Capstone Project 2: Topic Model Analysis of Berkshire Hathaway’s Annual Letters to Shareholders

Springboard Capstone Project II

Tom Halloin

Advice from Warren Buffett

“Read 500 pages every day. That's how knowledge works. It builds up, like compound interest. All of you can do it, but I guarantee not many of you will do it.”

The Problem

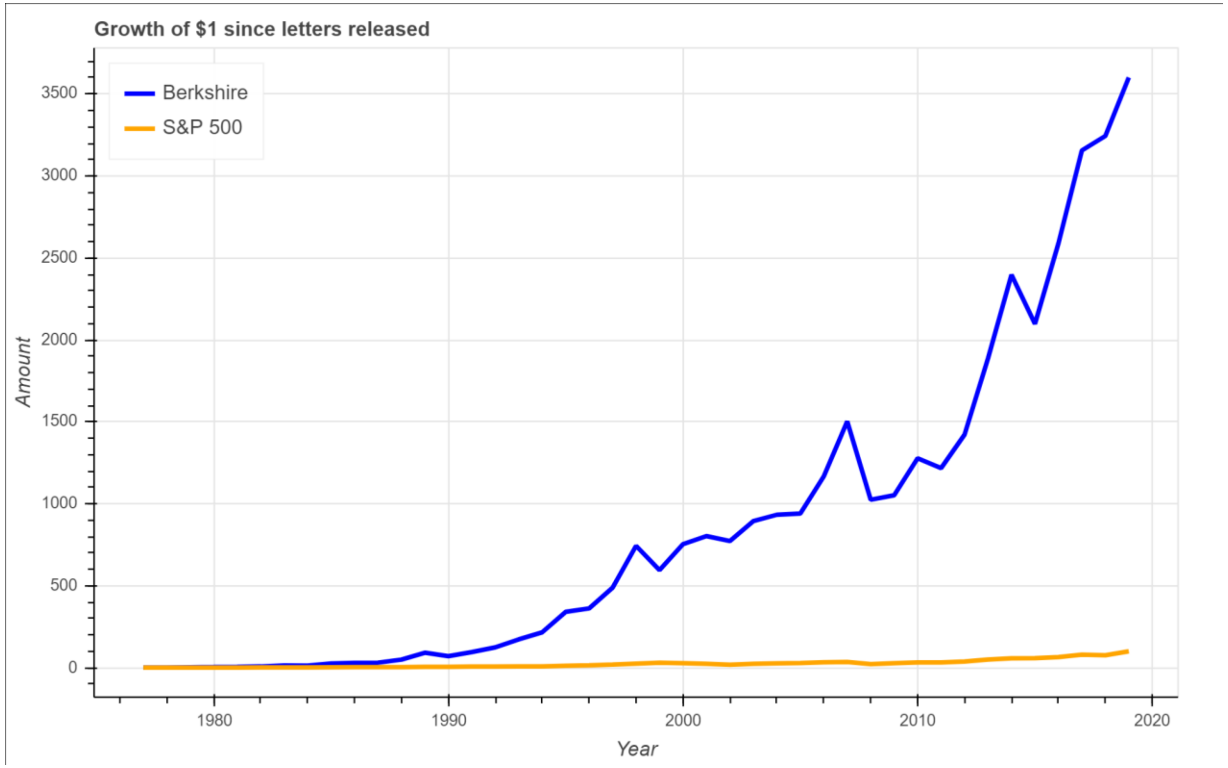
- Financial literature is long, confusing, and boring to most people.

- Solution: summarize documents and build a topic model using machine learning.

[Source: The Financial Brand]

Case study: Berkshire Hathaway

- Berkshire Hathaway is a holding company run by Warren Buffett.
- Berkshire's returns have trounced the S&P 500 since releasing letters in 1977.
- Dataset: 43 letters, 500,000 words, readability level of 58.



Exploration Libraries: SpaCy and Textacy

- SpaCy is an open-sourced library used to tokenize, parse, and tag text data.

- Textacy is built on Spacy and does pre- and post- processing for Spacy, including cleaning text, generating text statistics, and creating n-grams.

Acquiring, Cleaning, and Wrangling the Data

Letters dating back to 1977 are on Berkshire’s website in HTML and PDF files.

Challenges:

- Reading files from different formats.
- Splitting PDF files in order to parse information.
- Getting past spam blockers for multiple years.

Cleaning Letters for Humans

- Removing HTML
- Removing text not in original letters
- Removing symbols
- UTF-8 encodings
- Tabular data

```
<!-- Global site tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=UA-136883390-1"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());

  gtag('config', 'UA-136883390-1');
</script>
<HTML>
<HEAD>
  <TITLE>Chairman's Letter - 1977</TITLE>
</HEAD>
<BODY>
<P ALIGN=CENTER>
<B>BERKSHIRE HATHAWAY INC.</B>
</P>
<PRE>

<I>To the Stockholders of Berkshire Hathaway Inc.:</I>
```

Summarizing the Documents

- Three different summarizers: LexRank, TextRank, LSA.

- Use each method to find the top 5 sentences for each year and compare across years and summarization methods.

Comparing Summarizations: LexRank

- Finds the most relevant sentences by using weighted cosine similarities of TF-IDF vectors.

- Graph based ranking model similar to Google's PageRank algorithm.

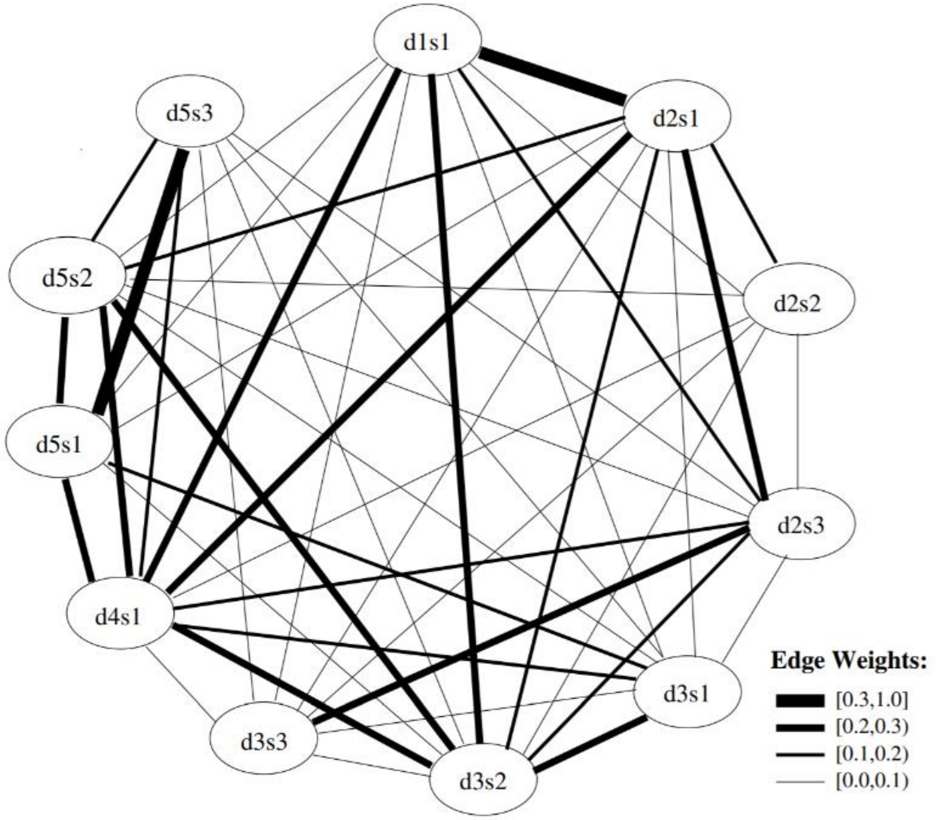


Figure 2: Weighted cosine similarity graph for the cluster in Figure 1.

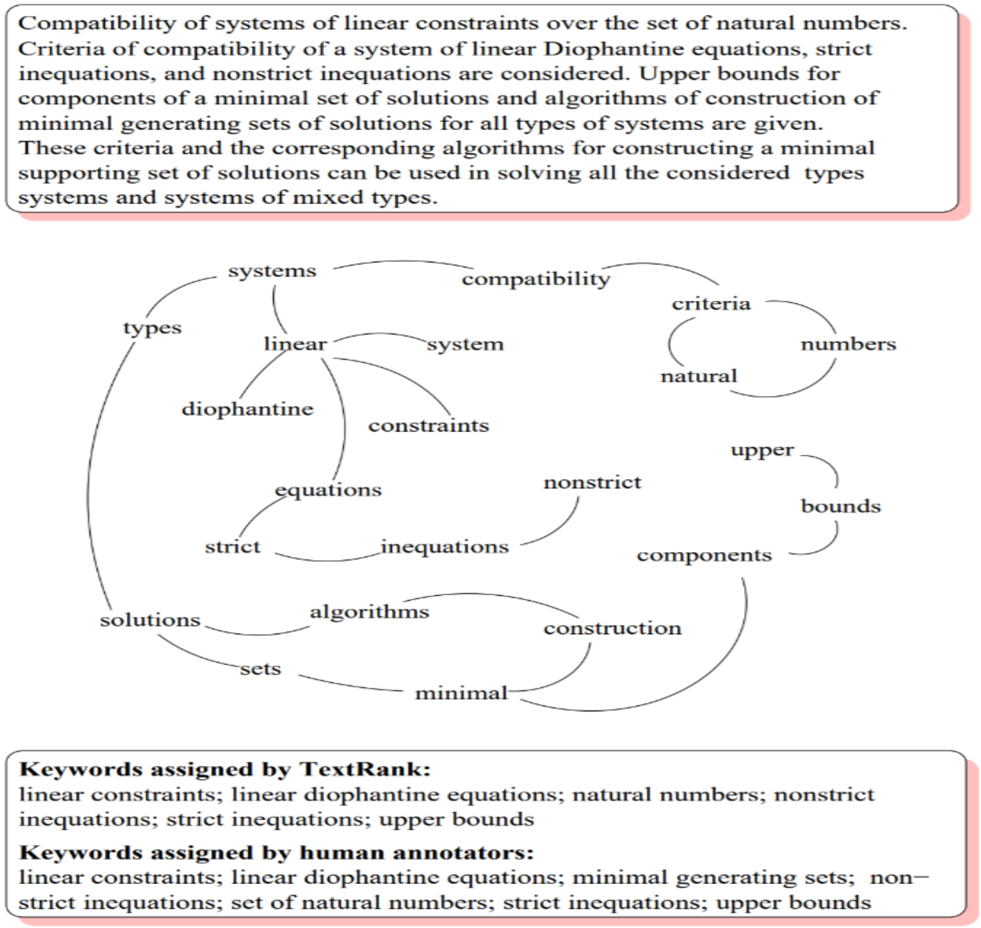
LexRank, the ugly

- “Berkshire's Share of Undistributed Berkshire's Approximate Operating Earnings Berkshire's Major Investees Ownership at Yearend (in millions) 1993 1992 1993 1992 Capital Cities/ABC, Inc. 13.0% 18.2% $ 83(2) $ 70 The Coca-Cola Company 7.2% 7.1% 94 82 Federal Home Loan Mortgage Corp. 6.8%(1) 8.2%(1) 41(2) 29(2) GEICO Corp. 48.4% 48.1% 76(3) 34(3) General Dynamics Corp. 13.9% 14.1% 25 11(2) The Gillette Company 10.9% 10.9% 44 38 Guinness PLC 1.9% 2.0% 8 7 The Washington Post Company 14.8% 14.6% 15 11 Wells Fargo & Company 12.2% 11.5% 53(2) 16(2)” – 1993 letter

Comparing Summarizations: TextRank

- Almost the same as LexRank, with the exception of different weights (log based).

- Weighs sentences based on the number of words two sentences have in common.



TextRank, the ugly

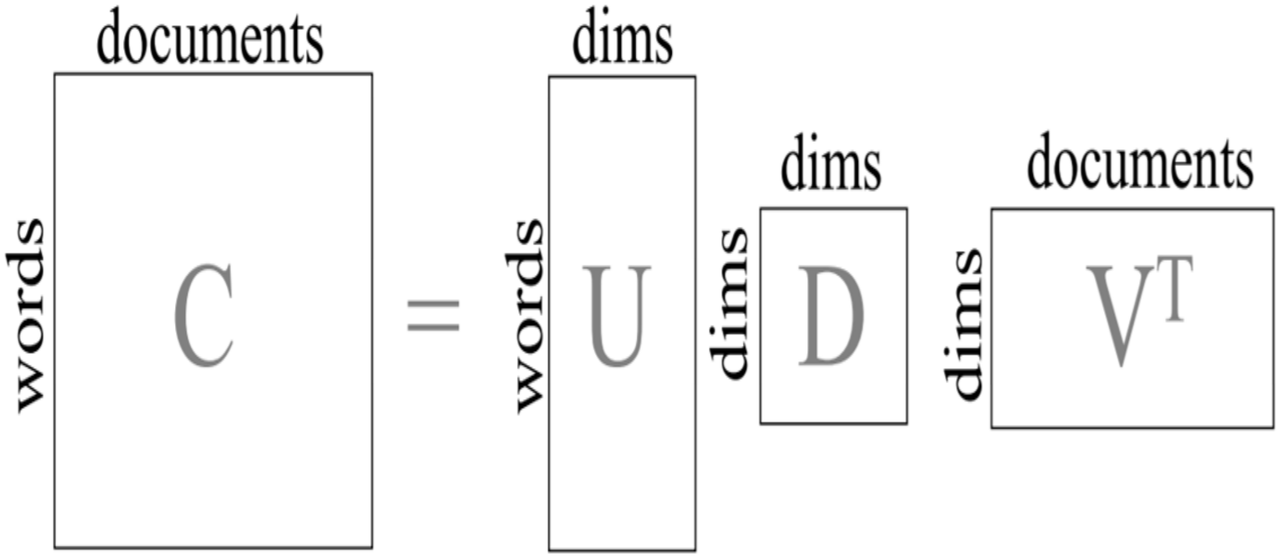
- Loading time was by far the longest of the three methods.

- “Berkshire's Share of Undistributed Berkshire's Approximate Operating Earnings Berkshire's Major Investees Ownership at Yearend (in millions) 1991 1990 1991 1990 Capital Cities/ABC Inc. 18.1% 17.9% $ 61 $ 85 The Coca-Cola Company 7.0% 7.0% 69 58 Federal Home Loan Mortgage Corp. 3.4%(1) 3.2%(1) 15 10 The Gillette Company 11.0% 23(2) GEICO Corp. 48.2% 46.1% 69 76 The Washington Post Company 14.6% 14.6% 10 18 Wells Fargo & Company 9.6% 9.7% (17) 19(3) Berkshire's share of undistributed earnings of major investees $230 $266 Hypothetical tax on these undistributed investee earnings (30) (35) Reported operating earnings of Berkshire 316 371 Total look-through earnings of Berkshire $516 $602 (1) Net of minority interest at Wesco (2) For the nine months after Berkshire converted its preferred on April 1 (3) Calculated on average ownership for the year We also believe that investors can benefit by focusing on their own look-through earnings.”

Comparing Summarizations: LSA

- LSA creates a term-document matrix consisting of word frequencies for each term in each document.

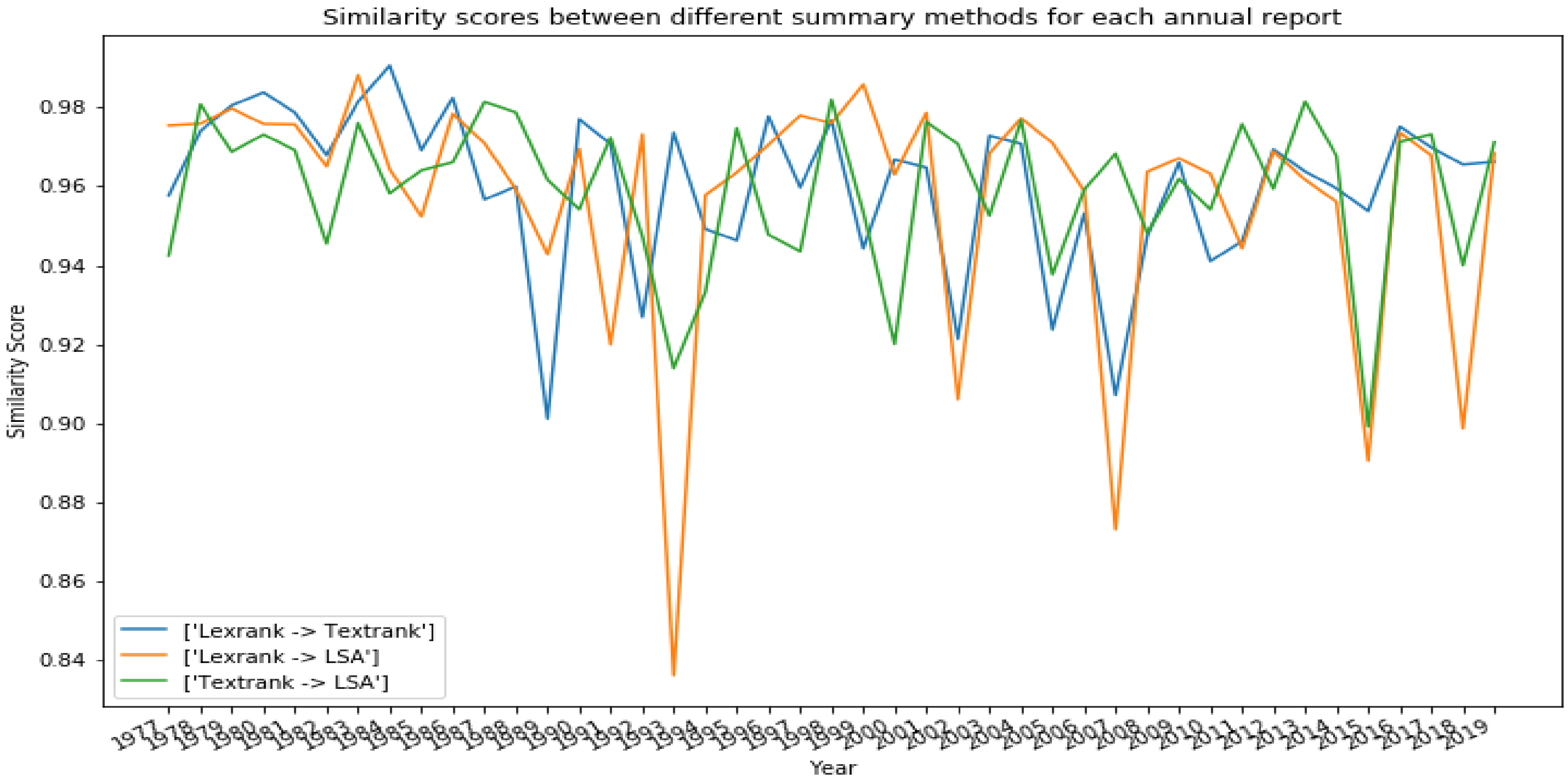
- Uses singular value decomposition to find most important sentences.



LSA, the ugly

- “You only learn who has been swimming naked when the tide goes out and what we are witnessing at some of our largest financial institutions is an ugly sight.”
  - – 2007 letter

Overall, summaries were fairly similar and did a good job summarizing.



Selecting Appropriate Algorithms

- This is an **unstructured** dataset.
- We are not trying to predict categories or quantities.
- In this case, look for clustering algorithms.
- For text data, this means topic modeling.

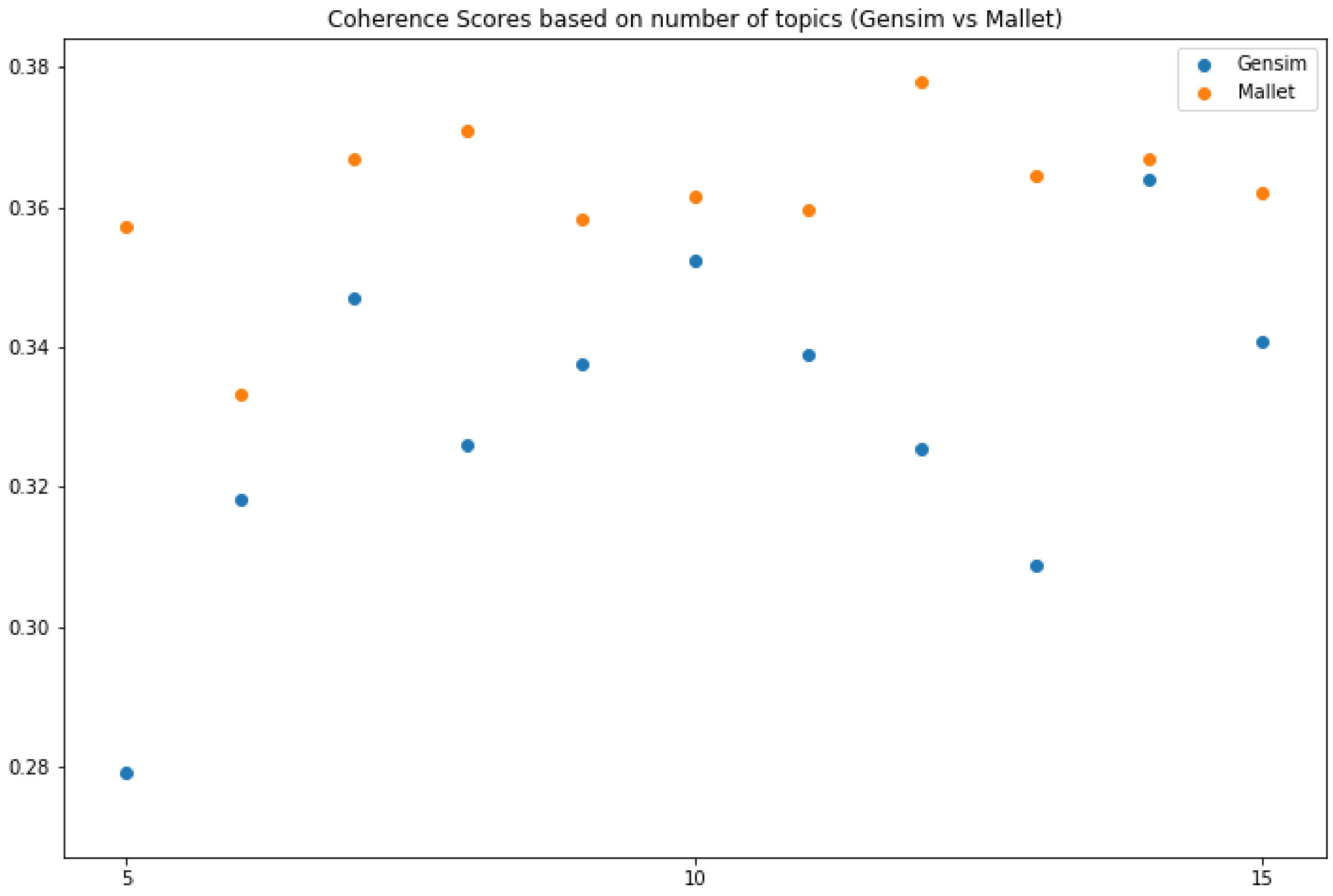
Topic Modeling Methods and Libraries

- Latent Dirichlet Allocation (LDA) tries to separate sets of observations into groups to explain similarity within the groups.

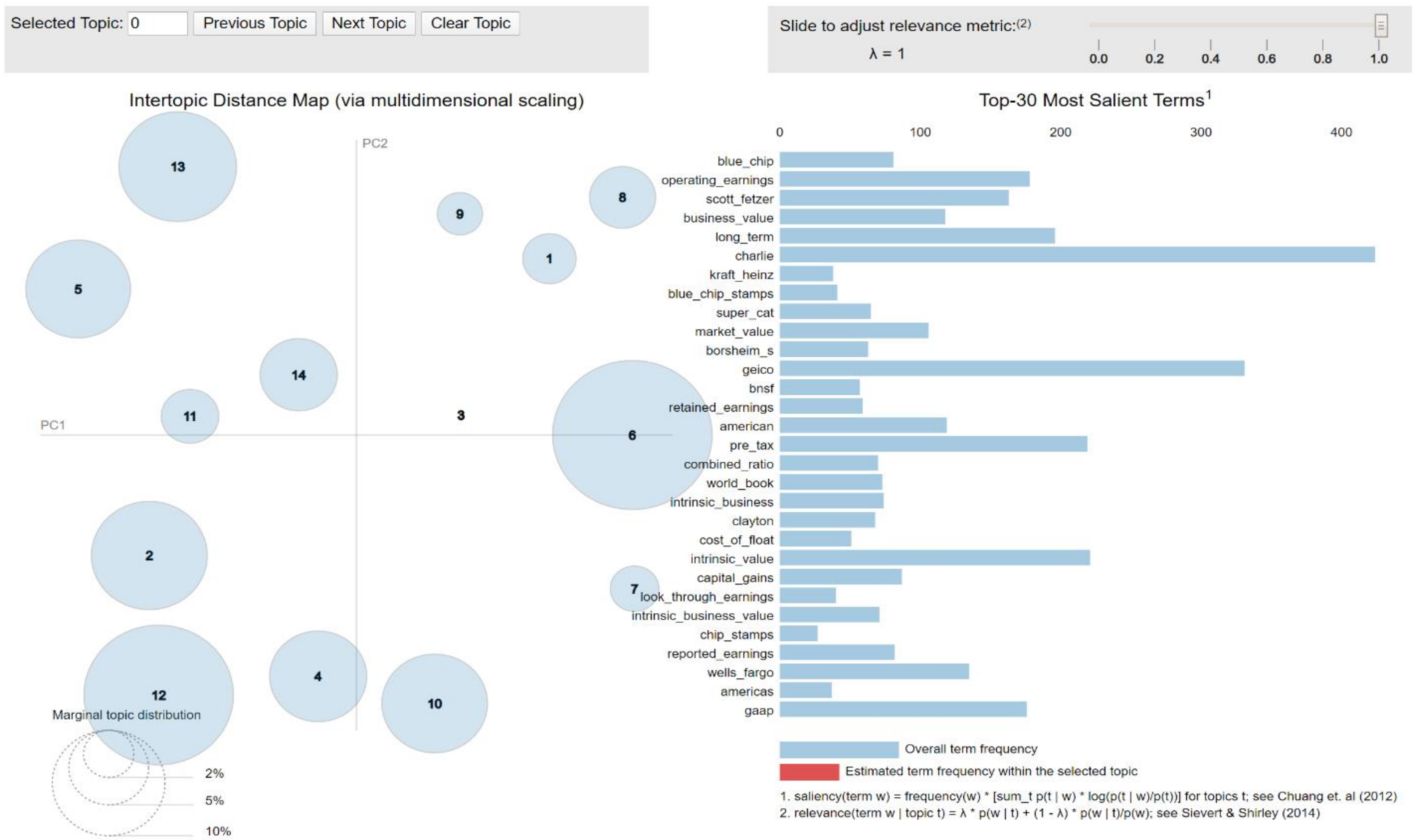
- Gensim LDA is an LDA implementation designed to provide an approximation to large datasets.

- Mallet LDA is similar to Gensim, but uses Gibbs Sampling to create evenly-distributed topics. Needs Java for extra processing power.

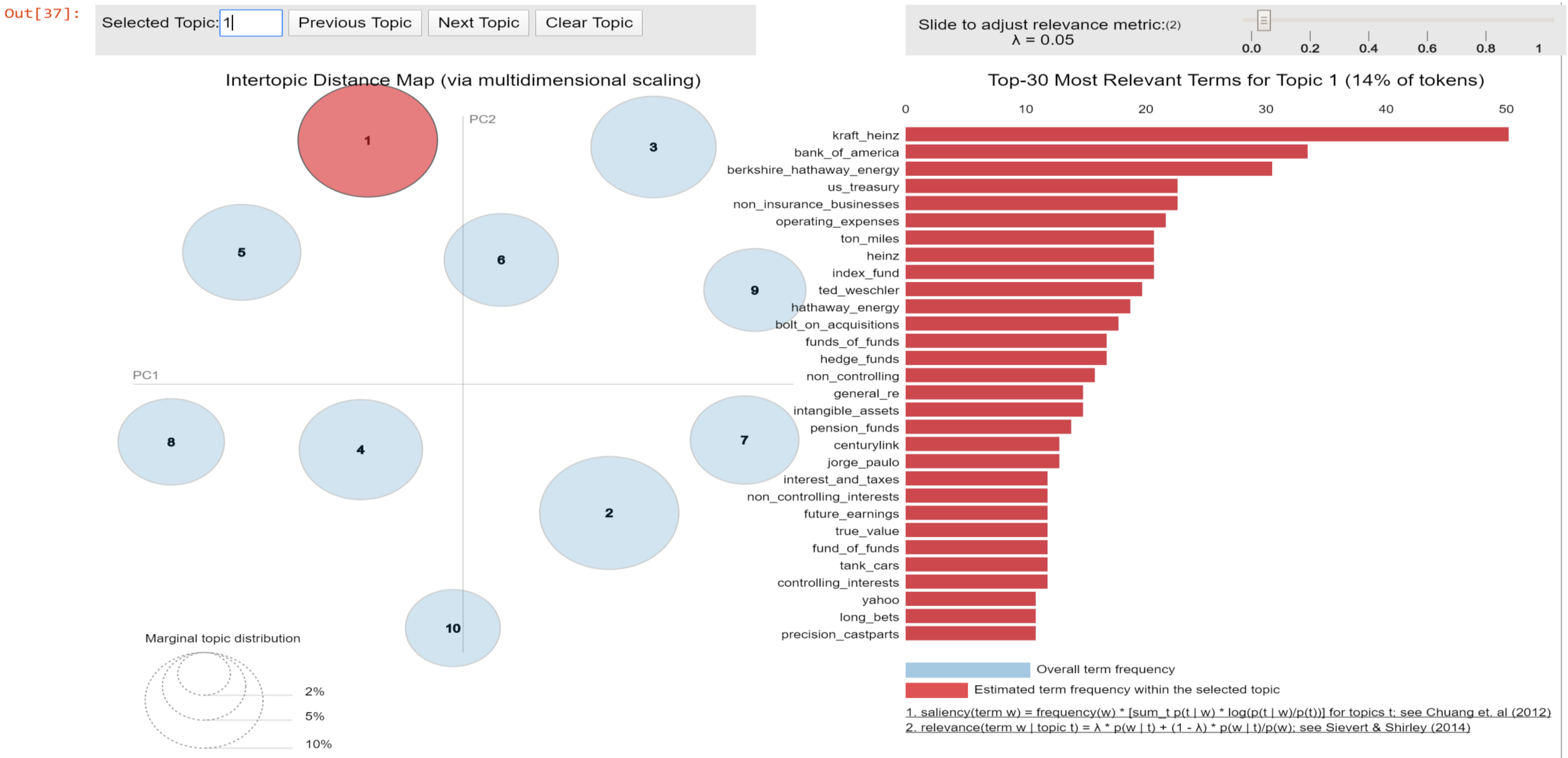
Measure Topic Quality: Coherence Score



Sample Topic Model - GensimLDA

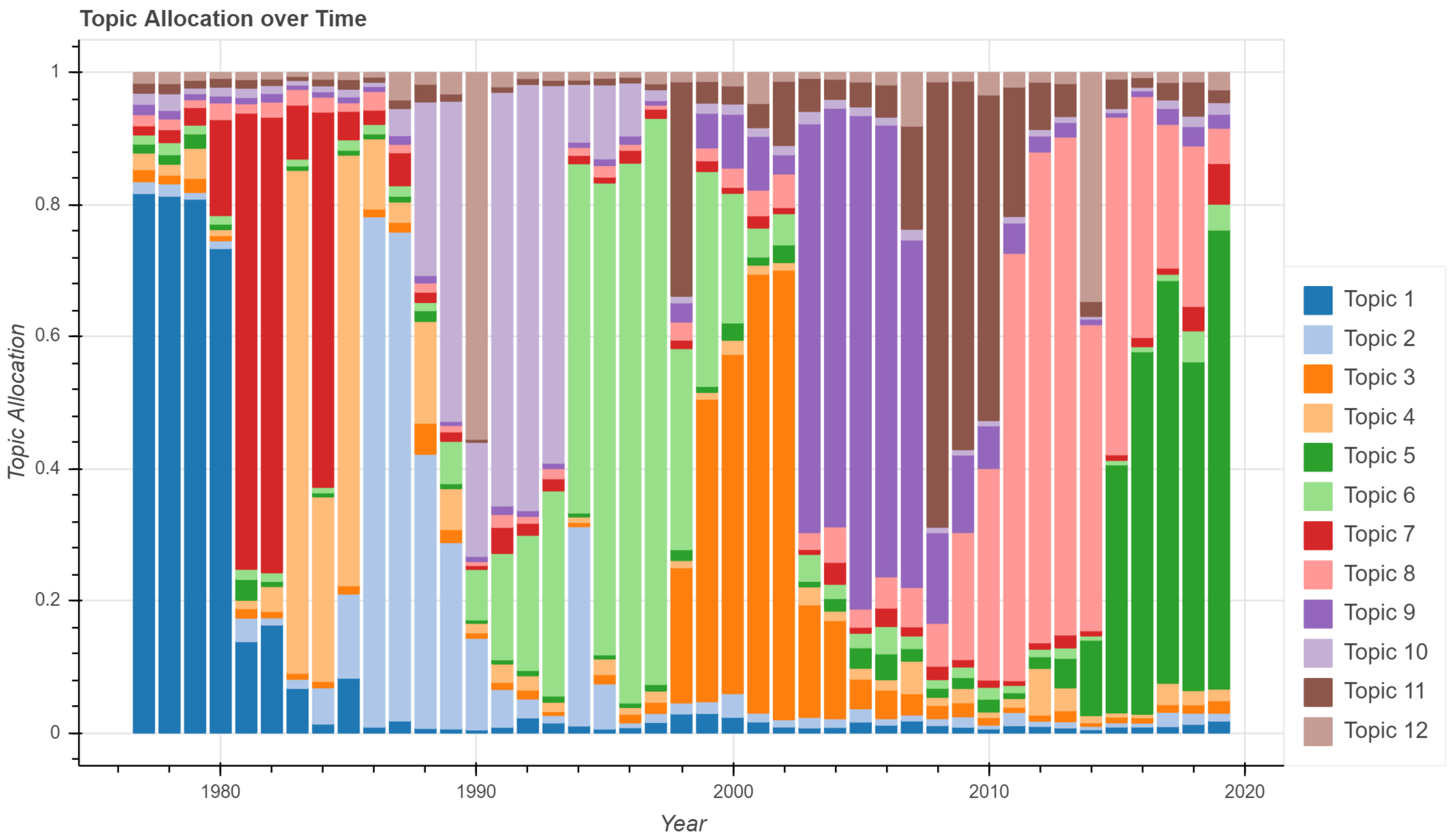


Why MalletLDA is Better



Mallet Topic Words (Lambda = .05)

Topic 1	Topic 2	Topic 3	Topic 4
Wesco Financial	Ralph Schey	Fractional Ownership	WPPSS
Illinois National	Scott Fetzer's	General Re	News
Earnings Per Share	Ralph	September 11th	Courier Express
Phil Liesche	Fechheimer	Executive Jet	The News
National Bank	Chuck	Owners Manual	Stan
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
Kraft Heinz	Major Investees	Non-controlled	BNSF
Hathaway Energy	Cat Business	Good Businesses	Marmon
Fund of Funds	Super Cat Business	Controlled Businesses	Operating Expenses
Hedge Funds	Geico's	Buffalo Evening	ton miles
Berkshire Hathaway Energy	Earnings Reported	Unusual Sales	Heinz
Topic 9	**Topic 10**	**Topic 11**	**Topic 12**
R.C. Willey	HH Brown	General Res	Stock Prices
New Jersey	Cost of Funds	Compounded annually	Contingency Reserve
Growth Rate	Preferred Stock	Black Scholes	Berkshire System
Electric Customers	H Brown	Kern River	Board of Directors
Qwest	RJR Nabisco	General Electric	Years Later


Topic Allocation over Time
Topic Allocation
1
0.8
0.6
0.4
0.2
0
1980
1990
2000
2010
2020
Year
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
Topic 9
Topic 10
Topic 11
Topic 12

Conclusion – Annotating the Topics

- Topic 1: Early Holdings: Banks and Stamps
- Topic 2: Scott Fetzer Company
- Topic 3: Aviation Businesses
- Topic 4: Bond Defaults and Newspapers
- Topic 5: Massive Funds and Businesses
- Topic 6: Insurance Underwriting
- Topic 7: Insurance Companies
- Topic 8: Industrial Holdings
- Topic 9: Annual Meeting at the Qwest
- Topic 10: Cowboy Boots, Junk Bonds and Mortgages
- Topic 11: Electricity and Reinsurance
- Topic 12: US Steel

Next Steps

- Applying techniques to new documents (10-K)
- Deep learning and abstractive summarization
- Sentiment Analysis

Special thanks to Liang Kuang from Springboard for his assistance and mentorship.