BAIT 508 Group Project: Industry Analysis

PROJECT OVERVIEW:

The goal of this project is to conduct an in-depth analysis of public US firms within selected industry sector(s) using various data analyses and natural language processing (NLP) techniques that we learned in BAIT 508. Each team will choose at least one industry sector to investigate and utilize multiple datasets to extract valuable industry insights from the data.

The project will utilize three datasets (located in the data folder):

- public_firms.csv
- major_groups.csv
- 2020_10K_item1_full.csv

DETAILED INSTRUCTIONS:

Part 1. Quantitative Analysis of the Industry Sector

A. **[Industry Sector Selection and Data Filtering; 20 points]**

1. The file "data/major_groups.csv" contains a list of major industry sectors and their corresponding codes (column "major_group"). Your first task is to **choose at least one industry** sector that interests your group. It is okay if multiple groups choose the same industry sector, so you don't need to coordinate with other groups.
2. Next, filter the data in "data/public_firms.csv" to only include the firms belonging to the industry sector(s) you have selected. You can use the "major_group" value, which corresponds to the first two digits of each firm's SIC code,[1] to identify relevant firms. For example, if you are interested in the "Business Service" sector and its "major_group" code is 73, you should retain all firms whose SIC codes start with 73.
3. Now, answer the following questions based on the filtered dataset:
    a. How many unique firm-year ("fyear") observations are there in the filtered dataset?
    b. How many unique firms are there in the filtered dataset?
    c. How many firms in the filtered dataset have records over all 27 years (1994-2020)?

B. **[Preliminary Analysis; 20 points]** Answer the following questions:

1. What are the top 10 firms with the highest stock price (column "prcc_c") in the year 2020?
2. What are the top 10 firms with the highest sales (column "sale") in the entire history of the dataset?
3. What is the geographical distribution (column "location") of all the firms? In other words, how many firms are there in each location? Please list the top 10 locations.
4. Create a line chart to show the average stock price (column "prcc_c") in the selected sector(s) across the years. If you have selected multiple sectors, draw multiple lines to show them separately.
5. Which firm was affected the most by the 2008 Financial Crisis, as measured by the percentage drop in stock price from 2007 to 2008?

[1] SIC stands for Standard Industry Classification. Please see https://www.sec.gov/corpfin/division-of-corporation-finance-standard-industrial-classification-sic-code-list for more details.

6. Plot the average Return on Assets (ROA) for the firms located in the "USA" across the years. ROA is calculated as ni/asset.

Part 2. Text Analysis on the Industry Sector

C. **[Text Cleaning; 10 points]** The file "data/2020_10K_item1_full.csv" contains a sample of 5,988 firms and their "item_1" content in their 10-K reports in the year 2020.[2] Load the dataset as a DataFrame and create a new column containing the cleaned text for each "item_1" content. Follow the steps below to clean the text:

1. Convert all words into lowercase.
2. Remove punctuations.
3. Remove stop words based on the list of English stop words in NLTK.

D. **[Keyword Analysis; 20 points]** Conduct keywords analysis on your selected industry sector(s). Follow the steps below to complete the analysis:

1. Create a new DataFrame that includes only firms in your selected industry sector(s). Ensure that you merge the 10-K data with the previous "public_firm.csv" data using an inner join.
2. Generate the top 10 keywords for each firm based on two different methods: word counts and TF-IDF score.
3. Create two wordclouds to visualize the keywords across all firms in the selected sector(s): one based on the word counts and another based on the TF-IDF scores.

E. **[Word embedding; 20 points]** Train a word2vec model and analyze word similarities.

1. Train a word2vec model with the **full 10-K sample** (e.g., "data/2020_10K_item1_full.csv"). Please use the **cleaned text** (e.g., results from Step C) for training.
2. Manually inspect the wordclouds you generated in D.3 and choose three representative keywords that are relevant to the industry sector of your interest. Utilize the trained word2vec model to find the most relevant five words for each of these three keywords.

Part 3. Comprehensive Analysis of One Sample Firm

F. **[Firm Analysis and Strategy Suggestion; 10 points]** This is an open question. Pick one firm that your group is interested in and try to analyze its market status. The ultimate goal is to provide **one valuable suggestion** to the firm based on your analysis. Some directions you might consider are, but not limited to:

1. Convert the keywords extracted in D.2 into word embeddings with the word2vec model trained in E.1. Add up the embeddings for each firm to create the firm-level embeddings. Use the firm-level embeddings to find the focal firm's competing firms (or most similar firms).
2. Compare the revenue, market share, and ROA of the focal firm to its competitors and provide suggestions accordingly.
3. Perform an analysis of the historical stock prices, ROA, revenue, and assets of the chosen firm. Investigate potential correlations and address noteworthy decreases and increases.

[2] Please refer to this wiki page to understand 10-K annual reports: https://en.wikipedia.org/wiki/Form_10-K

Note: Please focus on **one** direction and provide **one** suggestion to the firm. It is a busy time with many finals and projects. We don't want to overwhelm you with an extensive report 😊.

Project Report: Please write a report on the project considering the followings:

1. The report has to be self-explanatory. This means that you don't expect readers to check your codes or Jupyter notebooks. A general guideline would be: imagine that you will send this report to your potential employers for a job interview.
2. On the first page, indicate the team members (full name, email, student ID) and section number. Please describe each member's role in the group project.
3. Then describe the procedures (Steps A-F) and the results.
   a. Describe your code at a high level. You don't need to copy your entire code in the report, but you may include important code snippets for explanation.
   b. Please think about how to deliver information effectively. For example, attaching blurry screenshots will be a factor of deduction.

Submission Instructions:

1. This is a group project. Form a team of 2-3 within the same section.
2. In UBC Canvas, submit the following in a zip file:
   a. Your project report (.pdf or .docx)
   b. Python codes / Jupyter notebooks (.ipynb or .py)
3. Late submissions will not be accepted.
4. A plagiarism check will be conducted. Do not share code/data across different groups. Also, please put appropriate references in your code(s) and report if you use external sources (e.g., Stack Overflow, ChatGPT). Using the lecture codes provided by the instructor is permitted.
5. The use of ChatGPT or other Generative AI tools: You are **allowed** to use ChatGPT or other Generative AI tools for the project. However, please disclose which tools you used and how you used them (e.g., we used this prompt to get the initial code base; we used ChatGPT to correct our code, we used ChatGPT to copyedit our report, etc.). Please refer to the Course Outline for attribution guidelines.