Data analytics for The Observatory Review 《天文臺》

Introduction:

Since The Observatory Review is a newspaper containing lots of news about politics, famous people, and a unique analysis of the important events that happened from 1950 to 1985, research and analysis of the article and content of this newspaper can show us well the social situation at that time. To achieve that, we would like to use some modern methods to analyze the newspaper and get a basic knowledge of what was happening during that time. What we plan to do first is to use some OCR engines which could provide us with more accurate text files. After getting all the content in the TXT files, we would first go through the whole context to do some statistical analysis of the words. Then we would use ML and NLP algorithms to categorize the whole text, find out what is the focus of the newspapers each year and draw the heatmap. Finally, we are also curious about the relationship between events and persons during different periods.

Target:

1. Use the OCR engine to make more accurate text recognition for observatories.
2. Word cloud and frequency during the whole time.
3. Word cloud and frequency of each year.
4. Generate different topics of the whole paper and create a dictionary for each topic.
5. Word cloud and frequency for each topic cluster and find the focused topics of each year.
6. Analysis of the relationship and drawing relation diagram for events and persons in the newspaper.
7. Visualize all the outcome products.

Method:

1. Google vision OCR to recognize the newspaper.
2. Bertopic model for topics generation.
3. CkipTagger to do word segmentation and recognize the word role (noun, verb, adjective….).
4. WordCloud package in python to generate the word cloud.
5. Pandas and networkx package combine with CSV to draw the relational diagram.

Expected output:

1. More accurate OCR result in txt
2. Original data of all analyses.
3. WordCloud of each topic of each year.
4. Visualized results, which are aimed to contain the bar chart and leaderboard of the frequency of words and topics.
5. Heatmaps of the topics in different years.
6. Relational diagram showing the relationship between different events and persons.

Timeline:

October: Finish all the OCR jobs to turn the image into TXT files.

November: Tokenize all the words in the TXT file and extract the word existing most frequently every year and finish creating the word cloud.

December: Topic modeling, creating different categories, and relationship analysis.

January: Design the poster and construct the website.