we store it in the dictionary and convert it to json format. Finally, we obtain the data files like that.



**3 wordcloud**

In the third part, wordcloud.

There are also 3 sections in this part.

First, we need to extract words from the previous step. we use Jieba to help us to segment Chinese words. After that, we use tf-idf and tf algorithm to statistics word frequency.

About the wordcloud, we import open-source library WordCloud from github.

Finally, Analyze the trending with Vertical comparison & Horizontal comparison.

**Compare TF and TFIDF**

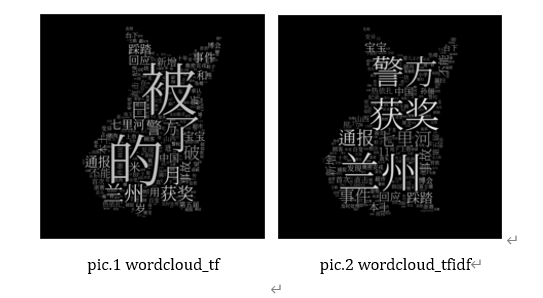
We compared the results of using TF and TFIDF.

For keyword extraction, the simplest method must be obtained by word frequency, and TF-IDF is a more advanced method.

When counting word frequency, we know that some meaningless words have very high word frequency. Obviously, these words cannot appear as keywords, so it is necessary to pay attention to word filtering.

Look at those two wordclouds.

When using Term Frequency algorithm, we can see some stop words such as “被”, “的”have high term frequency. The second word cloud which use TFIDF looks better.



**Vertical Comparation**

To compare the time trend of the hot search keywords,

We prepare the corpus use the hot search data of different days as, and want to get the keywords of a certain day.

From the word clouds, Itaewon Stampede was the main focus of social from October 31 to November 2.



**Horizontal Comparation**

we also compare the hot search data of Zhihu and Weibo on Nov2

We can see Zhihu has some professional terms in its hot words, while Weibo is more of a daily news

**4 Discussion**

**improvements**

Finally, we reviewed our project and felt the following improvements could be made:

First, Beautify word cloud: Add various colors to the word cloud or enriching the pattern of the word cloud.

we can use TextRank to extract keywords.

a comprehensive keyword extraction algorithm can be obtained by weighting the above algorithm