# Computational Analysis of Self-Reported Business Risk Factors

Samuel Grayson, Advisor: Vincent Ng

## Introduction

SEC requires businesses to release 10-K forms including:

- Financial data
  - Quantitative
  - Already well studied
- Risk Factors
  - Plain text essays
  - Less well studied due to unstructured nature

Example: “Successful operation… may be disrupted by… terrorism, cybersecurity attacks, and other local security concerns”
— Exxon Mobil

## Goal

1. Build a distributed system
2. Implement textmining algorithms
3. Extract insights from Risk Factors section for investors and researchers

This is difficult because:

- 10-K forms are human-readable but not machine-readable
- Big data: 10 mb / firm, 10,000 firms / year, 20 years = 2Tb
- Needs to be inexpensive

## Distributed System

- Google Cloud provides cheap on-demand compute resources
- Network-virtualized services in Kubernetes provides fault tolerance and flexible mapping to physical hardware
- Map/reduce model makes cluster easy to program

## Text mining AI

- Model each document as a bag of word-stems
- Use Porter Stemmer
  - fishing, fished, fisher ⇨ fish
- Tf/idf weights words by importance
- Learn a topic model on those word-stems
  - Operates by matrix factorization, not knowing meaning of words

| | “monetary” | “banking” | “tornado” | “hurricane” |
|---|---|---|---|---|
| doc #1 | 0 | 0 | 3 | 4 |
| doc #2 | 2 | 2 | 3 | 2 |
| doc #3 | 4 | 3 | 1 | 0 |

=

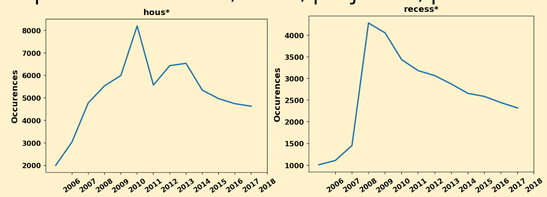
| | topic #1 | topic #2 |
|---|---|---|
| doc #1 | 1 | 0 |
| doc #2 | 1 | 1 |
| doc #3 | 0 | 1 |

x

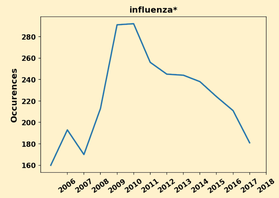
| | “monetary” | “banking” | “tornado” | “hurricane” |
|---|---|---|---|---|
| topic #1 | 0 | 0 | 3 | 3 |
| topic #2 | 3 | 3 | 0 | 0 |

## Results

Automatically identified topics:
Topic #2: patent, intellectual, technological, infringement
Topic #145: investor, asset, projected, portfolio




Terms related to housing crisis and recession go up during the crisis, and appear to have reached a new steady state. Perhaps the risk has always been at that level, but was previously unacknowledged.




Compare to Influenza which spiked around the discovery of Swine Flu in humans in 2009, but went back down. Perhaps there is no unacknowledged risk here, and the event was more

## Improvements

- Vectorize with neural word embeddings instead of word-counts
- LDA topic modeling instead of LSI
- Include N-gram tokens
- Use word-sense disambiguation

## Applications

- For investors: firms with orthogonal risk vectors might have uncorrelated returns (a diverse portfolio)
- For social scientists: investigate how topics change over time quantitatively
  - Apply same system to other corpora such as political news

## Conclusion

- Created cost-effective, fast, easy-to-program distributed system.
- Implemented state-of-the-art text mining

## References and acknowledgements

Shayan Monadjemi, Jack Sollows, and Micheal Seeligson guided me in developing this project.

15 U.S.C. § 78m. Oct. 2018.
Dask Development Team. Dask: *Library for dynamic task scheduling.* 2016.
*Learning Research* 3 (Jan. 2003), pp. 993–1022
Radim Řehůřek and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora.”
English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.*
John Gerdes Jr. “EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC’s
EDGAR database.” In: *Decision Support Systems* 35 (1 Apr. 2003), pp. 7–29.
Kay M. Nelson et al. “Virtual auditing agents: The edgar agent challenge.” In: *Decision Support*
*Systems* 28 (3 May 2000), pp. 241–253.
Peter Willett. “The Porter stemming algorithm: then and now.” In: *Program* 40 (3 July 2006), pp. 219–223.