Guiding questions for the check-in report

<https://www.overleaf.com/5427356137dzgttgsfttnt#66005f>

Vijay: Data acquisition and integration

* What data did you get (and, more generally, what data do we have)?
* What features are in the data (what columns/tags are there)?
* Where did you get the data?
* How much data is there?
* How did you get the data?
* How is the data stored?
* What is the process to get more data?
* What difficulties did you encounter in acquisition or integration?

Andrew: Data cleaning

* What needs to be cleaned from the data?
  + In the database of pulled data from the SEC website, we need to first remove a lot of the useless tags. There are about 521 distinct tags per company. After removing the redundant tags, we will then need to clean the information stored in the value column.
* Do the things that need to be cleaned differ based on the type of data?
  + The tags need to be cleaned based on whether they point to important information in the value column.
  + The value column then needs to be cleaned by removing all the unnecessary text and converting all the information into some kind of time series.
* Why do we need to clean the data?
  + Our data is full of junk data that we do not need. There is no reason to load or sift through all that data every time we run through the program.
  + It also helps us to focus on the big data points and not stuff that may or may not have an impact. This will drastically improve our computation time with not very much loss in our predictions.
* What is the process you are using to clean the data?
  + So far when trying to eliminate tags, we have been using a combination of sifting through the tags manually by searching for specific keywords and looking through the data myself.
  + We have not gotten to cleaning up the value column. We are probably going to be using some form of keyword analysis to pull dates and important data from the value table.
* Is there any way to automate the cleaning?
  + We have not found a way to isolate the important tags in an automated way, but once the important tags are found it shouldn’t be too big of a problem to automate the rest using some form of keyword analysis.
* What information is being retained after the cleaning process?
  + We want to keep dates, revenue, the direction of the company, etc.. Anything that could help us make better predictions. We might clean the data even further at a later date.
* What difficulties did you encounter in cleaning?
  + Our major difficulty was with finding all the important tags. We are doing most of that manually.

Harry: Sentiment analysis

* What is sentiment analysis?

Sentiment analysis is a process in which language is converted into a machine decipherable format usually a number that corresponds to how positive or negative the sentence is, but sentiment analysis is a broad research area and there is no one agreed upon technique that is best. There are ML approaches as well as more traditional approaches that map words to positive or negative values.

* What are we using sentiment analysis for in our project?

In our project sentiment analysis will play a valuable role in identifying whether sentences within SEC filings an overall positive or negative outlook. It is a reasonable assumption from there that sentences with a more positive outlook correspond to a more positive outlook for that company, and vice versa.

* What data are we using sentiment analysis on specifically?

As mentioned in the previous paragraph we will be using SEC filings with our sentiment analysis tool. The aggregated dataset which we have collected contains all the text within these reports that can provide a picture of the overall sentiment of the sentence. Combining this with every piece of text from the filing, we can achieve an accurate and comprehensive sentiment from the whole document. Using the whole document will allow sentences that are outliers in their sentiment as it relates to the rest of the document to not make an adverse impact in our modeling or analysis.

* What are we expecting to see from the sentiment analysis from the data?

We expect that overall document sentiment is highly correlated with stock performance, e.g. a company with a mediocre sentiment has an overall flat stock price not seeing much movement.

* What options do we have for sentiment analysis?

For sentiment analysis there are several tools each with their own theoretical benefits, our team has chosen two tools to explore these pros and cons with, VADER and FINBERT. VADER is a more traditional approach to sentiment analysis using human graded sentiments and their algorithm. FINBERT is a machine learning model using BERT embeddings fine-tuned on financial data with a sentiment analysis aspect on top. It is trained on a sentiment data set presumably from human graders.

* What are the pros and cons of each option for sentiment analysis?

VADER is a highly optimized algorithm that can allow us to process more data at a higher speed giving us the ability to run this tool on more companies giving us more data for our analysis. VADER is a standard tool and is known for its robustness, but it is not tuned for finance specifically. FINBERT on the other hand is a more modern approach using state-of-the-art techniques and could give us more accurate sentiments, but being around seven billion parameters it is highly likely that inference would take much longer, giving us less companies to work with.

* What difficulties did you encounter in sentiment analysis?

Sentiment analysis is a difficult task, and with the complex nature of language these tools will often misunderstand what certain words mean in context. This includes within sentences, but considering paragraphs and the whole document words can be misinterpreted easily. Still we believe that the average sentiment of the document is still accurate.

**Peter: Topic modelling and keyword analysis**

* What are topic modelling and keyword analysis?

Keyword analysis is a process which takes in a document or collection of documents and outputs the important or key words in the document. Topic modelling is a process which takes in a document or collection of documents and outputs collections of words which represent the overall topic or theme of the document.

* What are we using topic modelling and keyword analysis for in our project?

In our project, we are using keyword analysis and topic modelling to identify future directions and focuses that companies are working toward.

* What data are we using topic modelling and keyword analysis on specifically?

We are specifically applying keyword analysis and topic modelling on excerpts of publicly disclosed 8-K and 10-K reports obtained from the SEC.

* What are we expecting to see from the topic modelling and keyword analysis from the data?

The main thing we hope to see are new technologies such as cloud computing and generative AI. We also anticipate seeing more general concerns such as sustainability and diversity. Another thing we expect to see is more high-level descriptions such as risky or cautious.

* What options do we have for topic modelling and keyword analysis?

For keyword analysis, we plan to use term-frequency inverse-document-frequency (TF-IDF), though we could simply use term frequency. Term frequency is simply the count of how many times a term appears in a document. TF-IDF is a measure of how important a term is to a particular document.  
  
For topic modelling, we plan to use non-negative matrix factorization (NMF) or latent Dirichlet analysis (LDA). NMF is a technique which converts a document-term matrix into two lower-rank matrices: one that represents the relationship between documents and topics, and the other that represents the relationship between topics and terms. Furthermore, it outputs matrices which are non-negative which simplify interpretation for text analysis tasks. LDA is a technique which represents each document as a mixture of some number of topics which are computed as part of LDA via a generative, probabilistic model.

* What are the pros and cons of each option for topic modelling and keyword analysis?

Term frequency is useful because it is simple to compute, but it requires careful filtering of common words which do not contribute to future directions or focuses. TF-IDF is slightly more complex to compute, but it helps remove the bias that term frequency has toward irrelevant common words.  
  
NMF is useful because it is simple to calculate via linear algebra, but the topics which are recovered can be difficult to interpret due to the inherently linear algebraic nature of the decomposition. LDA is useful because the topics it computes are generally easier to interpret than the topics from NMF, but it can take more time to compute.

* What difficulties did you encounter in topic modelling and keyword analysis?

One major difficulty we have encountered is determining the appropriate thresholds for filtering out keywords and topics. This is because some common words which appear in many documents have not been removed. Another difficulty we have is in identifying topics from some of the topic decompositions – the collections of words returned from topics are not always focused on a specific topic as we would have hoped.

Remaining topics to be addressed:

Cross-correlation

* What is cross-correlation?
  + Cross-correlation is a statistical measure used to quantify the relationship between two time series, in this case, stock prices. It is used to assess the degree to which the movement of one stock (or asset) is related to the movement of another over time.
* What do we plan to use cross-correlation for in our project?
  + We plan on using cross-correlation on pairs of different stock prices to evaluate how coupled the price changes are between two stocks.
* What are we expecting to see from cross-correlation?
  + We expect to see that companies which have some sort of partnership to have highly coupled price changes (i.e. price changes which are have higher correlation than the average correlation).
* What limitations do we anticipate cross-correlation having in our use case?
  + Because companies may have only short partnerships, it is possible that the cross-correlation computation needs to be performed on a smaller window of the stock price data.

Tentative models and their evaluation (for each model):

* What is a high-level description of the model?

We have decided on using two models, one for interpretability and the other for predictive accuracy. The first model which we can interpret easily will be a random forest which falls into tree-based methods. The second model focusing on accuracy will be an LSTM which has shown success with predicting stock prices.

* How are we using each model on the data?
* What are the limitations of the model?

We are still evaluating model performance and limitations, at the time of submission.

* How will we evaluate the efficacy of each model?

For evaluating efficacy, we will use Mean Squared Error on a separate testing set that is approximately twenty percent of the dataset. This testing set will split daily pricing data as well as financial reports to ensure there is no leakage.

Hypotheses for patterns we will find from our data (for each overarching goal):

* What patterns do we expect to find?
* How will we check these hypotheses?

Questions from professor:

* Can we detect sudden changes in partnerships and what the partnerships change to?
* What kinds of baseline correlations can we expect for pairs of companies?
* Do we need to consider volatility of stocks by themselves (i.e. do we need to factor in that company X stock is more volatile than company Y stock)?

It could be helpful to consider the volatility of a company's stocks to make future predictions of where that stock might go. We may consider adding a volatility measure to our prediction algorithm in the future.