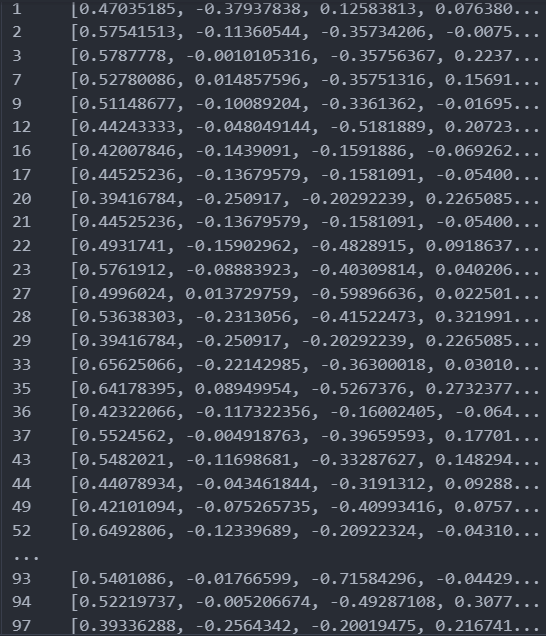
# Report: Analysis of News Data and A-Share Companies

Tianyi WANG, 21035840

In this project, we aim to analyze a dataset of news articles and identify which ones are relevant to a pre-defined list of A-share companies in China. We will be focusing on financial text analysis, knowledge graph construction, and knowledge-driven decision-making.

# Task 1

## Question 1

**1. Bert-Base-Chinese**  
First, I used a relatively traditional and simple Bert model as the baseline, which is Bert-Base-Chinese. It fits News Content with A-share listings by projecting the text into a vector space and then using cosine similarity to quantify the semantic similarity between them. The main issue with this approach is that the similarity cannot be learned but must be manually adjusted. Additionally, this method generates excessively large Bert vectors, around 13GB, resulting in prolonged computation times. Therefore, I only ran this method on a small sample.

This similarity is adjusted manually. So I chose a number based on my own fine-tuning, I set this parameter to a specific value – 0.82.

The selection of similarity can also be accomplished through the Similarity module in the text2vec package. However, due to its long running time, it takes approximately three hours to process every ten pieces of data. Therefore, this method has been abandoned.

**The Bert vectors transformed from the first 100 news contents after filtering**

**2. BERT-Base-Chinese-Finetuning-Financial-News-Sentiment-V2**

Second, I used an updated and specially fine-tuned Bert model, namely: hw2942/BERT-Base-Chinese-Finetuning-Financial-News-Sentiment-V2. Its main process involves using a Tokenizer to convert segmented text into a format that the model can understand, and then processing the text through its multi-layer Transformer architecture to capture the bidirectional relationships between words.

I chose this model as an improvement strategy because it is optimized for a specific domain (i.e., financial news), which means it requires fewer computational resources for related tasks. This allows it to process and analyze text more quickly. Additionally, it can converge to optimal performance at a faster rate, significantly reducing runtime, especially when dealing with 1 million data entries.

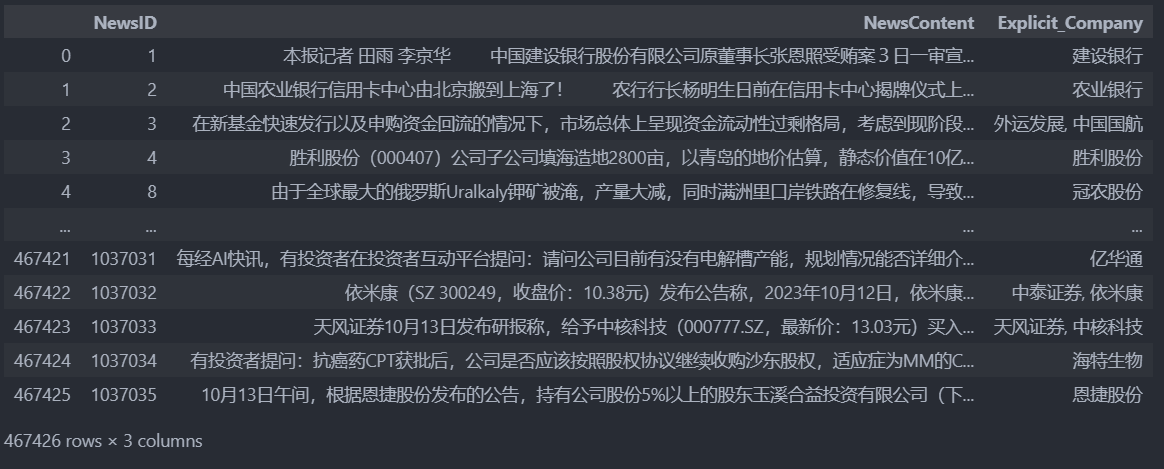
I did the Sentiment Polarity Analysis at the same time with this model.



**BERT-Base-Chinese-Finetuning-Financial-News-Sentiment-V2 Model Running Results (Colab)**

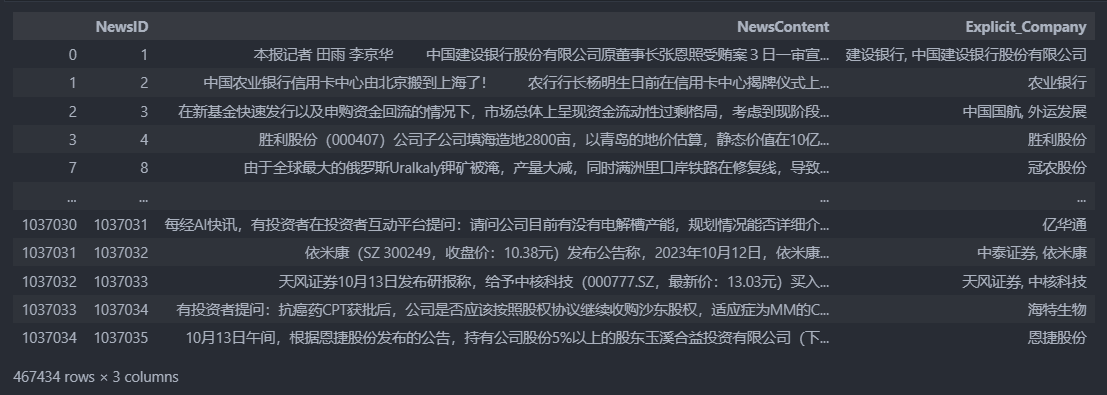
**3. Directly Matching (Brute Force)**

Third, I chose directly matching. This approach does not involve complex algorithms or intermediate steps. In directly matching, elements or data points are typically examined one by one to confirm their consistency or differences. Surprisingly, the direct matching method performed well in this task. I suspect this is probably due to the clear and standardized names of A-share companies, which made the straightforward direct matching method efficient. Additionally, the GPU resources used were average and could not fully meet the requirements of the Bert model.



**The Cleaning Results Obtained Through Directly Matching**

During this process, I also discovered that some bankrupt companies in the A-share list have prefixes like ST, SST, \*ST or S\*ST in their names, like ST南华, ST国重装, \*ST石化A and so on, which can affect direct matching. Therefore, I made certain adjustments to the company names in the A-share list in the improved version.



**The Cleaning Results Obtained Through Directly Matching After Second Checking**

## Question 2

Based on Q1's hw2942/bert-base-chinese-finetuning-financial-news-sentiment-v2, I defined a sentiment analysis function, which classified the sentiment of the text as positive or negative based on the probabilities generated by the model. This function was applied to a DataFrame containing financial news content. The analysis process was timed, and a snapshot of the results was printed to verify accuracy.

As for data labeling, I used the model from Hugging Face(https://huggingface.co/Raychanan/bert-base-chinese-FineTuned-Binary-Best) to label Positive as 1 and Negative as 0, while some data that could not be determined were marked as Neutral. During the processing, news associated with the Neutral label were removed.

I also tried another method for sentiment label classification of data. I performed sentiment analysis on neutral data and distributed it proportionally to the positive and negative labels to achieve a binary classification approach.



**The Second Approach of Sentiment Polarity Analysis**

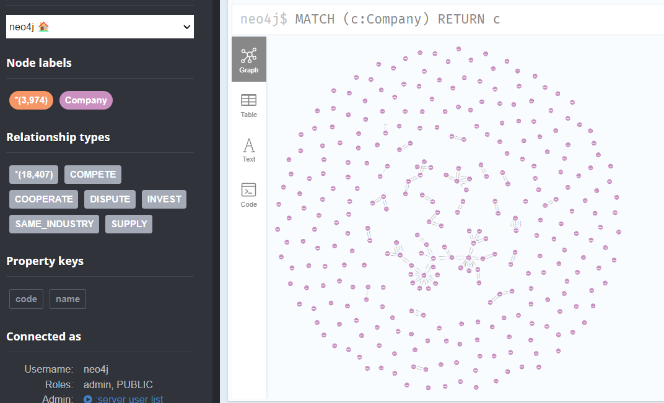
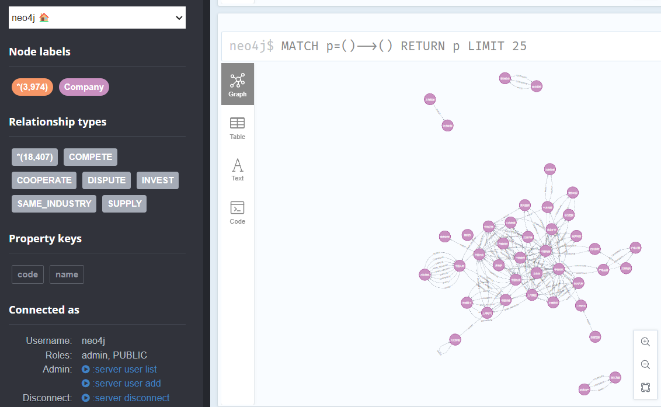
# Task 2

## Question 3

In the Neo4j graph database, each company was represented as a node with properties such as name and code. The script iterated through each row in the company data and creates these nodes using the MERGE command, ensuring no duplicates are created.

The script then processed different types of relationships between companies. For unidirectional relationships like 'invest' and 'supply', I extracted the start and end IDs, converted them into stock codes, and created directed edges (relationships) in the graph accordingly.

It is worth noting that I represented the bidirectional relationships using two unidirectional relationships combined. This was a workaround found due to persistent errors encountered when initially using bidirectional relationships.

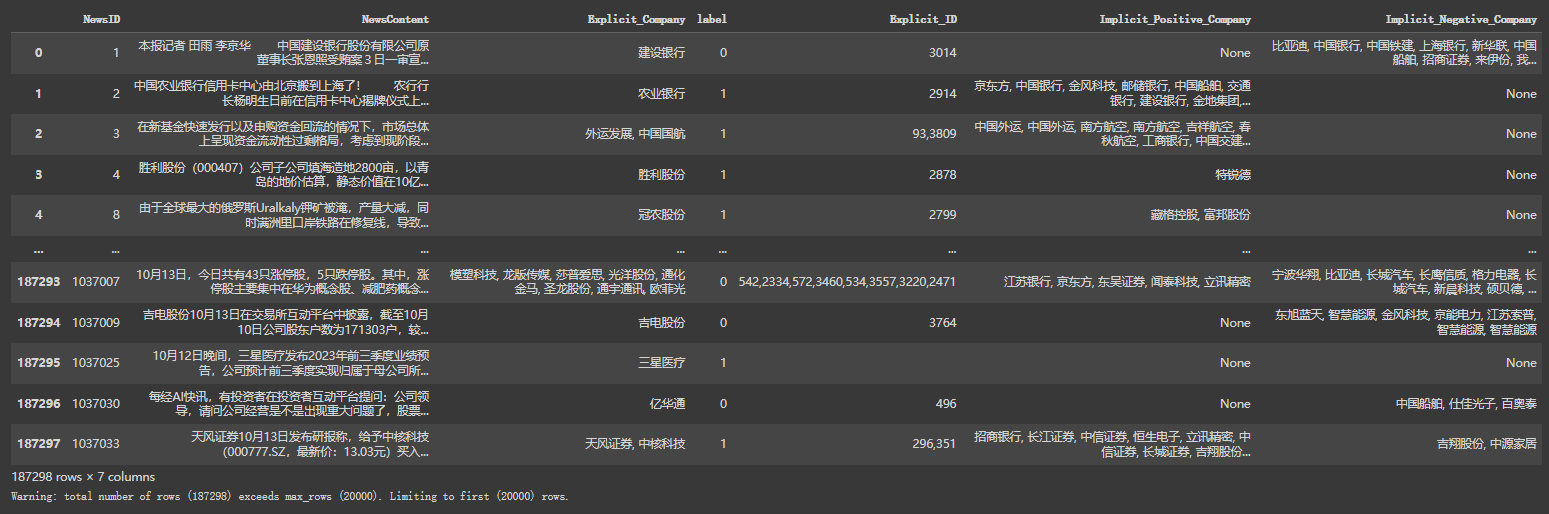
****

**"Supply" and "Invest" are unidirectional, while all the others are bidirectional.**

## Question 4

This analysis method efficiently processes and integrates data from files to map out company relationships. Using pandas, it combines these files into a unified data structure. Key to this process is categorizing relationships as positive 1 for cooperative interactions and negative 0 for competitive ones, with extraneous data like time information removed for clarity.

A significant part of the method involves mapping company names to their unique IDs, facilitating standardized data analysis. Further depth is added by incorporating sentiment analysis, converting company names to IDs within this context.

In summary, this method merges various data sources and sentiment analysis to reveal intricate business relationship dynamics. The outcome is a structured and informative dataset that elucidates the complex web of company interactions.

**The Judging Results of Knowledge-Driven Financial Analysis (Colab)**

It is noteworthy that this dataset contains only 187,298 entries. This is because the method of removing neutral news was adopted in the previous steps. If we were to apply the proposed method of matching neutral news based on their degree of positivity or negativity, we would obtain a richer dataset. However, due to time constraints in processing, only this approach is demonstrated here.

Reference (Task1 Question1): https://github.com/dongrixinyu/JioNLP/wiki/%E6%AD%A3%E5%88%99%E6%8A%BD%E5%8F%96%E4%B8%8E%E8%A7%A3%E6%9E%90-%E8%AF%B4%E6%98%8E%E6%96%87%E6%A1%A3#user-content-%E5%88%A0%E9%99%A4%E6%96%87%E6%9C%AC%E4%B8%AD%E7%9A%84%E5%BC%82%E5%B8%B8%E5%AD%97%E7%AC%A6

https://github.com/fighting41love/funNLP

https://huggingface.co/Raychanan/bert-base-chinese-FineTuned-Binary-Best