﻿**DSAA 5002 - Data Mining and Knowledge Discovery in Data Science**

**Project**

**50015940 Jiaxiang Gao**

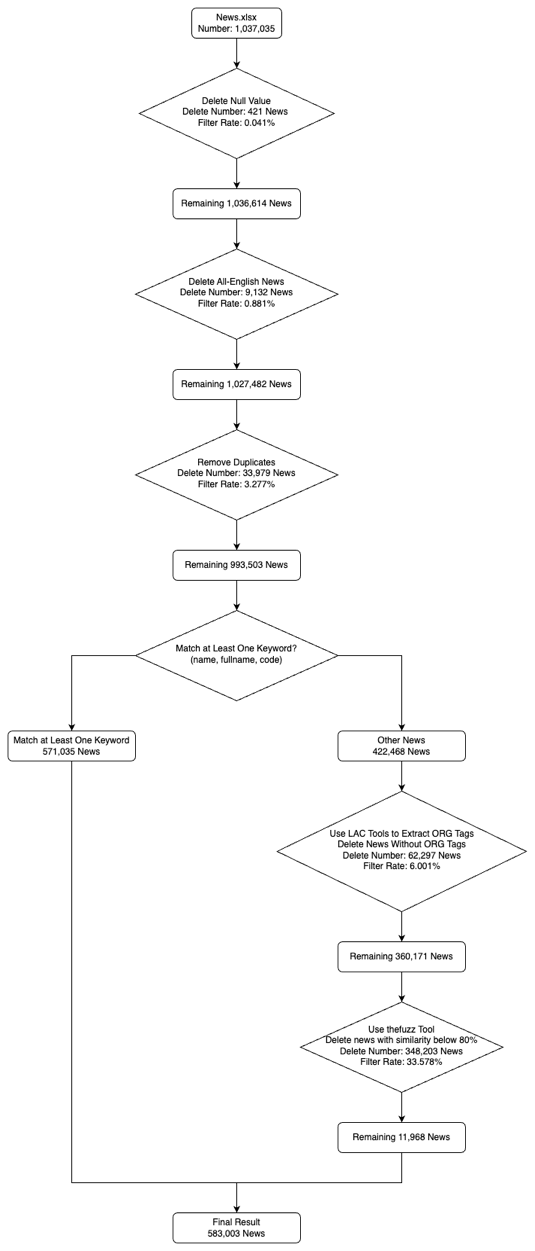
The below report briefly describes my process. For a detailed view of the process, you can check the Code section, which clearly introduces each step using markdown.

All the code for this project can be found in this GitHub repository:

[Github: HKUSTGZ\_DSAA5002\_project\_Financial\_Text\_Analysis\_Knowledge\_Graph](https://github.com/GDistories/HKUSTGZ_DSAA5002_project_Financial_Text_Analysis_Knowledge_Graph)

Additionally, due to file upload limitations, you can find the complete project, including files saved during the process, at this OneDrive link:

[OneDrive: HKUSTGZ\_DSAA5002\_project\_Financial\_Text\_Analysis\_Knowledge\_Graph](https://hkustgz-my.sharepoint.com/:f:/g/personal/jgao329_connect_hkust-gz_edu_cn/Ek4_vx1VgO1GmE5Drt_jvHkB6IkSW5yClHpiLSi0zldpjA?e=a2Izmt)

**Q1: Data Preprocessing - Noise Removal**

Here's a specific description of my data processing for Q1:

Original news data: 1,037,035.

1. Deleted the NewsContent column.
2. Removed null values.

* Data deleted in this process: 421
* Remaining data: 1,036,614
* Filter rate: 0.041%

1. Removed non-Chinese news.
   * Data deleted in this process: 9,132
   * Remaining data: 1,027,482
   * Filter rate: 0.881%
2. Removed special characters and numbers from the news.
3. Deduplicated the news.

* Data deleted in this process: 33,979
* Remaining data: 993,503
* Filter rate: 3.277%

1. Used regular expressions to remove certain special characters from the A-share list's name and divided the dataset into complete matches and other news based on whether at least one of the name, fullname, and code appeared in the news. Special characters: r'\*?ST', r'^PT', r'^S', r'B股$', r'B$', r'A$'

* Complete match news count: 571,035
* Other news count: 422,468

1. Used Baidu's LAC library to further process other news, extracting the ORG tag and deleting news that did not extract the ORG tag. [1][2]
   * + Data deleted in this process: 62,297.
     + Remaining data: 360,171
     + Filter rate: 6.001%
2. Used the thefuzz library to calculate text similarity between the extracted ORG tag and the name, deleting news with similarity below 80%. [3]

* Data deleted in this process: 348,203.
* Remaining data: 11,968
* Filter rate: 33.578%

1. Merged the news obtained with the complete match news from step 7 to get the result.

* Final data count: 583,003

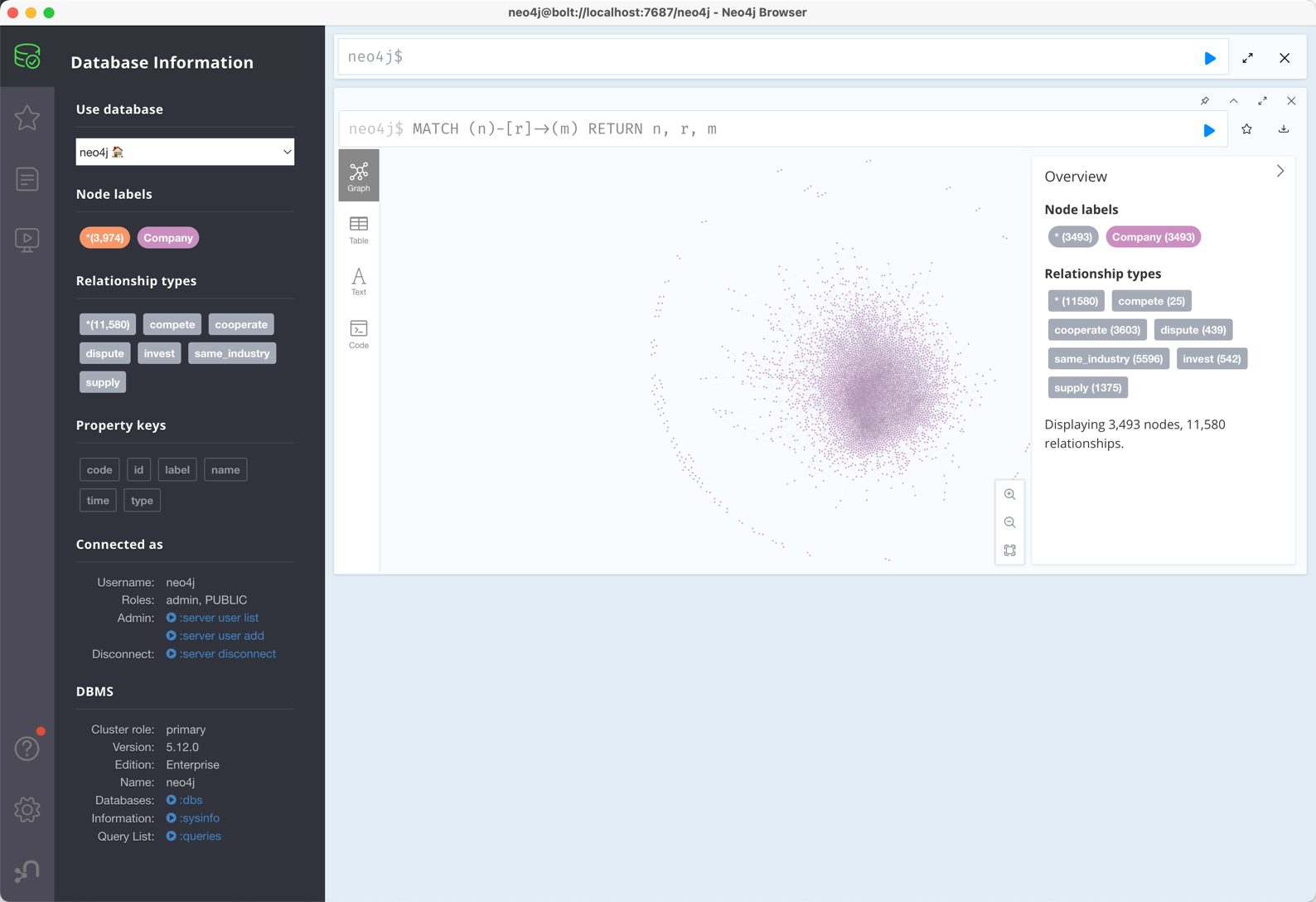
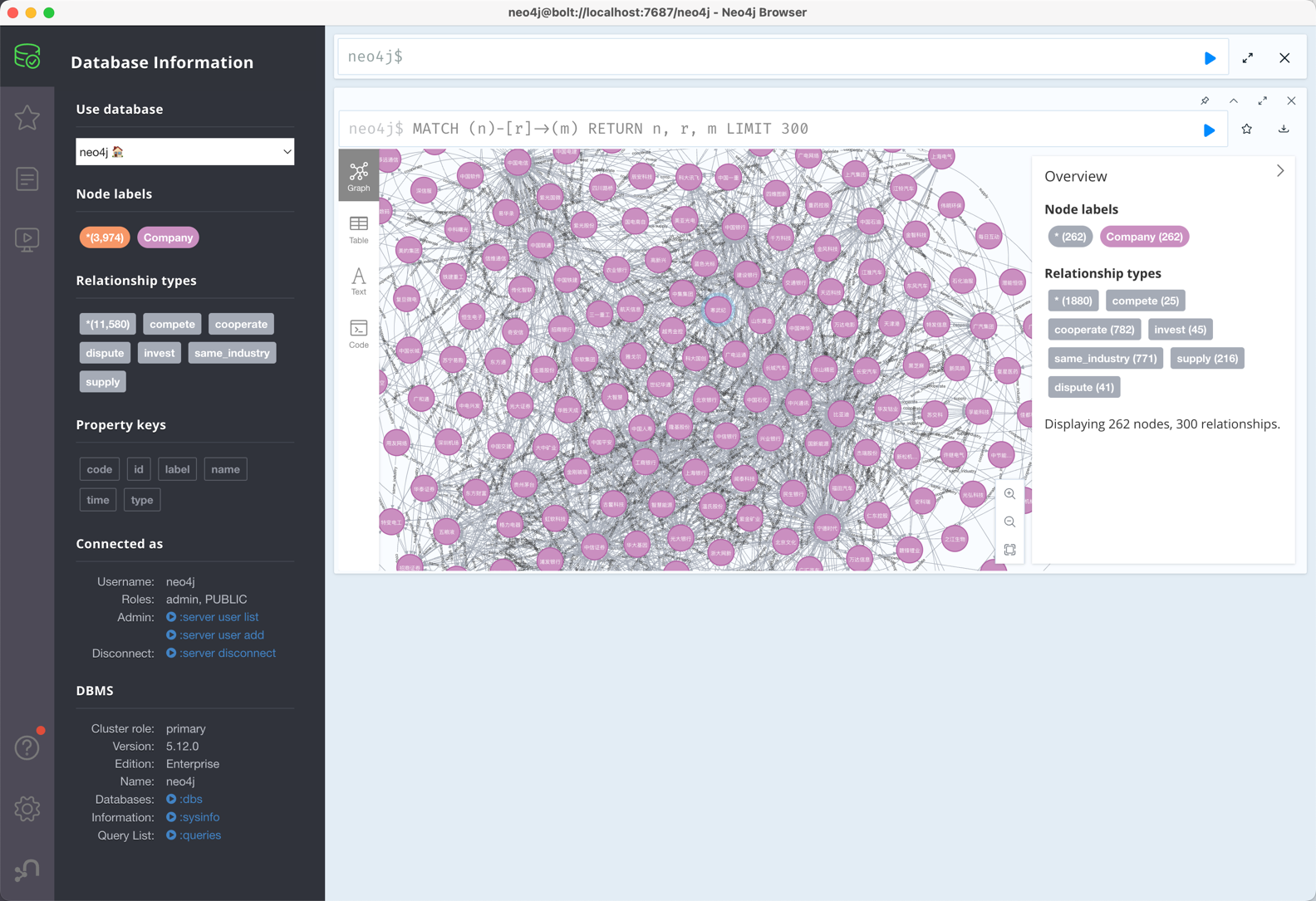
**Q2: ﻿Data Analysis - Text Knowledge Mining**

In this question, I used the IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment model [4] to conduct sentiment analysis on the obtained news. Due to computational resource limitations, I divided the dataset into 12 parts, calculated the sentiment in parallel, and then merged them into the result.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | ASAP-SENT | ASAP-ASPECT | ChnSentiCorp |
| Erlangshen-Roberta-110M-Sentiment | 97.77 | 97.31 | 96.61 |
| Erlangshen-Roberta-330M-Sentiment | 97.9 | 97.51 | 96.66 |
| Erlangshen-MegatronBert-1.3B-Sentiment | 98.1 | 97.8 | 97 |

According to the official documentation, the 110M Erlangshen-Roberta-110M-Sentiment model offers a good balance between performance and resource consumption. While it has fewer parameters compared to larger 330M and 1.3B models, this 110M model demonstrates quite high accuracy in terms of performance. Furthermore, the smaller model size translates to lower operational costs and faster processing speeds, which can alleviate resource constraints for me.

**Q3: ﻿Constructing a Knowledge Graph**



**References:**

[1] Baidu. (2021). *Baidu/Lac Baidu NLP: Word segmentation, part of speech tagging, named entity recognition, word importance*. GitHub. https://github.com/baidu/lac

[2] Jiao, Z., Sun, S., & Sun, K. (2018). C*hinese Lexical Analysis with Deep Bi-GRU-CRF Network. arXiv preprint arXiv:1807.01882*. <https://arxiv.org/abs/1807.01882>

[3] Seatgeek. (2021). *Seatgeek/thefuzz: Fuzzy string matching in Python*. GitHub. https://github.com/seatgeek/thefuzz

[4] Idea-CCNL. (2022). *Idea-CCNL/erlangshen-roberta-110m-sentiment · hugging face*. IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment · Hugging Face. https://huggingface.co/IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment