Midterm Proposal: 10-K filing analysis with NLP

By:
Aditya Natham, Akaash Dash, Sapan Patel, Max Zhao

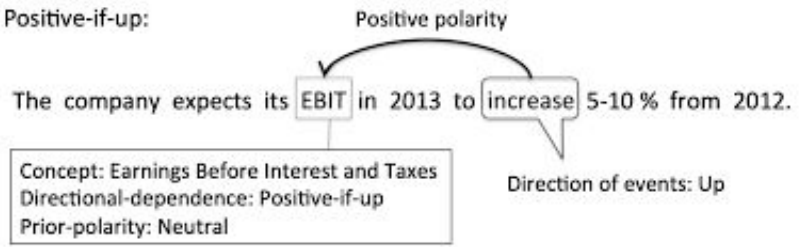
Introduction to SEC 10-K filing

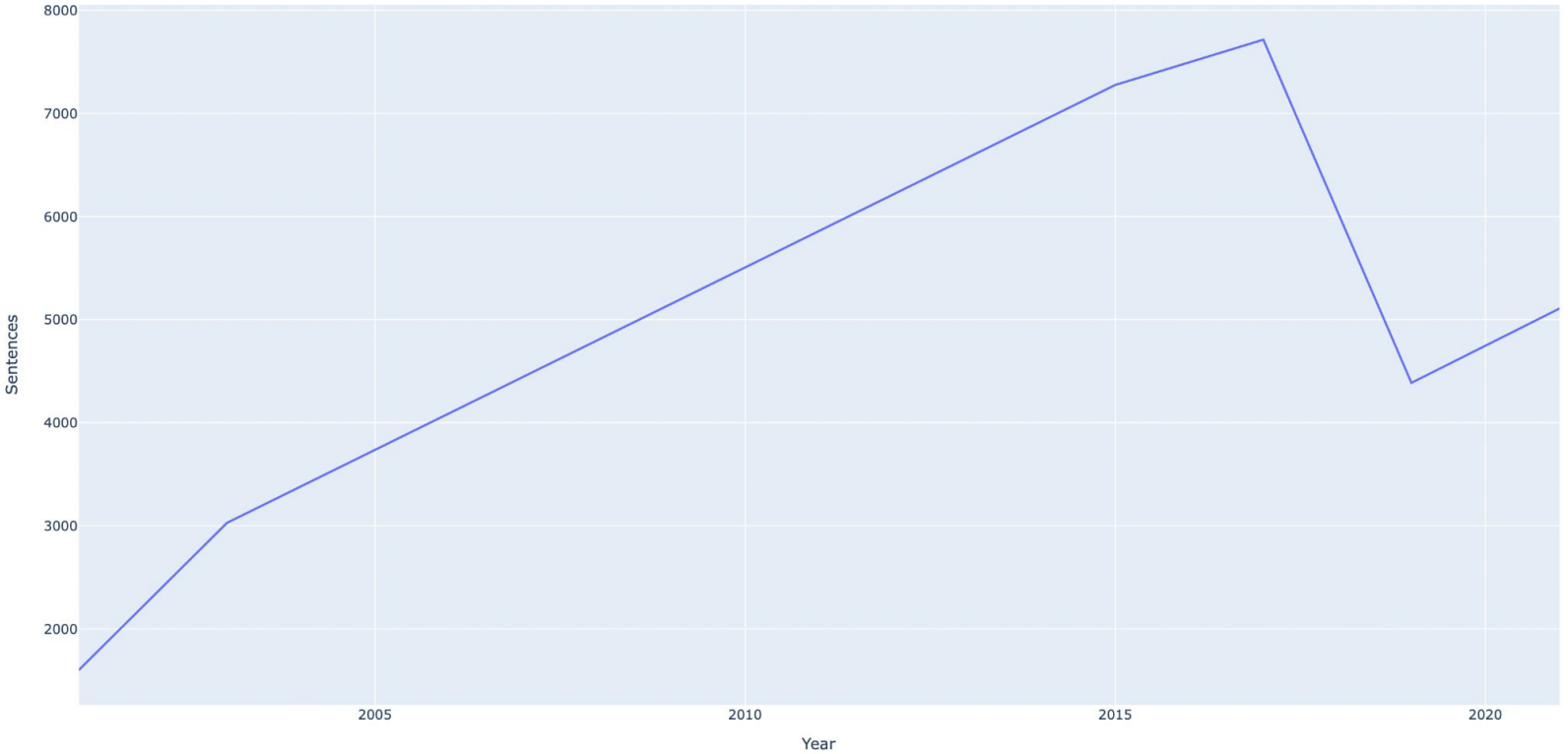
- Every publicly traded company must annually file a 10-k
- Comprehensive report on relevant financial data and risks
- Much too complicated and time consuming for a retail investor to read
- A 10-k is organized into 5 sections: business, risk factors, selected financial data, management discussion and analysis, and financial statements and supplementary data

10-K Structure	
Section	**Purpose**
Business	• Description of the company's history, key business divisions, product/service offerings, and the market(s) it operates in
Risk Factors	• Information regarding the most significant risks to the company, such as new market entrants or the threat of disruption
Management Discussion and Analysis (MD&A)	• Management commentary on the company's fiscal year performance – will address the positive takeaways, plus the mitigating risk factors
Financial Statements	• The audited financial statements of the company, namely the income statement, cash flow statement, and balance sheet
Supplementary Disclosures	• To further clarify the financial statements, the financials are accompanied by a section with footnotes (i.e. full disclosure)

Problem Statement

- There exists an “information gap” between the 10-k and actually understanding the contents
- Over time, companies have started padding their financial filings in order to appeal to investors
- This causes huge bloating and increased difficulty to interpret
- NLP is becoming increasing prevalent in the industry and we apply it to analyze 10-k filings





Example of 10-K padding from JPMorgan Chase

Note: This graph is flawed

Introduction to project - General Overview

1. Ensure we can extract all entities from the 10-k and can produce a word cloud from it
2. Build a website that allows a user to select a ticker and filing year and view that word cloud
3. Create custom labelling functions to extract relationships from the 10-k
4. Integrate the relationships with the entities to build a knowledge graph and place it on the website

5. Eventually use these NLP with ML models to predict potential market movements

End Goal

Semester:

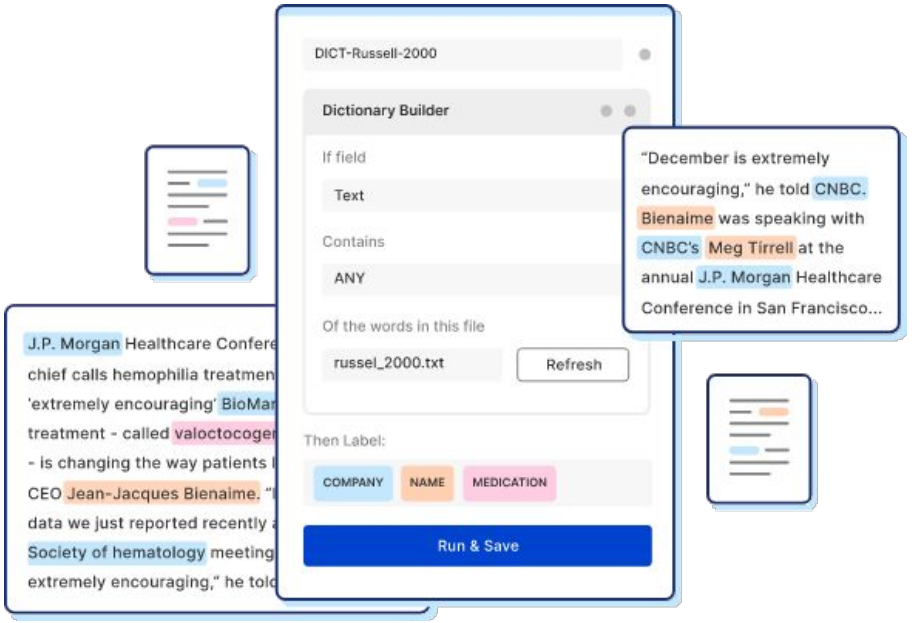
- Build small-scale web page
- User can view an interactive knowledge graph with subset of entities and entity relationships

Overall:

- A tool that retrieves 10-K filing and analyzes information
- User can select any publicly traded company and retrieve filing from the past 25 years
- From the filing, the user can view an interactive knowledge graph with key points and market predictions

Named Entity Recognition

- NLP technique that identifies and categorizes named entities into predefined categories
- Uses machine learning algorithms trained on annotated datasets



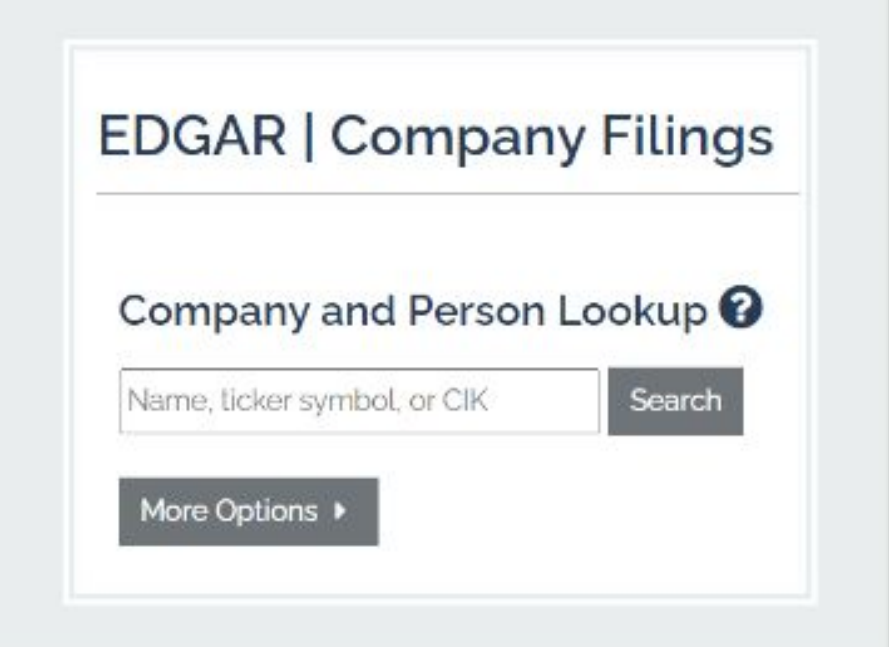
Word cloud

- Provides simple data visualization
- Easy to understand
- Good for users from non-technical background

JP Morgan Annual Report Fiscal Year 2021

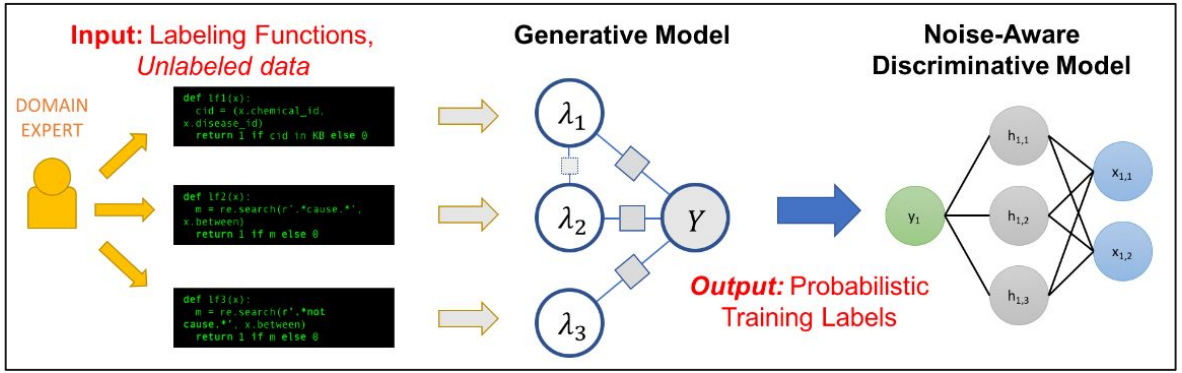
Interactive webpage

- Similar premise to SEC's EDGAR tool but with extended capabilities
- Goal is to provide quick functional data analysis for users



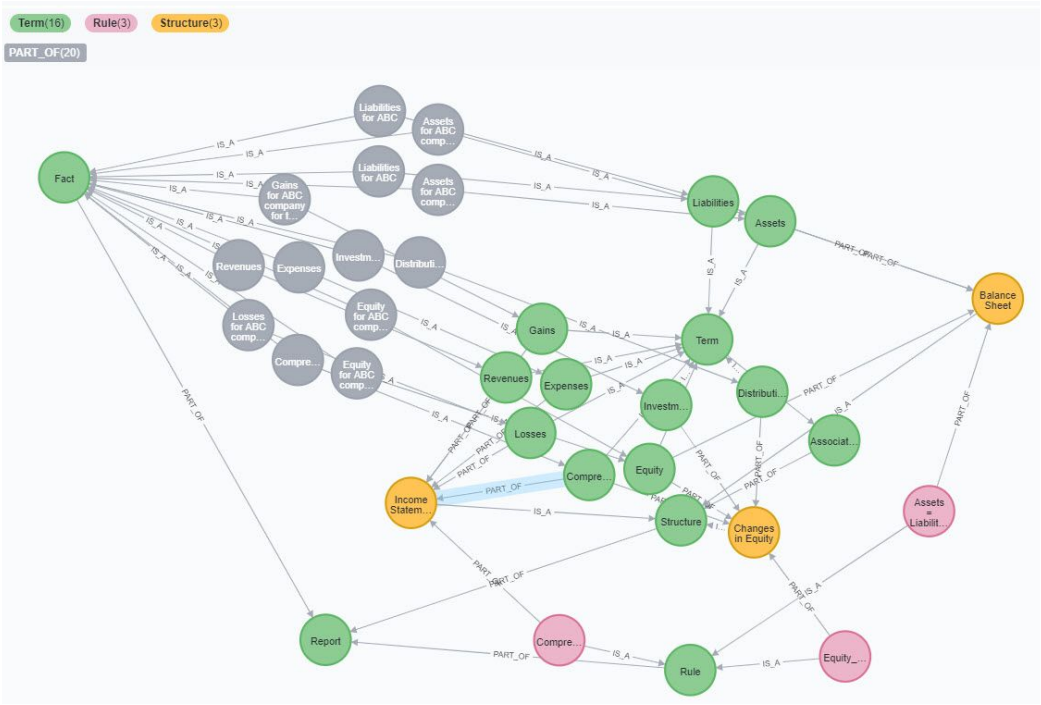
Relationship Labeling Functions

- Identifying and categorizing relationships between named entities
- Done by writing multiple labeling functions
- Long term goal is to use the labeling functions to train weak supervision model



Knowledge Graph

- Culmination of NER and relationship extraction
- Provides visualization for entities and relationships
- Presents data in a more structured format compared to word cloud



Estimated Timeline

Week 1: Introduction to VIP, logistics, and project options
Week 2: Make team and finalize project topic
Week 3: Start draft proposal and finalize project methodologies and project plan
Week 4 - 5: Get familiar with finance, research and understand a 10-k filing, and create project structure
Week 6: Finalize draft proposal
Week 7: Present proposal
Week 8-9: Create an interactive single page website that allows a user to select a company and year, and display a word cloud of the entities extracted from that company's 10-k filing.
Week 9 -10: Create relationship extractor in order to connect the entities in the word cloud to flesh it out into a knowledge graph.
Week 14 : Create an interactive single page website that allows a user to select a company and year, and display a knowledge graph of the entities and relationships extracted from that company's 10-k filing.
Week 15: Create presentation and present project