DSAA 5002 - Data Mining and Knowledge Discovery in Data Science

(Fall Semester 2023)

Project

Deadline: 5 Dec 2023 11:59pm

Full Mark: 100 Marks

This project consists of two tasks, totaling four questions. Please ensure that you complete them in sequence. Through this project, you will gain knowledge related to financial text analysis, knowledge graph construction, and knowledge-driven decision-making. You are required to submit via Canvas. You may submit several times to correct some mistakes, but please make sure that each submission is a WHOLE submission. Please start early. For each day overdue, the grade for this project will decrease by 10%.

You are required to submit **a ZIP file** via Canvas, which should include the following files:

- (40 marks) report.pdf. Please refer to Appendix 1.
- (20 marks) Task1.xlsx
- (20 marks) Task2.xlsx
- (20 marks) Code. Please refer to Appendix 2.
- (0 marks) Any other materials that you consider crucial.
- Name your zip file**: 5002Project_ID_.zip,** where ID is your student number.

Task 1 (50 marks) Data Preprocessing and Analysis

Background: Assuming you are a sentiment analyst at a securities firm, your task is to assess the impact of each news article on the A-share listed companies explicitly mentioned. For instance, on October 14, 2022, the China Securities Journal(中国证券报) reported the following:

2022-10-14	截至 10 月 13 日，包括贵州茅台、今世缘、水井坊等多家酒企披露了前三季度业绩预告或经营数据公告。从目前已披露的数据来看，前三季度酒企业绩表现稳定。

This news explicitly mentions three companies: 贵州茅台(600519.SH), 今世缘(603369.SH), and 水井坊(600779.SH), and the impact appears to be positive. "Positive" indicates that this news appears to positively affect the company's stock price.

Question 1(20 marks): Data Preprocessing - Noise Removal

Input: News.xlsx, which includes **1,037,035** pieces of news. Please download it from:
https://docs.google.com/spreadsheets/d/1VAzteetSSc9WOCne_u6-5oFt_6rIMR5E/edit?usp=share_link&ouid=112799952654350672254&rtpof=true&sd=true

Description: This is an open-ended question. Given the input, we consider any news that does not mention any of the China A-share listed companies to be **noise**. Your task is to **remove rows** in the data table News.xlsx that do not mention any of the China A-share listed companies(as provided in A_share_list.json). An intuitive approach is to use a rule-based strategy, such as building a dictionary of the names of listed companies and then conducting a brute force search. However, news articles may not always provide companies' full names, as is the case with news from the Shanghai Securities Journal(上海证券报):

2022-10-14	市场震荡之际，知名投资人动作频频。段永平透露，10 月 12 日买入了茅台

This news clearly has a positive impact on 贵州茅台(600519.SH). However, the original text uses the abbreviation "茅台,"

which does not match "贵州茅台." Similar examples include "中石油", "建行", "海尔" and so on. While more complex rule-based methods can be applied, we also **encourage** you to explore some methods based on **similarity**. Similarity-based methods involve projecting text or segmented text into an N-dimensional space, creating an N-dimensional vector, and then calculating the distance between different texts in the vector space. You can choose any text vectorization method that interests you, such as using a deep learning-based Chinese pre-trained BERT model. Feel free to check out Appendix 3 for additional recommended reading materials.

Submission (20 marks): You should document the process in your report.pdf file. For detailed requirements, please refer to Appendix 1.

Question 2(30 marks): Data Analysis - Text Knowledge Mining

Input: The news data was cleaned through Question 1.

Description: Building on Question 1, we assume you have obtained a clean dataset where each news mentions at least one A-share listed company. In this question, your objective is to determine the **sentiment polarity** of each news text. This task can be treated as a binary classification problem, with Class 0 indicating "negative" and Class 1 indicating "positive". You can refer to the *submission_excel_sample/**sample_Task1.xlsx*** file for more information. You can use any method to achieve this, including but not limited to:

a) Building a small dataset, manually annotating the data, and training a classifier for sentiment analysis. It's important to vectorize the text for this task. For details on text vectorization, please refer to Appendix 3.

or b) Using existing Chinese sentiment analysis libraries.

or c) Any other techniques.

Submission1 (10 marks): You should document the process in your report.pdf file. For detailed requirements, please refer to Appendix 1.

Submission2 (20 marks): For Task 1, you are required to submit an **EXCEL file** named **Task1.xlsx**, which should include the following content:

a) You can refer to the *submission_excel_sample/**sample_Task1.xlsx*** file for more information.
b) You only need to provide the data after noise removal.
c) Task1.xlsx should have four columns: *NewsID*, *NewsContent*, *Explicit_Company*, and *label*.
d) *NewsID* and *NewsContent* contain the same data as in New.xlsx after the noise removal.
e) In the *Explicit_Company* column, list the full names of the companies explicitly mentioned in the news, separated by English commas, for example, "贵州茅台,中国铝业,德必集团". The *label* column indicates the sentiment polarity of NewsContent, where 0 represents negative sentiment, and 1 represents positive sentiment.
f) You should strive to ensure the accuracy of *Explicit_Company* and *label* as the TA will randomly sample your submitted Excel file and grade it based on accuracy.
g) Please submit the codes.

Task 2 (50 marks) Application of Knowledge Graph

Background: In addition to explicitly mentioning listed companies, each news article may also implicitly impact the other companies, either positively or negatively. For instance, on October 14, 2022, the China Securities Journal(中国证券报) reported the following:

2022-10-14	截至 10 月 13 日，包括贵州茅台、今世缘、水井坊等多家酒企披露了前三季度业绩预告或经营数据公告。从目前已披露的数据来看，前三季度酒企业绩表现稳定。

This news explicitly mentions three companies: 贵州茅台(600519.SH), 今世缘(603369.SH), and 水井坊(600779.SH), and the impact appears to be positive. However, other companies such as 五粮液 (000858.SZ), 洋河股份 (002304.SZ), 泸州

老窖 (000568.SZ), and 山西汾酒(600809.SH) might also be positively affected, as they belong to the same industry as 贵州茅台(600519.SH). Conversely, this news might have a negative impact on 贵绳股份(600992.SH) and 宁德时代(300750.SZ), as 贵绳股份 has a dispute with 贵州茅台, and 宁德时代 competes with 贵州茅台.

In the above analysis, expressions like "belong to the same industry," "have a dispute," and "compete" can be considered as forms of **knowledge**. The most well-known data structure for representing knowledge is a **knowledge graph**, where nodes represent entities, and edges represent relationships. Each relationship connects two entities and can be represented using a triple (S, P, O) = (Subject, Predicate, Object). For example, "贵州茅台(600519.SH) and 贵绳股份(603369.SH) have a dispute" can be represented as a triple (600519.SH, "dispute", 603369.SH), and "宁德时代(300750.SZ) competes with 贵州茅台(600519.SH)" can be represented as a triple (300750.SZ, "compete", 600519.SH).

Fortunately, a research team has studied the relationships between A-share listed companies and provided the data in the *KnowledgeGraph* folder. In the following questions, you are required to use all the data from the *KnowledgeGraph* folder.

Question 3: (10 marks) Constructing a Knowledge Graph

Input: ALL files under the *KnowledgeGraph* folder

Description: To construct a knowledge graph, you will need to use **Python** and **Neo4j** database. To create the knowledge graph, you should install the Neo4j graph database on your computer and use Python to connect to it. Then, you can construct the knowledge graph based on the input files. Please search for relevant content on your own. In the knowledge graph you build, the node type is "company," and there are six types of edges: "compete," "cooperate," "dispute," "invest," "same_industry," and "supply." Edges can be directed, meaning from S to P, or undirected (bidirectional).

Submission (10 marks): You should document the process, and plot the knowledge graph in your report.pdf file. For detailed requirements, please refer to Appendix 1.

Question 4 (20 marks): Knowledge-Driven Financial Analysis

Input: (1) ALL files under the *KnowledgeGraph* folder, and (2) your Task1.xlsx in task 1.

Description: Please, based on the rules described in Table 1, identify **ALL** implicit companies corresponding to each company of *Explicit_Company* in your own Task1.xlsx file. Categorize them into *Implicit Positive Companies* and *Implicit Negative Companies*.

Table 1: Correspondence between company relationships and news sentiment. For example, taking "same_industry" as an example, if there is a "same_industry" relationship between companies A and B, then the impact of news on A and B is the same, either positive (1) or negative (0). If companies A and B have a "compete" relationship, then the impact of news on A and B is opposite.

Relation between company A and B	The impact of news sentiment on companies	Case 1		Case 2	
		New impact on A	New impact on B	New impact on A	New impact on B
compete	opposite	1	0	0	1
cooperate	same	1	1	0	0
dispute	opposite	1	0	0	1
invest	same	1	1	0	0
same_industry	same	1	1	0	0
supply	same	1	1	0	0

Submission: (20 marks): For Task 2, you are required to submit an **EXCEL file** named Task2.xlsx, which should include the following contents:

a) You can refer to the *submission_excel_sample*/**sample_Task2.xlsx** file.
b) It should have six columns: *NewsID*, *NewsContent*, *Explicit_Company*, *label*, *Implicit_Positive_Company*, *Implicit_Negative_Company*.
c) The first four columns are the same as in Task1.xlsx.
d) The *Implicit_Positive_Company* column should list implicitly mentioned companies with a positive impact by the news, and the *Implicit_Negative_Company* column should list implicitly mentioned companies with a negative impact.
e) *Explicit_Company, Implicit_Positive_Company and Implicit_Negative_Company* can overlap.
f) Please submit the codes.

Appendix 1. Report Requirements

Your PDF report should be a **maximum of 5 pages** with no specific formatting requirements, but you must clearly indicate the question numbers. If you have used any code from Github in your project, you MUST declare it in this report. Otherwise, it will be considered as CHEATING.

Your report should include, but is not limited to, the followings:

(a) (15 marks) Task1Q1 - Present the noise removal strategies you've tried and analyze the results. Please provide as much detail as possible about your strategy, including equations, images, and so on.

(b) (5 marks) Task1Q1 - Introduce the filter rates of each strategy, i.e.,

$$filter\ rate = \frac{\#\ \text{filtered news}}{\#\ \text{Total news}}$$

(c) (10 marks) Task1Q2 - Present and analyze the strategies you've tried. Please provide as much detail as possible about your strategy, including equations, images, and so on.

(d) (10 marks) Task2Q3 - Plot the knowledge graph you independently constructed. Your images should clearly include all edge types, node types, and their quantities, as shown in Figure 1.

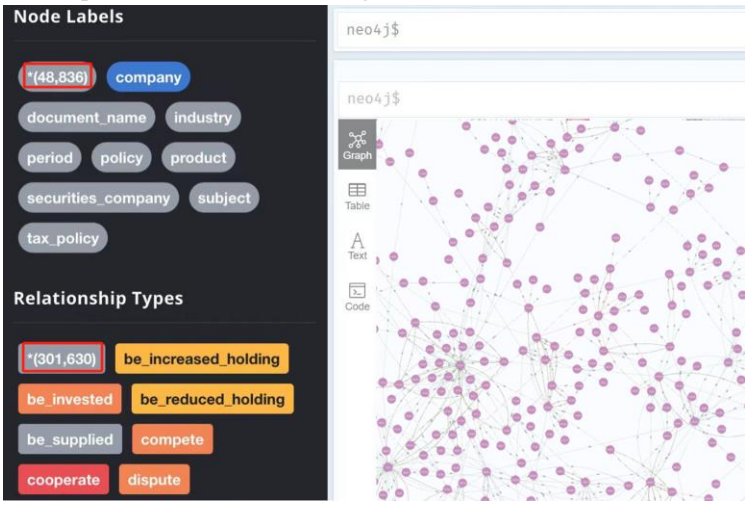

Figure 1. An example of a knowledge graph visualized using Neo4j (unrelated to this project)

Appendix 2. Code Requirements

1. You may include one or more Python files, but please include a *README* file to clearly outline the contents of the files.
2. You are NOT allowed to plagiarize your classmates' code. If detected, this project will be graded as ZERO.
3. If you have used any code from Github in your project, you MUST declare it in the report
4. Please provide all necessary files to run your code, including but not limited to .xlsx, .csv, etc. Also, provide the *requirements.txt* file for your code.
5. In your code, it is necessary to clearly indicate the corresponding question and task. You are suggested to use ''' ''' symbols

to add comments within the code.

Appendix 3. Recommended reading materials

1. Text vectorization -
   a) https://www.uni-mannheim.de/media/Einrichtungen/dws/Files_Teaching/Data_Mining/Slides/DM05-Text-Mining.pdf
   b) BERT: https://huggingface.co/bert-base-chinese
2. Sentiment analysis - Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey[J]. Ain Shams engineering journal, 2014, 5(4): 1093-1113.
3. What is financial knowledge graph? - Wang W, Xu Y, Du C, et al. Data set and evaluation of automated construction of financial knowledge graph[J]. Data Intelligence, 2021, 3(3): 418-443.
4. Where does the knowledge graph data come from? – please give it a star: https://github.com/K-Quant/HiDy

Feel free to reach out to the TA via email
if you have any questions about the project: jli226@connect.hkust-gz.edu.cn.

Have fun!