CCM Report

Group Introduction

Group_Name: FullgerV50

Group_Number: 5

Class: [Wed 1415 周三]CMM 云计算

Members Introduction

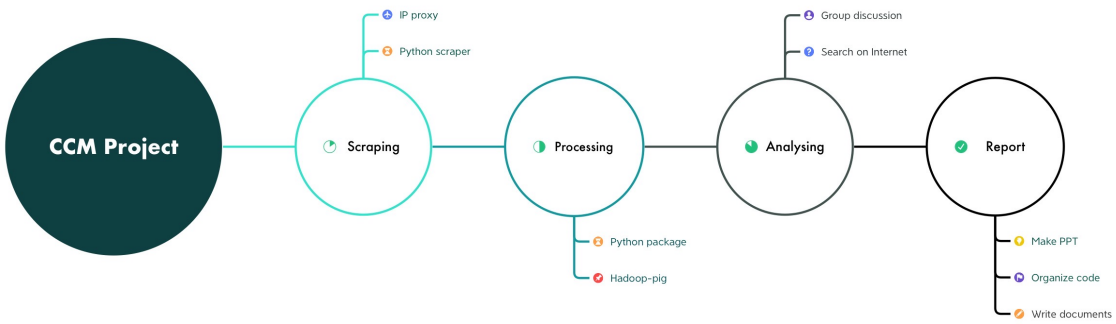
EN_NAME	CH_NAME	Student ID	分工
Sebastian	殷健朝	20201003028	Team Organiser; News Collector; Double-check the consistency between the result and the news; Translater; Docking between the teacher and the team members.
Alex	饶文均	20201002954	Define the direction of the project; Do the scraping job; Draw the flow charts and schematic diagrams in the report; Map word clouds; Organize report writing.
Vivi	吴婉琳	20201003130	Write the visualization process part of the report and analysis the process of making word clouds.
Mody	古津铭	20201003113	Integrate and analyse extension of the research in two presentations; Be responsible for mapping and visualizing the data analysis; Using a communication perspective to analyse the research content in response to public opinion trends.
Kmpa	林健杰	20201003054	Process raw data; Help to find knowledge; Code organizing; Help to write the document.

Project Description

This project aims to scrap the news released by BBC News in its official account of Facebook and analyzes the recent(As of December 5th,2022) changes in public opinion and the resons for such changes.

1. Data Quality: It should be ensured that the news data scraped is of high quality and up to date.
2. Data Analysis Tools: Appropriate data analysis tools should be selected for analysis so as to better extract valuable information.
3. Security: When accessing news data, pay attention to protecting the privacy and avoid being attacked or monitored by hackers.
4. Reliability of Results: During the analysis, the reliability of results shall be ensured to avoid deviation due to poor data quality or improper analysis methods.

This is our workflow diagram



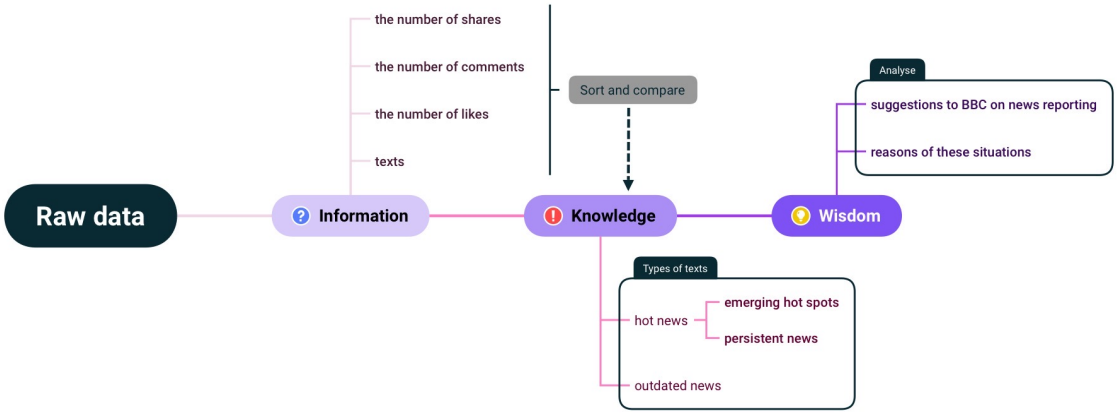
Input Data

The CSV file obtained by the Facebook scraper, including 52 fields such as timestamp, user ID, text, likes, shares and so on.

Output Data

- Top 50 words in all news texts;
- Top 50 words in all news texts sorted according to the number of likes.
- Top 50 words in all news texts sorted according to the number of comments.
- Top 50 words in all news texts sorted according to the number of shares.

This is the flow chart of our data processing



Target

Obtain the recent(As of December 5th,2022) news with the highest topic/the most popular news/the news that is gradually losing popularity(outdated) by analysis and compare the output results. This will make some suggestions for the netword media, which can be about the news content or about the time and frequency of news release.

Processing Procedure:

Data Scraping

Scraping Facebook through the Facebook Scraper package, proxy is used to prevent accounts from being banned due to a large number of accesses. Using proxy during scraping has the following advantages:

1. Hide the real IP address: Proxy can be used to hide the real IP address, which is very helpful to prevent being banned or attacked by the target website.Using proxy can make it much more difficult for the scraper to be detected by the scraper system.
2. Circumvention of regional restrictions: Facebook has regional restrictions, which prevent normal access to Chinese Mainland. Proxy can be used to circumvent this restriction by changing the region of the proxy server.
3. High speed of scraping: Proxy can be used to send requests to faster servers to speed up the scraping sequence.
4. High reliability and concurrency: Using proxy can make the scraper more reliable since it can automatically switch IP addresses, so as to avoid the paralysis of the entire scraper due to the blocking of a single IP address.
5. High concurrency: Using proxy allows the scraper to use multiple IP addresses to scrape at the same time, which can greatly improve the concurrency of the scraper.

Data Extraction

	post_id	text	post_text	shared_te	original_	time	timestamp	image	image_low	images	images_de	images_lo	images_lo	video	video_dur	video_hei	video_id	video_qua	video_siz	video_thu	video_wat	video_wid	likes	comments	shares	post_url	link	links	user_id
1																													
2	0	5.49E+14	"Lego brings people together and encourage	"Lego bri	BBC.COM Baby hor christene d in UK's ' first Lego		00:38.0	1.67E+09	https://e:	[]	[]	['https:/	[None]										592	291	79	https://fi	https://bl	[{'link':	1.001E+
3	0	5.49E+14	"They're able to access things that emotional	"They're able to access things that emotional			04:28.0	1.67E+09	https://s	[]	[]	['https:/	[None]	https://scontent-seal-1.xx.fb	5.634E+14					https://scontent-seal-1.xx.fb			1307	161	228	https://facebook.co		[{'link':	1.001E+
4	0	5.49E+14	"We will not accept this ceiling."	"We will	BBC.COM Russia rejects oil price cap imposed		59:37.0	1.67E+09	https://e:	[]	[]	['https:/	[None]										908	673	29	https://fi	https://bl	[{'link':	1.001E+

From the data we scraped, we can see that there are close to 30 columns of data in fb_scrapped_bbc.csv, which are 'Unnamed: 0', 'post_id', 'post_text', 'shared_text','original_text', 'time','timestamp','image','image_lowquality','images','images_description','images_lowquality','images_lowquality_description','video','video_duration_seconds','video_height','video_id','video_quality','video_size_MB','video_thumbnail','video_watches','video_width','post_url','link','links','user_id','username','user_url','is_live','factcheck','shared_post_id','shared_time','shared_user_id','shared_user

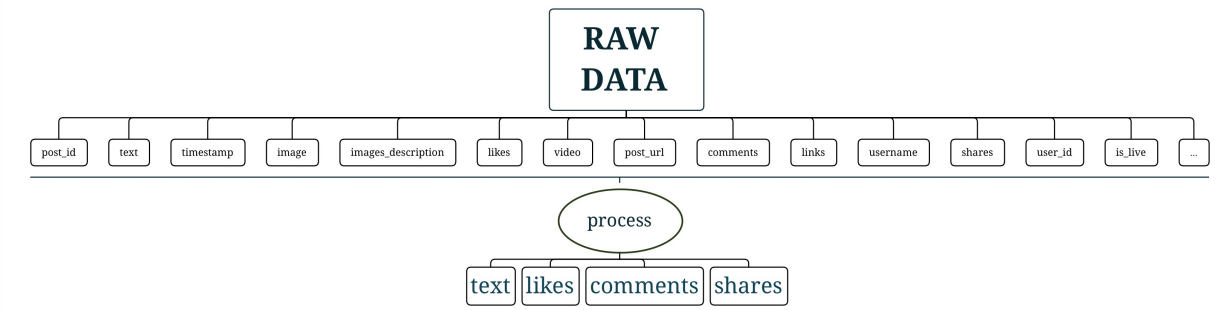
name','shared_post_url','available','comments_full','reactors','w3_fb_url','reactions','reaction_count','with', 'page_id', 'sharers', 'image_id', 'image_ids', 'was_live'.

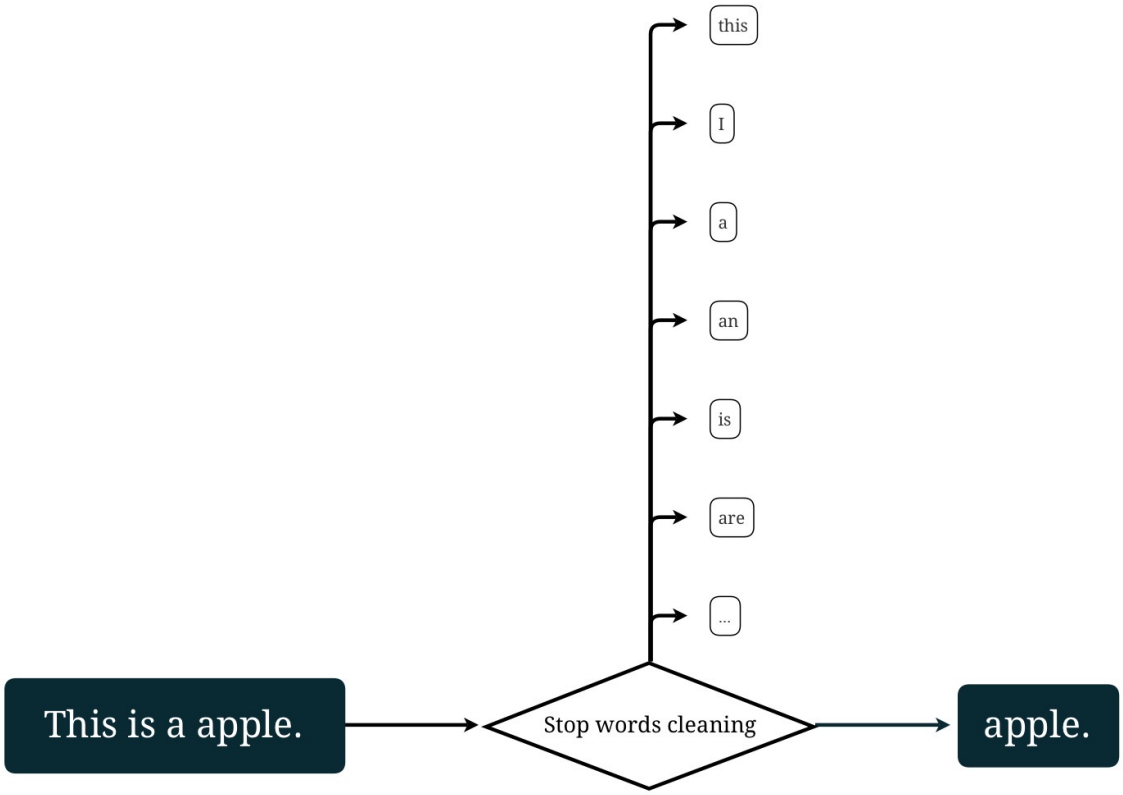
According to the goals we set, we only need the four columns of likes, shares, comments and texts from the fb_scrapped_bbc.csv file. We are going to use python to process the fb_scrapped_bbc.csv file to get these four columns of data and save them to news.csv. Here are some of our results.

	A	B	C	D
1	text	likes	comments	shares
2	This is the moment a chimpanze	63000	3800	13000
3	"Cleaning up after football matches	105000	3900	7300
4	"An incredible human being and revered	6178	2100	5700
5	One of the young filmmakers said they	14395	1600	5400

Data Cleaning

We need to clean the data in the texts column by removing the stop words and some special symbols. We used a stop words list found on github to clean the data. The cleaned data is put into news.csv. Here is stop words list we used and some of the result .





Sort and count the words

First we put news.csv to hadoop.

Then,we used pig to sort the texts column in news.csv by likes,comments, and shares.

Here is the codes for using pig and the results.Of all the sorted results, we only extract the first 50 results from the largest to the smallest.

The original order

```
A = LOAD '/root/news_.csv' USING PigStorage(',') AS
(text:chararray,likes:int,comments:int,shares:int);
DUMP A
B = filter A by $0 is not null;
C = FOREACH B GENERATE REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE
(REPLACE(REPLACE(REPLACE($0,'\\"', ','),' ', ','),'\\_', ','),'\\-', ','),'\\-', ','),'\\?', ','),'\\=', ','),'\\.',','),'.CNBC.COM',','),'\\/',',') as (line);
words = foreach C generate flatten(TOKENIZE(line)) as aword;
grwords = group words by aword;
MapRed = foreach grwords generate group, COUNT(words) AS cnt;
SortMapRed = order MapRed by cnt DESC;
R = LIMIT SortMapRed 50;
STORE R INTO 'all_top50' USING PigStorage(',');
```

```
('world', 936)
('cup', 585)
('uk', 558)
('president', 234)
('wales', 207)
('qatar', 198)
('police', 180)
('time', 162)
('england', 162)
('family', 153)
('covid', 144)
('football', 144)
('ukraine', 135)
('fans', 135)
('protests', 135)
('woman', 126)
('russia', 126)
('killed', 126)
('russian', 126)
('dies', 117)
('twitter', 117)
('china', 117)
('death', 108)
('iran', 108)
('feel', 99)
('city', 99)
('prince', 99)
('set', 99)
('game', 99)
('staff', 99)
('power', 99)
('war', 90)
```

Sort by likes

```
A = LOAD '/root/news_.csv' USING PigStorage(',') AS (text:chararray,likes:int,comments:int,shares:int);
B = filter A by $1 is not null;
E = order B by likes DESC;
C = FOREACH E GENERATE REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE
(REPLACE(REPLACE($0,' ', ','),'\\_', ','),'\\-', ','),'\\-', ','),'\\?', ','),'\\=', ','),'\\.',','),'.CNBC.COM',','),'\\/',',') as (line);
words = foreach C generate flatten(TOKENIZE(line)) as aword;
grwords = group words by aword;
MapRed = foreach grwords generate group, COUNT(words) AS cnt;
SortMapRed = order MapRed by cnt DESC;
R = LIMIT SortMapRed 50;
STORE R INTO 'likes_top50' USING PigStorage(',');
```

```
('world', 210)
('cup', 154)
('win', 42)
('germany', 42)
('time', 35)
('england', 35)
('birth', 28)
('uk', 28)
('shock', 28)
('beat', 28)
('fifa', 28)
('qatar', 28)
('baby', 28)
('japan', 21)
('met', 21)
('opener', 21)
('argentina', 21)
('portugal', 21)
('cristiano', 21)
('ronaldo', 21)
('score', 21)
('alcohol', 21)
('2022', 21)
('team', 21)
```

Sort by comments

```
A = LOAD '/root/news_.csv' USING PigStorage(',') AS (text:chararray,likes:int,comments:int,shares:int);
B = filter A by $2 is not null;
E = order B by comments DESC;
C = FOREACH E GENERATE REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE
(REPLACE(REPLACE($0,' ', ','),'\\_', ','),'\\-', ','),'\\-', ','),'\\?', ','),'\\=', ','),'\\.',','),'.CNBC.COM',','),'\\/',',') as (line);
words = foreach C generate flatten(TOKENIZE(line)) as aword;
grwords = group words by aword;
MapRed = foreach grwords generate group, COUNT(words) AS cnt;
SortMapRed = order MapRed by cnt DESC;
R = LIMIT SortMapRed 50;
STORE R INTO 'comments_top50' USING PigStorage(',');
```

```
('world', 29)
('cup', 22)
('qatar', 8)
('england', 7)
('uk', 6)
('germany', 5)
('win', 5)
('fifa', 4)
('wales', 4)
('armband', 4)
('alcohol', 3)
('stadiums', 3)
('fan', 3)
('told', 3)
('players', 3)
('onelove', 3)
('shock', 3)
('japan', 3)
('fans', 3)
('time', 3)
('wear', 3)
('beat', 3)
('west', 3)
('twitter', 3)
('ambulance', 3)
('tournament', 2)
('birth', 2)
('support', 2)
('rainbow', 2)
```

Sort by shares

```
A = LOAD '/root/news_.csv' USING PigStorage(',') AS (text:chararray,likes:int,comments:int,shares:int);
B = filter A by $3 is not null;
E = order B by shares DESC;
C = FOREACH E GENERATE REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE
(REPLACE(REPLACE($0,' ', ','),'\\_', ','),'\\-', ','),'\\-', ','),'\\?', ','),'\\=', ','),'\\.',','),'.CNBC.COM',',','\\/',',') as (line);
words = foreach C generate flatten(TOKENIZE(line)) as aword;
grwords = group words by aword;
MapRed = foreach grwords generate group, COUNT(words) AS cnt;
SortMapRed = order MapRed by cnt DESC;
R = LIMIT SortMapRed 50;
STORE R INTO 'shares_top50' USING PigStorage(',');
```

```
('world', 105)
('cup', 70)
('win', 30)
('germany', 25)
('uk', 25)
('moment', 20)
('dies', 20)
('family', 20)
('message', 20)
('qatar', 20)
('shock', 20)
('birth', 15)
('japan', 15)
('baby', 15)
('time', 15)
('children', 15)
('argentina', 15)
('club', 15)
('score', 15)
('chimpanzee', 10)
('mum', 10)
('met', 10)
('newborn', 10)
('mahle', 10)
('chimp', 10)
```

Visualization process

In this part, we decide to visualizes the data that has been cleaned to produce a word cloud graph. In this task, we use python for word cloud graph production and the advantages of making word clouds are as follows:

1. Visually more impactful. Word cloud charts are more attractive and visually impactful than bar charts, histograms and word frequency statistics tables, and to some extent cater to people's fast-paced reading habits.
2. More direct in content. The word cloud map itself is a highly condensed and streamlined processing of text content, which can reflect the content of a specific text more intuitively, saving the reader's time to a certain extent and allowing the reader to get a glimpse of the main information of the text data in a short period of time, which is of great use for our subsequent analysis.

Because the data comes from facebook, we designed the word cloud as facebook's logo in order to make the data more visual. We have not removed the BBC from the list of stop words because it is a visual representation of the data coming from the BBC's Facebook account. This is a example.

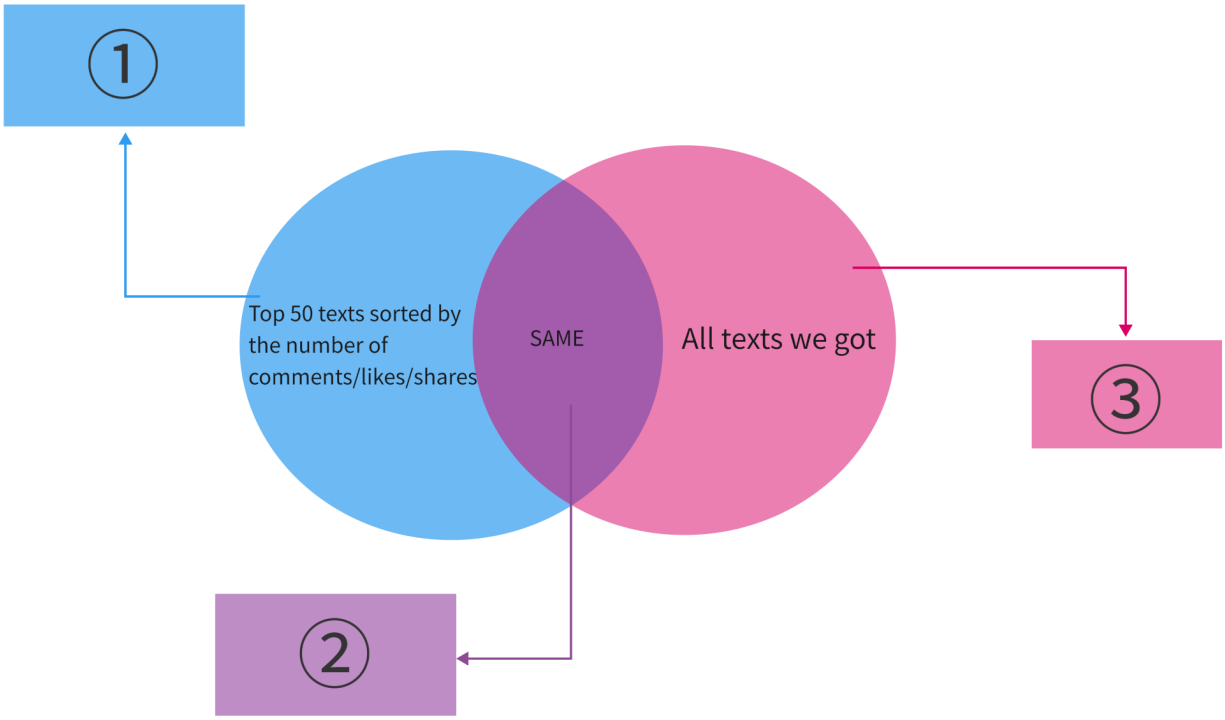
Analysis process

Data comparison

We consider that the obtained text data can be divided into several types:

- Emerging hot spots
- Persistent news
- Outdated news

Overlapping listA (top 50 words of all texts) and listB/C/D (top50 words of texts sorted by likes/comments/shares) , the result obtained is shown in the following figure.



Part ① is the emerging hot spots, part ② is the persistent news, and part ③ is the outdated news.

Analysis of public opinion trends

Three kinds of keywords

As mentioned above, according to the three parts we got, we can take a look at the trending topics and the outdated ones.

- The emerging hot spots

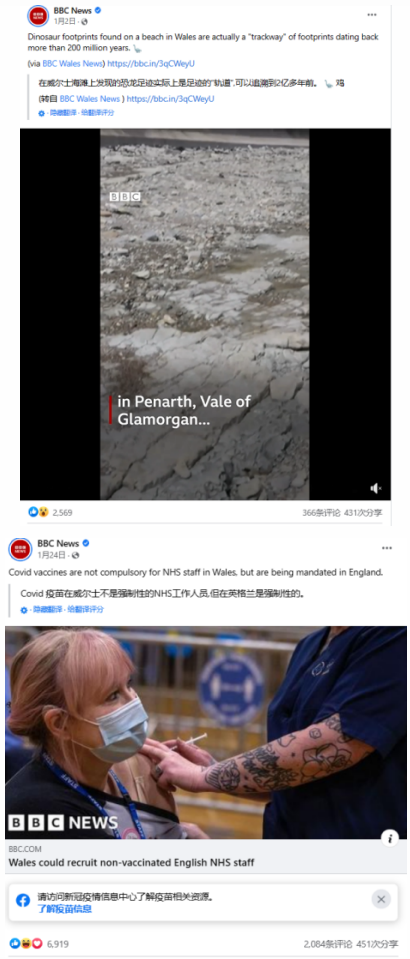
  Based on the list of keywords derived from the data analysis, followed by a keyword search, we can come up with the high-hitting news published by the bbc.

  Some of the partial contents are as follows:

- The persistent news

  Our data analysis summarize the same keywords in the top 50 words in likes, comments, shares and all texts. These keywords represents the persistent news.

  Some of the partial contents are as follows:



- The outdated news

  There's some keywords that only exist when we sort the data by all texts. We regard these keywords as "outdated".

  Some of the partial contents are as follows:

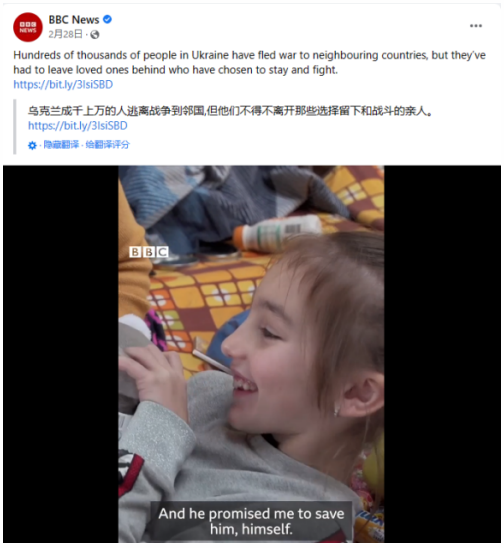
The difference between the concerns of the official and the mass

During the course of the study, there's one thing that caught our focus. When we talk about BBC in the Chinese context, it is often described as 'controversial' and 'inauthentic'. But when we take a international view of what the BBC publishes, it is still highly authoritative and referable in the international context. The discussion, based on different positions and contexts, is always reminding us to find a better way to analyse the BBC's news' stand and authenticity.

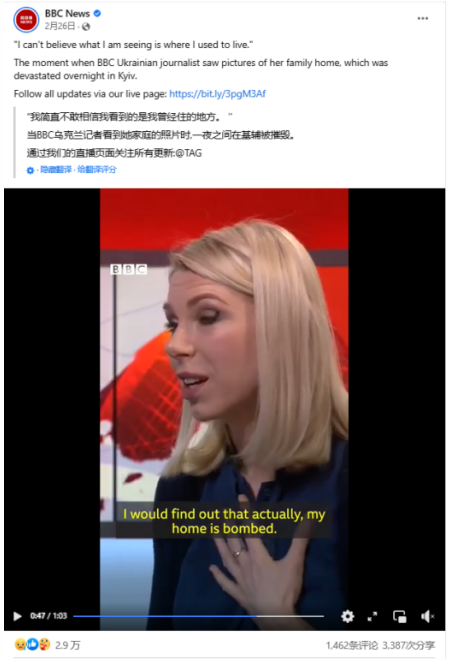
So, what follows is a summary and analysis of our information on the same topic as discussed in the national press and BBC.

Topic 1: The Russia-Ukraine war

- #Putin says Ukraine side has deviated from consensus back to dead-end status# ——CCTV News, China



- BBC



- “China's position on the crisis in Ukraine has always been clear. We call on the parties concerned to achieve a ceasefire and stop the war through dialogue and negotiation, and to find a solution that takes into account the reasonable security concerns of all parties as soon as possible, and we hope that the international community will create conditions and space for this purpose.” —— People's Daily, China

- “Since the outbreak of the Russia-Ukraine conflict, the US has spread a series of lies and used the Ukraine crisis to smear China”“Who is America’s enemy in earth?” ——People's Daily, China

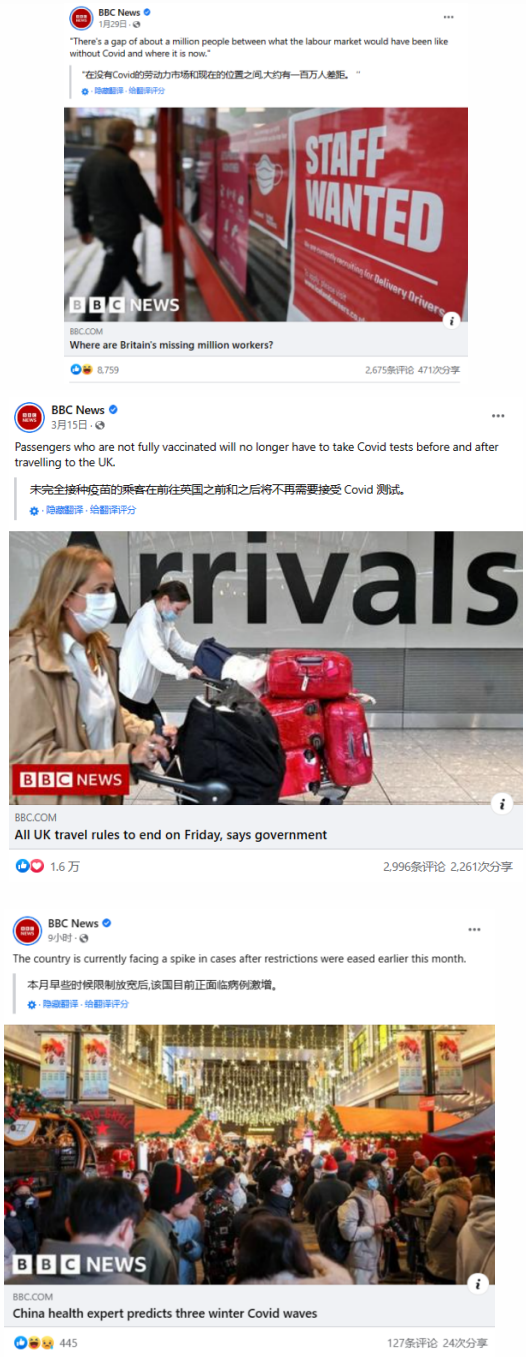
As we can see from the comparison, BBC, representing the Western media, is biased in favour of the Ukrainian side in the Russian-Ukrainian war, while the Chinese media trying to keep a neutral stand and promotes peaceful negotiations, which is in line with the international image of the Chinese government. At the same time, the Chinese media will clearly express their doubts and objections to the US side, which is to a certain extent in line with the international situation.

Topic 2: The covid-19 pandemic

- "Since the outbreak of Covid-19, we have insisted on putting people first and life first, on preventing external imports and internal rebound, on dynamic clearance, and on constantly adjusting prevention and control measures according to the time and situation. We now have achieved significant strategic results in the prevention and control of the epidemic." ——Xinhua News Agency, China

- # Zhao Lijian responds to US epidemic death toll reaching 1 million #. ——People's Daily, China

- BBC



In the case of Covid-19, until the recent relaxation of the policy on the pandemic, the Chinese media continued to promote a 'dynamic zero' policy on the outbreak. At the same time, the Chinese side has always strongly condemned the indifferent attitude of Western countries (especially the US) towards the outbreak, as represented by the Chinese spokesperson's speech towards the deaths in the US. In contrast to the Chinese media, the BBC, representing the Western media, has long since reduced the frequency of its coverage of Covid-19, which means that the West has lowered the priority of the epidemic. Following the recent relaxation of China's policy on the pandemic, the BBC's coverage of Covid-19 has focused on its impact on China.

This contrast reflects the different policies of China and the West regarding the epidemic. At the same time, it also reflects the international tensions that have kept the Chinese media and the Western media in an "undercurrent" of the conflict and political tension. So as all of the topics mentioned above.