Chapter 1. Introduction *(3 pages)*

1.1 Motivation (1 page)

[Financial reports are a great source of information. Information is embedded in texts in a linear and scattered fashion. Experienced financial analysts accumulate sector knowledge through years of following certain companies' performance; they could 'see' the long term trends by 'connecting the dots'. This project inspires to mimic the analysts' work flow, automating the text analytics process and systematically extracting information at scale. With graph-based data model and algorithms, new insights on performance drivers can be discovered across different industries and over an extended period of time.]

1.2 Objectives (1 page)

[The final output would be a connected graph which enable users to visualize clusters of companies which are similar in terms of performance drivers. Discover new investment opportunities and quick overview of comparable universe. Visualization of trends instead of having to read annual reports in serials of text.]

1.3 Outline (0.5 page)

[…]

Chapter 2. Background and Related Work *(15 pages)*

2.1 Financial Reporting (2 pages)

[Regulatory requirement, EDGAR database, Management Discussion & Analysis, text analytics application in finance; various researches on reporting language vs financial performance, credit attribution, etc.]

2.2 NLP and Information Extraction (9 pages)

[Statistical methods, document representation, bag of words, template filling, word embeddings, linguistics features, ML-based language models, etc.]

2.3 Clustering Algorithms (4 pages)

[Unsupervised learning, K-Means, DBSCAN, Self Organizing Map]

Chapter 3. [Conceptualization/Framework] *(15 pages)*

3.1 Overview and Objectives (1 page)

3.2 Data Model (5 page)

[Graph representation, causality modelling, connectivity across sectors and through time]

3.3 System Requirement (5 page)

[Causality extraction, word embeddings, semantic clustering, similarity measures]

3.4 Queries and Visualization (4 page)

[Temporal trends, similarities between companies, prediction - identify leader and laggers affected by the same factors with a time lag]

Chapter 4. Experimental Results *(20 pages)*

4.1 Data Collection (2 page)

[…]

4.2 System Implementation (10 page)

[…]

4.3 Evaluation (4 page)

[…]

4.4 Use Case Demos (4 page)

[…]

Chapter 5. Conclusion *(2 pages)*

5.1 Discussion (1 page)

[…]

5.2 Future Work (1 page)

[…]

[DRAFT] Chapter 3

3.1 Overview and Objectives

The aim of the system is to provide a holistic view of financial performance drivers in S&P 500 companies. The fundamental question we would like to answer are: what are the key drivers of financial performance? How does the trend change with time? Which companies share similar underlying drivers, especially when they do not belong to the same sector?

The primary source of information is derived from the quarterly reports filed by the listed companies on the SEC EDGAR system. The final output is a database containing explanation drivers for financial performances in a map format for better visualization, which enables intuitive exploration for various end applications.

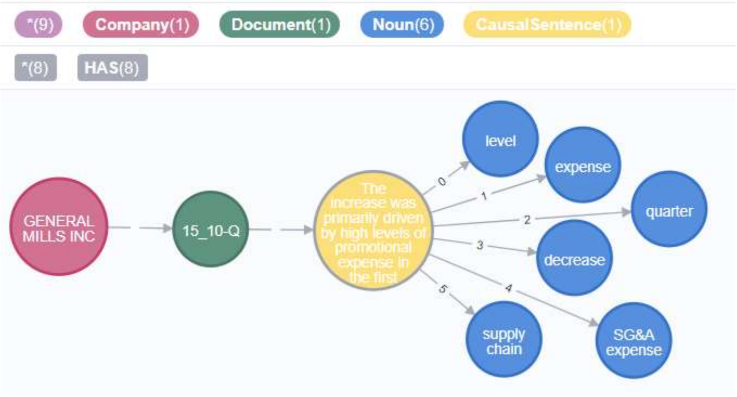
[… more on system requirements …]

3.2 Data Model

The overall data model schema is envisioned as:

(Company) -*[has]*-> **(Document)** -*[has]*-> **(CausalSentence)** -*[has]*-> **(Noun)**

Keys: **(Node)** -*[edge]*-> **(Node)**



[Figure xxx. Caption xxx]

In our model, a company is perceived as a collection of all its financial documents (i.e. 10K and 10Q reports) over an extended period of time. Conceptually, it can be represented by a hypergraph with a hyperedge encapsulating all the associated documents. In practice, a hypergraph can be implemented as a directed graph consisting of a synthetic parent node (representing a company) connected to all child nodes (representing documents), with labelled edges that specify the reporting year. Each company node also has properties such as {company name, ticker, industry classification, etc.}. Each document node has properties such as {file id, file type, index location for MD&A section, etc.}.

$$Company := \{ Document\_i \},\ i \in N;$$

Similarly, each document can be perceived as a collection of sentences. In this project, we focus on only the relevant sentences that contain causal information on our chosen topic of interest. More specifically, we select sentences with explicit causal markers (e.g., 'due to', 'driven by', 'attributable to', etc.), which express a causal relationship between the performance (effect clause) and the drivers (causal clause). We encode each sentence as a node with connection to the document to which they belong to.

$$Document := \{ CausalSentence\_j \}, j \in N$$

$$CausalSentence := \{ Noun\_k \}, k \in N$$

Next, noun phrases in the causal clause of each sentence are also represented as nodes, which are connected to the sentence node. The edge between the noun node and the sentence node has a property that encodes the order of sequence according to which the noun appears in the clause. For example, the sentence *"The increase was primarily driven by high levels of promotional expense in the first quarter of fiscal 2015, a decrease in SG&A expenses, and lower supply chain costs"* contains the following noun phrases (excluding adjectives) in the causal clause:

[1. level, 2. expense, 3. quarter, 4. decrease, 5. SG&A expense, 6 supply chain costs]

The edge connecting noun node 'level' to the sentence node has a property value of 1; the edge associated with 'expense' has a property value of 2 and similarly for the rest of the nouns in the list.

Clustering algorithms can then be performed on the noun nodes to form semantically similar concept groups. Due to the hierarchical structure of our data model, each company can be effectively perceived as a collection of patterns formed by the concept clusters ordered in a specific sequence. (For instance, pattern 1 = [product, demand, location], pattern 2 = [commodity, price, weather]). These patterns are also time-stamped as they can be traced back to a financial report corresponding to a specific time period. This would ultimately enable us to find similarity measures between different companies through time.

3.3 System Requirement

There are five stages in the system. Overview of the pipeline:

1. Obtaining raw input text data and preprocessing
2. Causality extraction
3. Semantic clustering
4. Query in graph database
5. Output visualization and application cases

Financial documents as Input

- EDGAR database accessibility: automatic download on scale is possible

- Each document has a time period associated with it.

- Financial reports are well structured: most relevant is Management Discussion and Analysis, in which the management discusses the historical financial performance and attribute to various factors such as … ….

- Narrowing down to the MDA section of each financial document, which can be easily identified by regex patterns; this is to mimic real life analyst behavior – if I am interested in finding out. … I will go directly the MDA section.

- Language in financial reports – concise and formal. Clearly articulated. Sometimes complex sentences and long – multi-layered reasoning. Tends to follow certain patterns when it comes to explanation

- Once the MDA sections are extracted, the next phase is identify sentences with causality expressions and with relevant topics

Causality Extraction

- Causality extraction could be perceived as a classification task. Supervised learning if gold standard labels are available. Or rule-based heuristic approach ← baseline

- System requirements: precision is (probably) more important than recall in this project. Hence, focus on explicit causal indicators rather than implicit causal relationships.

- The sentence candidate pool can be narrowed down further by screening for selected topics of interest (such as sale/revenue, margin/profitability, etc.)

- A causal sentence can be modelled as (cause) –[casual indicator]->(effect); template filling; consideration including voice (active vs passive), POS tags, connector types (verb vs non-verb), etc.

- More complicated cases: multiple causalities in the same sentence (more than one cause/effect in a single sentence), causality expressed over multiple sentences (co-reference resolution), etc.

- Identification of effect and extraction of casual phrases in terms of nouns (TBD: only nouns or adj + nouns?)

- Rationales: basic concepts are expressed fundamentally in nouns (noun bias - linguistic backup?)

Semantic Clustering

- Categorized all noun phrases into concept groups for summarization and visualization

- Effectively reducing dimensionality: number of clusters rather than number of unique nouns. Some nouns are synonyms that express the same concept.

- Choice of word vectorization: word2vec, GloVe, BERT embeddings, etc. semantically encoded and trained on large corpus (rather than TFIDF over the just the financial document vocab)

- Choice clustering algorithm:

  - K-means –> need to preset number of clusters
  - Self Organizing Map (SOM) → need to preset grid size, i.e. # of neurons
  - DBSCAN -> Not sure if cluster has similar density

- Other consideration for clustering:

  - Static vs online incremental clustering
  - Quality of clustering – how to evaluate and benchmark?
  - How to decide on the optimal number of clusters? How to decide if two clusters are semantically different enough? hierarchical classification?
  - How to decide when to update clusters – splitting, merging, when to form a new cluster?
  - How to classify a new node: nearest neighbor or distance to cluster center?

Query in Graph Database

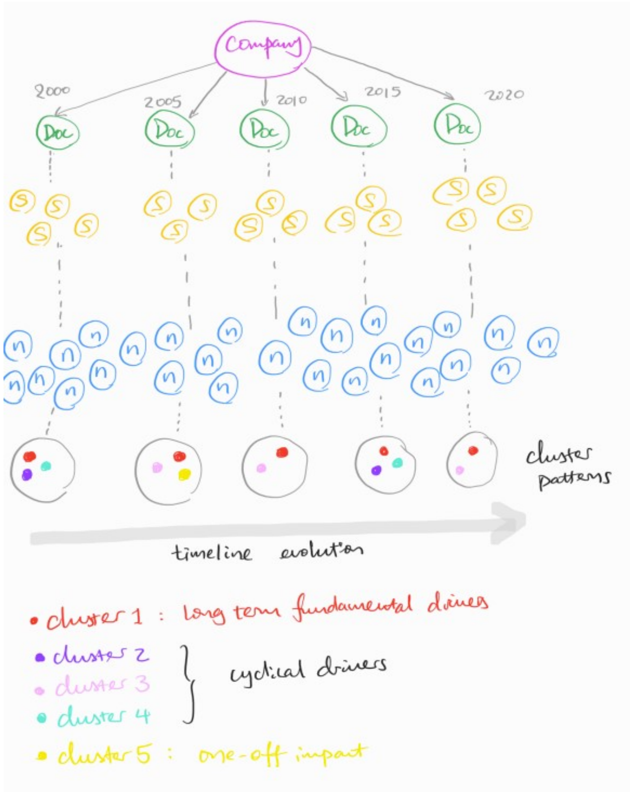
Design codes in cypher for specific queries, such as:

- For a set of input drivers (e.g. chip shortage), which companies are most affected by them?
- For a selected company, what other companies share the most similar divers with it? (i.e. comparable universe)

Need to think more on: How to use cluster patterns for similarity measures, how to do ranking, etc.

Visualization and Use Case

- Temporal trends: time evolution view of which factors have been affecting company's performance
- Discovery of companies driven by similar underlying drivers, (especially when companies do not belong to the same sector -> new insight for investment opportunities)
- Prediction: identify leader and laggers affected by the same factors with a time lag.



Some preliminary statistics on samples:

- Based on a sample of 10 randomly selected companies
- a total of 738 financial documents are collected
- 13,983 causal sentences identified
- 5,973 unique noun terms extracted
- These noun terms can be further clustered into [50?] semantically similar concept groups.

Feature selection: marker for causal sentences (explicit causal verbs and non-verb phrases), topic of interests (financial performance)

Unsupervised learning: clustering algorithms (K-Means, Self Organizing Map)

Graph database for storage and query: Neo4j