Individual Project – Free Topic

Sentiment Analysis in management discussion and analysis (MD&A) from company's SEC filing
Chengbo Jiang
chengbo6@illinois.edu

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Net ID: Chengbo6, this will be an individual project so Chengbo Jiang will be served as the captain and other roles.

2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

- Free Topic: Sentiment Analysis in management discussion and analysis (MD&A) from company's SEC filing.
- The task is to extract the management and analysis (MD&A) sections of text from public companies' 10K (Annual) and 10Q (quarterly) filings from U.S. Securities and Exchange Commission (SEC) database.
- After extracting the texts, I will try to classify the sentiment whether they are positive or negative using the methods and algorithms learned from CS410.
- Positive and negative can be determined via counting positive or negative words from dictionary.
- Other than accounting information on those filings, text information could be valuable but overwhelming. Classifying these texts information in positivity or negativity will streamline the analytical procedures of public companies as well analysts' understanding. For example, if we can filter the negatives so that we can concentrate more on analyzing the positive ones.
- The planned approach:
  1. Write a web crawler to download the reports from SEC's website.
  2. Extract the MD&A information sections from the downloaded reports.
  3. Sentiment analysis using methods such as Naive Bayes, Multinomial Naive Bayes etc.
- Tools: Python (metapy, nltk, numpy, beautifulsoup etc.)
- Datasets: The EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database from SEC.
- Expected Outcome: Expect to accurately classify the textual data in their respective categories.
- Evaluation: Since I will use different methods, I will compare the results from different methods.

3. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

1. Write a web crawler to download the reports from SEC's website. 40 Hours
2. Extract the MD&A information sections from the downloaded reports. 10 Hours
3. Sentiment analysis using methods such as Naive Bayes, Multinomial Naive Bayes etc. 40 Hours