Report

Xing Daiyan

Task1Q1

In order to find the news we want, we can firstly utilize LAC (Lexical Analysis of Chinese) for named entity recognition to uncover the mentioned organization names in each news article like Figure 1. This will reduce our subsequent processing time while also improving accuracy.

dfnew

	NewsID	Title	NewsContent	NewsSource	NamedEntityRecognition
0	1	建设银行原董事长张恩照一审被判15年	本报记者 田雨 李宗华 中国建设银行股份有限公司原董事长张恩照受贿案3日一审宣判，北京市第...	中国证券报	['中国建设银行, 北京市第一中级人民法院, 中国建设银行股份有限公司']
1	2	农行信用卡中心搬到上海踢	中国农业银行信用卡中心由北京搬到上海了！农行行长杨明生日前在信用卡中心揭牌仪式上表示，此...	人民日报	['农行, 农行信用卡中心, 农行金穗资讯卡, 中国农业银行信用卡中心']
2	3	外运发展：价值型盈筹股补涨要求强烈	在新基金快速发行以及申购资金回流的情况下，市场总体上呈现资金流动性过剩格局，考虑到现阶段投...	杭州新希望	['大韩航空, 中国国航, 新韩投资, 中外运致豪, 四川航空集团有限责任公司, 四川航空集团公司']
3	4	胜利股份：稳步走强形成标准上升通道	胜利股份（000407）公司子公司缅海造地2800亩，以青岛的地价估算，静态价值在10亿元...	源达投资	['深圳前山集团, 新加坡港务集团, 青岛胜通海岸营业发展有限公司, 胜利股份, 山东胜利钢铁...'
4	5	【插股快讯】恒指收市报18960点 成交467亿港元	全景网11月30日讯 外围股市活跃，带动港股今早强好，恒指高开后反覆上升，最高升252点，...	全景网	['中国财险, 工行, 中行, 建行, 交行, 中移动, 恒地, 恒隆地产, 招行']
...	...	...	...	...	...
1037030	1037031	亿华通：公司电解槽相关产品目前还处于产品的研发及测试阶段 尚未实现批量销售	每经AI快讯，有投资者在投资者互动平台提问：请问公司目前有没有电解槽产能，规划情况能否详细介...	每日经济新闻	['亿华通']
1037031	1037032	依米康：接受中泰证券调研	依米康（SZ 300249，收盘价：10.38元）发布公告称，2023年10月12日，依米康...	每日经济新闻	['中泰证券']
1037032	1037033	天风证券给予中核科技买入评级 核电行业景气上行 公司有望乘风而起	天风证券10月13日发布研报称，给予中核科技（000777.SZ，最新价：13.03元）买入...	每日经济新闻	['中核科技, 天风证券']
1037033	1037034	海特生物：公司在抗癌药CPT获批后 会考虑适时开展CPT在海外的临床开展求上市	有投资者提问：抗癌药CPT获批后，公司是否应该按照股权协议继续收购沙东股权，适应症为MM的C...	界面新闻	['海特生物']
1037034	1037035	恩捷股份：股东合益投资部分股份补充质押	10月13日午间，根据恩捷股份发布的公告，持有公司股份5%以上的股东玉溪合益投资有限公司（下...	证券日报	['000股公司股份, 恩捷股份, 玉溪合益投资有限公司']

1037035 rows × 5 columns

Figure 1

And we have observed that in the 'A_share_list', there are company names and their corresponding fullnames. We will combine these two lists to create a new name list called 'a_share_keywords'.

① Rule-based strategy

For each row in the 'dfnew' dataframe, if any entity in the 'NamedEntityRecognition' column appears in the 'a_share_keywords' list, the row data will be kept. Otherwise, it will be filtered out.

pd.read_excel('/content/drive/MyDrive/data/df_filtered_rule.xlsx')

	NewsID	Title	NewsContent	NewsSource	NamedEntityRecognition	Explicit_Company
0	1	建设银行原董事长张恩照一审被判15年	本报记者 田雨 李宗华 中国建设银行股份有限公司原董事长张恩照受贿案3日一审宣判，北京市第...	中国证券报	['中国建设银行', '北京市第一中级人民法院', '中国建设银行股份有限公司']	中国建设银行股份有限公司
1	3	外运发展：价值型盈筹股补涨要求强烈	在新基金快速发行以及申购资金回流的情况下，市场总体上呈现资金流动性过剩格局，考虑到现阶段投...	杭州新希望	['大韩航空', '中国国航', '新韩投资', '中外运致豪', '四川航空集团有限责任公...	中国国航
2	4	胜利股份：稳步走强形成标准上升通道	胜利股份（000407）公司子公司缅海造地2800亩，以青岛的地价估算，静态价值在10亿元...	源达投资	['深圳前山集团', '新加坡港务集团', '青岛胜通海岸营业发展有限公司', '胜利股份'...	胜利股份
3	8	南风化工：钾肥三雄之一 被"中化"相中	由于全球最大的俄罗斯Uralkali的钾矿被淹，产量大减，同时属内陆口岸铁路在修复线，导致中...	银河证券	['山西运城盐湖', '运城盐湖', '中化集团', '山西钾肥有限公司', '运城盐湖资源'...	冠农股份
4	10	招商轮船上市定位清析 运行区间5-6元	招商局能源运输股份有限公司A股于12月1日上市流通，证券简称为“招商轮船”，证券代码601...	金通证券	['招商局能源运输股份有限公司', '中远集团', '中石化集团', '中金公司', '中化...	招商局能源运输股份有限公司,中金公司,招商轮船
...	...	...	...	...	...	...
334483	1037031	亿华通：公司电解槽相关产品目前还处于产品的研发及测试阶段 尚未实现批量销售	每经AI快讯，有投资者在投资者互动平台提问：请问公司目前有没有电解槽产能，规划情况能否详细介...	每日经济新闻	['亿华通']	亿华通
334484	1037032	依米康：接受中泰证券调研	依米康（SZ 300249，收盘价：10.38元）发布公告称，2023年10月12日，依米康...	每日经济新闻	['中泰证券']	中泰证券
334485	1037033	天风证券给予中核科技买入评级 核电行业景气上行 公司有望乘风而起	天风证券10月13日发布研报称，给予中核科技（000777.SZ，最新价：13.03元）买入...	每日经济新闻	['中核科技', '天风证券']	中核科技,天风证券
334486	1037034	海特生物：公司在抗癌药CPT获批后 会考虑适时开展CPT在海外的临床开展求上市	有投资者提问：抗癌药CPT获批后，公司是否应该按照股权协议继续收购沙东股权，适应症为MM的C...	界面新闻	['海特生物']	海特生物
334487	1037035	恩捷股份：股东合益投资部分股份补充质押	10月13日午间，根据恩捷股份发布的公告，持有公司股份5%以上的股东玉溪合益投资有限公司（下...	证券日报	['000股公司股份', '恩捷股份', '玉溪合益投资有限公司']	恩捷股份

334488 rows × 6 columns

Figure 2

$$filter\ rate = \frac{1037035 - 334488}{1037035} = 0.677457$$

② Similarity-based strategy

The similarity-based strategy is an improvement over the rule-based strategy because it allows for similarity calculations to be performed between entities and company names. In this case, we are using "company names" to refer to both full company names and their abbreviations. This is because abbreviations of company names are often similar to their corresponding full names. By employing similarity calculations, we can effectively handle cases where entities and company names do not match exactly.

For each row in the 'dfnew' dataframe, if any entity in the 'NamedEntityRecognition' column appears in the 'a_share_keywords' list, the entity will be added the ‘Explicit_Company’. If it doesn't appear, then the similarity between that entity and each company name in the 'company_names' list will be compared.

We define a function called ‘calculate_similarity’ to calculate the similarity scores between an entity and a set of company names. It utilizes the TF-IDF vectorization approach and cosine similarity to measure the similarity between texts：

- The TfidfVectorizer object is created with the configuration to consider n-grams of 1-2 characters. The entity and company names are combined into a single list called corpus to construct the text corpus. The ‘fit_transform’ method is then used to convert the text corpus into a TF-IDF matrix, where each element represents the TF-IDF weight of each term in the corresponding document.
- Next, the ‘cosine_similarity’ function is applied to calculate the cosine similarity between the entity and each company name. It compares the TF-IDF vector of the entity with the TF-IDF matrix of each company name, resulting in a similarity matrix.
- Finally, the function returns the first row of the similarity matrix, which represents the similarity scores between the entity and each company name. This indicates the degree of similarity between the entity and each company name based on the TF-IDF weights.

We set the similarity threshold to 0.7(this may filter out a lot of news) and add the ‘Explicit_Company’ with the highest similarity. When we have finished iterating over each entity in every row and removed duplicate occurrences of 'Explicit_Company' in each row, it means that the filtering process is complete like Figure 3

df_filtered_sim

	NewsID	Title	NewsContent	NewsSource	NamedEntityRecognition	Explicit_Company
0	1	建设银行原董事长张恩照一审被判15年	本报记者 田雨 李京华 中国建设银行股份有限公司原董事长张恩照受贿案3日一审宣判，北京市第...	中国证券报	['中国建设银行', '北京市第一中级人民法院', '中国建设银行股份有限公司']	建设银行
1	2	农行信用卡中心搬到上海啦	中国农业银行信用卡中心由北京搬到上海了！农行行长杨明生日前在信用卡中心揭牌仪式上表示，此...	人民日报	['农行', '农行信用卡中心', '农行金穗资记卡', '中国农业银行信用卡中心']	农业银行
2	3	外运发展：价值型监筹股补涨要求强烈	在新基金快速发行以及申购资金回流的情况下，市场总体上呈现资金流动性过剩格局，考虑到现阶段权...	杭州新希望	['大韩航空', '中国国航', '新韩投资', '中外运敦豪', '四川航空集团有限责任公...	中国国航
3	4	胜利股份：稳步走强形成标准上升通道	胜利股份（000407）公司子公司现海监地2800亩，以青岛的地价估算，静态价值在10亿元...	源达投资	['深圳市山集团', '新加坡港务集团', '青岛胜通海岸置业发展有限公司', '胜利股份'...	胜利股份
4	8	南风化工：钾肥三雄之一 被"中化"相中	由于全球最大的俄罗斯Uralkaly钾矿被淹，产量大减，同时湄洲里口岸铁路在修复线，导致中...	银河证券	['山西运城盐湖', '运城盐湖', '中化集团', '山西钾肥有限公司', '运城国资委'...	冠农股份,雅化集团
...	...	...	...	...	...	...
396376	1037031	亿华通：公司电解槽相关产品目前还处于产品的研发及测试阶段 尚未实现批量销售	每经AI快讯，有投资者在投资者互动平台提问：请问公司目前有没有电解槽产能，现划情况能否详细介...	每日经济新闻	['亿华通']	亿华通
396377	1037032	依米康：接受中泰证券调研	依米康（SZ 300249，收盘价：10.38元）发布公告称，2023年10月12日，依米康...	每日经济新闻	['中泰证券']	中泰证券
396378	1037033	天风证券给予中核科技买入评级 核电行业量气上行 公司有望乘风而起	天风证券10月13日发布研报称，给予中核科技（000777.SZ，最新价：13.03元）买入...	每日经济新闻	['中核科技', '天风证券']	中核科技,天风证券
396379	1037034	海特生物：公司在抗癌药CPT获批后 会考虑适时开展CPT在海外的临床开课水上市	有投资者提问：抗癌药CPT获批后，公司是否应该按照股权协议继续收购少东股权，适应层为MM的C...	界面新闻	['海特生物']	海特生物
396380	1037035	恩捷股份：股东合益投资部分股份补充质押	10月13日午间，根据恩捷股份发布的公告，持有公司股份5%以上的股东玉溪合益投资有限公司（下...	证券日报	['000股公司股份', '恩捷股份', '玉溪合益投资有限公司']	恩捷股份

396381 rows × 6 columns

Figure 3

$$filter\ rate = \frac{1037035 - 396381}{1037035} = 0.617775$$

Also, I have found that using ‘wiki.zh.text.model’ for similarity calculation is much more accurate and may be possible to preserve more news. (https://github.com/AimeeLee77/wiki_zh_word2vec)

③ **Elasticsearch**

Elasticsearch is a popular open-source search and analytics engine. It is built on top of the Apache Lucene library and provides distributed, real-time search and analytics capabilities. Elasticsearch is designed to handle large volumes of data and is commonly used for various use cases such as full-text search, log analytics, and data visualization.

Due to its distributed architecture, Elasticsearch allows for parallel processing of data across multiple nodes, effectively utilizing resources and improving throughput, thereby significantly reducing runtime and enhancing processing efficiency when dealing with large datasets.

Task1Q2

① **SnowNLP**

We can create a SnowNLP object named s using the SnowNLP library used for sentiment analysis of the text. And then use s.sentiments method is called to obtain the sentiment score of the text. The sentiment score ranges from 0 to 1, where a score closer to 1 indicates positive sentiment, and a score closer to 0 indicates negative sentiment. A conditional statement is used to check if the value of s.sentiments is greater than 0.5. If the sentiment score is greater than 0.5, the sentiment variable is assigned a value of 1, indicating positive sentiment. Otherwise, it is assigned a value of 0, indicating negative sentiment.

	NewsID	NewsContent	Explicit_Company	label
0	1	本报记者 田雨 李京华 中国建设银行股份有限公司原董事长张恩照受贿案3日一审宣判，北京市第...	建设银行	0
1	2	中国农业银行信用卡中心由北京搬到上海了！ 农行行长杨明生日前在信用卡中心揭牌仪式上表示，此...	农业银行	1
2	3	在新基金快速发行以及申购资金回流的情况下，市场总体上呈现资金流动性过剩格局，考虑到现阶段权...	中国国航	1
3	4	胜利股份（000407）公司子公司填海造地2800亩，以青岛的地价估算，静态价值在10亿元...	胜利股份	1
4	8	由于全球最大的俄罗斯Uralkaly钾矿被淹，产量大减，同时满洲里口岸铁路在修复线，导致中...	冠农股份,雅化集团	1
...	...	...	...	...
396376	1037031	每经AI快讯，有投资者在投资者互动平台提问：请问公司目前有没有电解槽产能，规划情况能否详细介...	亿华通	1
396377	1037032	依米康（SZ 300249，收盘价：10.38元）发布公告称，2023年10月12日，依米康...	中泰证券	1
396378	1037033	天风证券10月13日发布研报称，给予中核科技（000777.SZ，最新价：13.03元）买入...	中核科技,天风证券	1
396379	1037034	有投资者提问：抗癌药CPT获批后，公司是否应该按照股权协议继续收购沙东股权，适应症为MM的C...	海特生物	1
396380	1037035	10月13日午间，根据恩捷股份发布的公告，持有公司股份5%以上的股东玉溪合益投资有限公司（下...	恩捷股份	1

396381 rows × 4 columns

② **CFSD (Chinese financial sentiment dictionary)**

We use jieba for word segmentation on each piece of news. We utilize the Chinese financial sentiment lexicon CFSD, which contains positive and negative words. We calculate the proportion of positive and negative words in each news article. If the proportion of positive words is greater than the proportion of negative words, we label it as 1; otherwise, we label it as 0. (https://zhuanlan.zhihu.com/p/82666889)

③ **Using API**

Many companies provide APIs for text sentiment analysis, such as Baidu API, which can perform sentiment analysis (determine the sentiment polarity category of the text and provide corresponding confidence scores. Sentiment polarity can be positive, negative, or neutral), as well as multi-entity sentiment analysis (for subjective textual passages in specific scenarios, automatically identify the core entity words in the text and determine the sentiment and corresponding confidence score for each entity word).

In my view, multi-entity sentiment analysis is very important when multiple company entities appear in a news article and express different sentiments simultaneously.

(https://cloud.baidu.com/product/nlp_apply/sentiment_classify?_=1700550236539)

④ **Machine learning/Active learning**

We can use machine learning to predict the class label for text and we can further improve the performance using active learning.

- Split the text data into labeled and unlabeled sets. The labeled set contains samples with known class labels, while the unlabeled set consists of samples without labels.
- Train a machine learning model, such as SVM, using the labeled data.
- Apply the trained model to predict class labels for the unlabeled data.
- Use an active learning strategy (like most uncertain) to select a subset of the unlabeled

samples for annotation. The adding the class labels to the labeled dataset. And retrain the model using the updated labeled dataset, including the newly annotated samples.

- Iteration: Repeat steps 3-4 until the model's performance reaches a satisfactory level or the desired amount of labeled data is reached.

Task2Q3

The graph drew by using Python and Neo4j has directed edges and represents six types of relationships. The number of nodes and the number of edges for each type of relationship, and the final graph are shown in the Figure 5.

Surprisingly, we found that 11044 is less than the sum of the relation list. Upon further investigation, we discovered that this is because there were originally bidirectional competitive relationships within the list and duplicates, as shown in Figure 4. However, when importing the relationships, only one of the same relationships may have been retained.

Figure 4

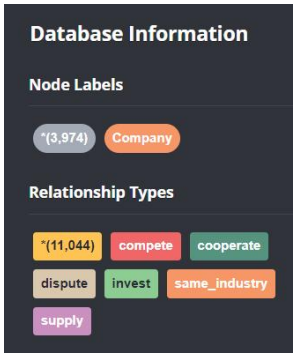
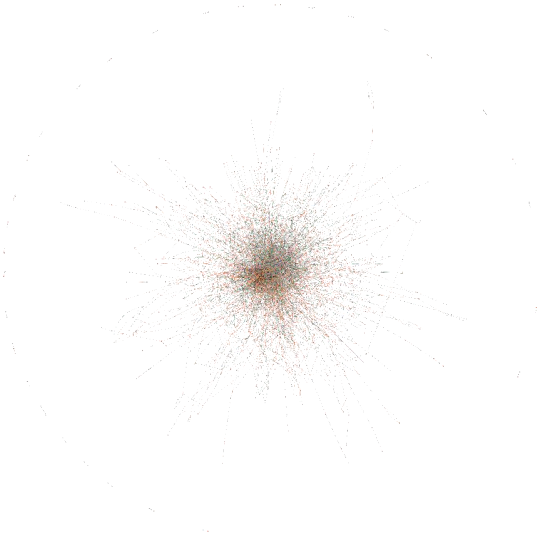


Figure 5



Task2Q4

① based on the graph in Task2Q3

Since we have already established company nodes and their corresponding relationships based on Neo4j and Python in Task 2Q3, mapping the company names to their IDs will allow us to query the respective affiliated companies, thus fulfilling the requirements of this task.

② based on all files under the KnowledgeGraph folder and Task1.xlsx

We read the data file and the knowledge graph file. 'company_df' represents the company nodes data read from a CSV file, and 'relation_dfs' represents the relationship data read from different CSV files.

Then we define two functions, one called 'process_relation' which is used to process each relationship data and add the relevant companies to the result based on the relationship type and another is called 'process_row' which is applied to process each row of data.

- In the 'process_relation' function, the first step is to obtain the relation type and store it in a variable. Then, the multiple company names in the 'Explicit_Company' column are split into a list ('explicit_companies'). Next, by searching for the IDs of the companies in 'company_df' that match the 'explicit_companies', the matched IDs are stored in the 'matched_ids'. Then, the function filters the relationships that are related to the matched IDs and stores the IDs of the other end of the relationships in the 'related_ids'. Finally, based on the relation type, the function determines whether to add the related companies to the

‘Implicit_Positive_Company’ or the ‘Implicit_Negative_Company’.

- In the ‘process_row’ function, we create the empty lists, called ‘implicit_positive_companies’ and ‘implicit_negative_companies’, to store the implicit positive and negative companies. The function iterates over the different relationship data in ‘relation_dfs’ and passes the ‘implicit_positive_companies’ and ‘implicit_negative_companies’ to the ‘process_relation’ function for processing. After processing all the relationship data, duplicate values are removed from ‘implicit_positive_companies’ and ‘implicit_negative_companies’ using list(set(...)).

When storing the result, we noticed that the "None" values should be replaced with the string "None" to avoid any blank content in the task2 table. We can achieve this by using result.fillna("None") to replace the "None" values with the string "None" in the result dataframe. The final result is shown in Figure 6.

```
result
✓ 0.0s                                                    Python
```

	NewsID	NewsContent	Explicit_Company	label	Implicit_Positive_Company	Implicit_Negative_Company
0	1	本报记者 田雨 李京华 中国建设银行股份有限公司原董事长张恩照受贿案 3 日一审宣判，北京市第...	建设银行	0	捷捷微电,任子行	交通银行,佳都科技,中国石化,我爱我家,晨鸣纸业,贵州茅台,中信银行,南天信息,特发信息,中...
1	2	中国农业银行信用卡中心由北京搬到上海了！ 农行行长杨明生日前在信用卡中心揭牌仪式上表示，此...	农业银行	1	交通银行,神州信息,润建股份,牧原股份,紫金银行,南天信息,中信银行,东方财富,优博讯,神州...	ST云维
2	3	在新基金快速发行以及申购资金回流的情况下，市场总体上呈现资金流动性过剩格局，考虑到现阶段权...	中国国航	1	中国电建,南方航空,兴业证券,寒武纪,中国石化,吉祥航空,山航B,春秋航空,中国卫通,农业银...	None
3	4	胜利股份（000407）公司子公司填海造地2800亩，以青岛的地价估算，静态价值在10亿元...	胜利股份	1	新疆浩源,特锐德	None
4	8	由于全球最大的俄罗斯Uralkaly钾矿被淹，产量大减，同时满洲里口岸铁路在修复线，导致中...	冠农股份,雅化集团	1	富邦股份,东华科技,天齐锂业,立讯精密,南宁百货,永兴材料,宁德时代,西部黄金,藏格控股,德...	None
...	...	...	...	...	...	...
396376	1037031	每经AI快讯，有投资者在投资者互动平台提问：请问公司目前有没有电解槽产能，规划情况能否详细介...	亿华通	1	福田汽车,东风汽车,仕佳光子,飞龙股份,百奥泰,宝泰隆,中国船舶,东旭光电	None
396377	1037032	依米康（SZ 300249，收盘价：10.38元）发布公告称，2023年10月12日，依米康...	中泰证券	1	贵州茅台,乐歌股份,光大证券,璞泰来,涪陵榨菜,中金公司,翔丰华,长江证券,东方证券,金融街...	西水股份,华谊嘉信
396378	1037033	天风证券10月13日发布研报称，给予中核科技（000777.SZ，最新价：13.03元）买入...	中核科技,天风证券	1	国信证券,人福医药,复旦微电,久远银海,中科软,中金公司,长江证券,西南证券,建龙微纳,华泰...	吉翔股份,三特索道,中源家居
396379	1037034	有投资者提问：抗癌药CPT获批后，公司是否应该按照股权协议继续收购沙东股权，适应症为MM的C...	海特生物	1	海尔生物,药明康德	None
396380	1037035	10月13日午间，根据恩捷股份发布的公告，持有公司股	恩捷股份	1	福耀玻璃,太平洋证券,德尔股份,平安银行,天赐材料,宁德时	None

Figure 6

Later, I discovered that if we store the relationships of each company as two dictionaries, one for the "same relationship" and the other for the "opposite relationship" (with the key being the company and the value being the corresponding company), and then read the file for Task 1, the speed will be significantly improved in all cases. This is because it eliminates the need for nested loops to iterate through each relationship, instead allowing direct lookups using dictionaries.