Introduction

What (if any) information can be extracted from the prose, specifically in the risk disclosure section, of a company's 10K which can be of use to better investment decisions?

Generally speaking, to simulate a collection of stock prices, one needs two inputs: E[Return] & a Covariance Matrix. My project will begin by modeling the covariance matrix using traditional backward-looking data and seeking improvement from these base models by engineering the covariance matrix values (specifically the non-diagonal entries) through textual analysis of company's 10K risk disclosure section.

Finance-Portfolio Optimization & Text Analysis

On Portfolio Management & Optimization:

- PCA in Equity Portfolio Management
- Mean Variance Optimization & CAPM

On Language Analysis in Finance

- 'The Use of Word Lists in Textual Analysis' Journal of Behavioral Finance (2015)
- 'Text-Based Network Industries & Endogenous Product Differentiation' NBER (2012)
- 'When is a Liability Not a Liability' Journal of Finance (2011)

Risk Disclosures

- SEC Update on Risk Disclosures
- Investopedia: Types of Risk Factors
- 'The Benefits of Specific Risk-Factor Disclosures' Review of Accounting Studies (2016)
- 'On the Predictive Ability of Narrative Disclosures in Annual Reports' European Journal of Operational Research (2010)

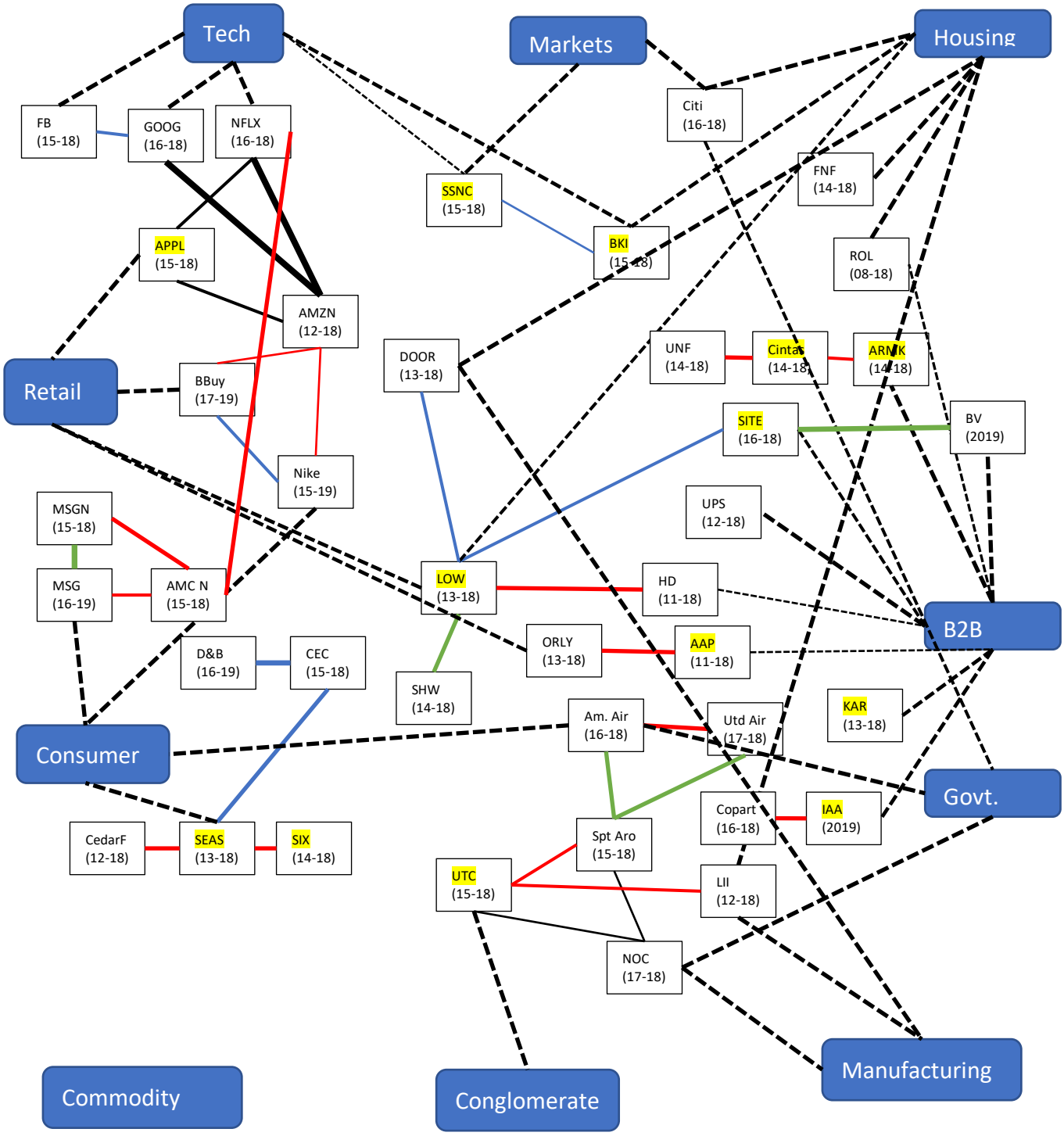
Machine Learning Methodology

- 'Machine Learning in Automated Text Categorization' Sebastini (2002)
- Text Categorization w/ LDA

SEC EDGAR-10Ks & Risk Disclosure

Data Pipeline:

Collected Data to Date:



Libraries/Tools

- Pandas-datareader
- Quandl (Possibly)
- BeautifulSoup
- NLTK

- SciKitLearn
- PyPortfolioOpt
- https://github.com/DarthQadir/Natural-Language-Processing-with-LDA-and-Text-Clustering/blob/main/NLP_LDA_Text_Clustering.ipynb

Methodologies

Database storage—may make sense to build relational database.

TARGET METRIC: Mean Absolute Percentage Error

- Traditional Risk Assessment & Portfolio Building:
  - PCA
  - Factor Analysis
  - Mean-Variance Optimization
- Language Analysis
  - Vectorization Strategies:
  - Simple Modeling Options:
  - Latent Dirichlet Allocation

Expected Work Flow

- Collect Raw Data
  - Stock Prices
  - Basic Company Information
  - 10K files
- Standard Portfolio Builds
- Process 10Ks: Grab Risk Disclosure Section & Organize into Headline:Detail
- Process Text
- Text-Based Models
- Apply to Portfolio Management