Understanding financial reports using topic models

Presented by Cesar Osorio
December 11, 2015

Agenda

1. Data
2. Methodology
3. Topic Modeling
4. What topics did I find?
5. How topics evolve over time?
6. How similar are the financial reports if we delete the most distinctive topics?

Financial Reports

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION

Washington, D.C. 20549

FORM 10-K

.3 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
14

or

)N 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from to

Commission file number:
1-6523

Exact name of registrant as specified in its charter:

Bank of America Corporation

State or other jurisdiction of incorporation or organization:
Delaware
IRS Employer Identification No.:
56-0906609
Address of principal executive offices:
Bank of America Corporate Center
100 N. Tryon Street
Charlotte, North Carolina 28255
Registrant's telephone number, including area code:
(704) 386-5681
Securities registered pursuant to section 12(b) of the Act:

Part I

Bank of America Corporation and Subsidiaries

Item 1. Business

Bank of America Corporation (together, with its consolidated subsidiaries, Bank of America, we or us) is a Delaware corporation, a bank holding company (BHC) and a financial holding company. When used in this report, "the Corporation" may refer to Bank of America Corporation individually, Bank of America Corporation and its subsidiaries, or certain of Bank of America Corporation's subsidiaries or affiliates. As part of our efforts to streamline the Corporation's organizational structure and reduce complexity and costs, the Corporation has reduced and intends to continue to reduce the number of its corporate subsidiaries, including through intercompany mergers.

Bank of America is one of the world's largest financial institutions, serving individual consumers, small- and middle-market businesses, institutional investors, large corporations and governments with a full range of banking, investing, asset management and other financial and risk management products and services. Our principal executive offices are located in the Bank of America Corporate Center, 100 North Tryon Street, Charlotte, North Carolina 28255.

Bank of America's website is www.bankofamerica.com. Our Annual Reports on Form 10-K, Quarterly Reports on Form 10-Q, Current Reports on Form 8-K and amendments to those reports filed or furnished pursuant to Section 13(a) or 15(d) of the Securities Exchange Act of 1934 (Exchange Act) are available on our website at http://investor.bankofamerica.com under the heading Financial Information SEC Filings as soon as reasonably practicable after we electronically file such reports with, or furnish them to, the U.S. Securities and Exchange Commission (SEC). In addition, we make available on http://investor.bankofamerica.com under the

Noninterest Income

Table 3 Noninterest Income

(Dollars in millions)	2014	2013
Card income	$ 5,944	$ 5,826
Service charges	7,443	7,390
Investment and brokerage services	13,284	12,282
Investment banking income	6,065	6,126
Equity investment income	1,130	2,901
Trading account profits	6,309	7,056
Mortgage banking income	1,563	3,874
Gains on sales of debt securities	1,354	1,271
Other income (loss)	1,203	(49)
Total noninterest income	**$ 44,295**	$ 46,677

Data

Financial reports of :

- Bank of America (BOA)

- J.P. Morgan (JPM)

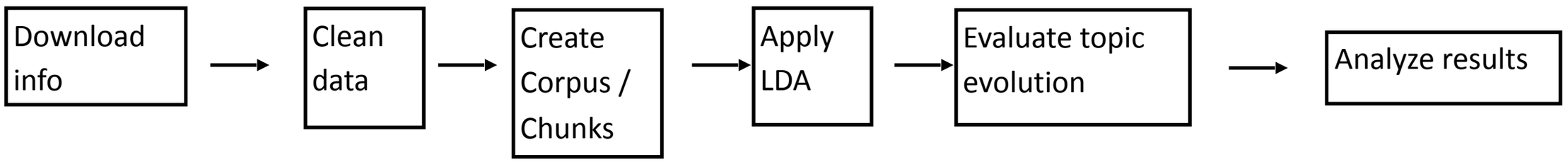
- Citibank (CITI)

Period: 2004 - 2014

Total: 33 reports / each with 70k - 100k words

This information available at the Securities and Exchange Commission (SEC)'s website : https://www.sec.gov/edgar/searchedgar/companysearch.html

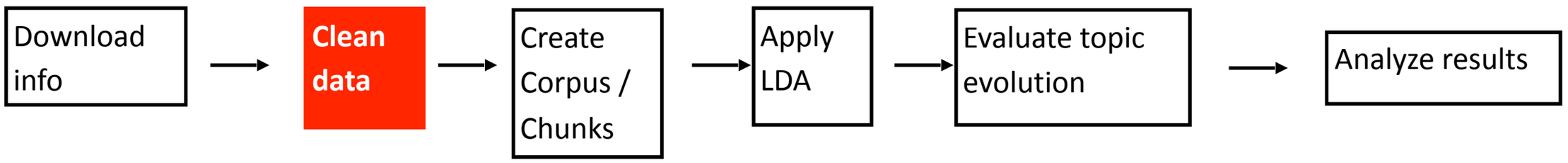
Methodology



80% of time

20% of time

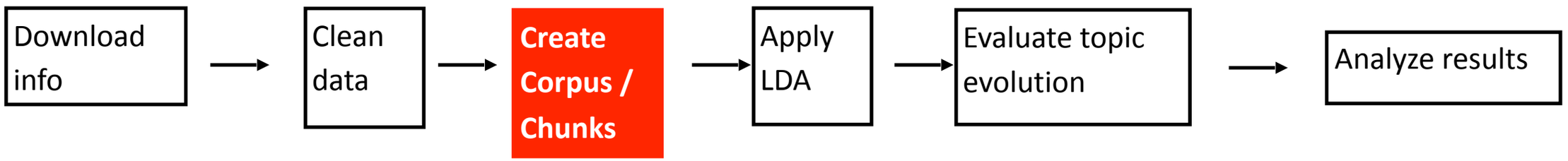
Methodology



Step 2: Eliminate non-relevant information in the financial reports (clean data)

1. Delete HTML tags , XML code and other tags using the “Beautiful Soup” Python library
2. Eliminate non-letters (numbers and symbols for example)
3. Delete words of 1 or 2 characters
4. Transform all the documents to lower case
5. Remove stop words

Methodology

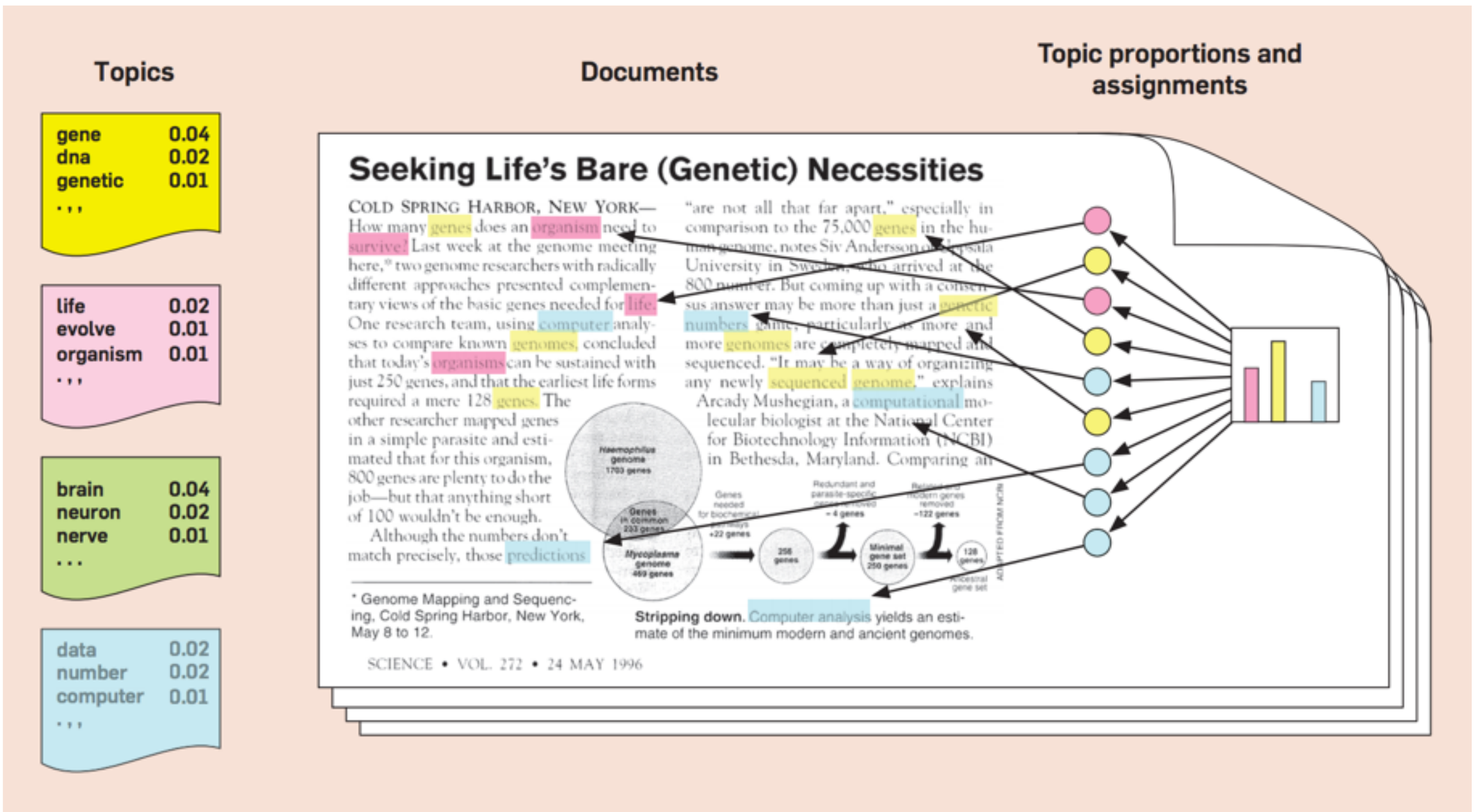


Step 3: Use NLP techniques to transform documents (create corpus)

Using “SciKit Learn” library in Python:

1. Using the cleaned documents from Step 2, read through the 33 reports and create a chunk every 1,000 words (each financial report is split to 70 to 100 chunks for a total of 3,049 chunks).
2. Transform the chunks to a dtm matrix, but eliminate words that are not at least in 30 chunks or 1% of the total chunks (this process deletes 7000 words approximately) and also words that occur in 70% or more of the chunks (only 33 words are deleted)

Topic Model

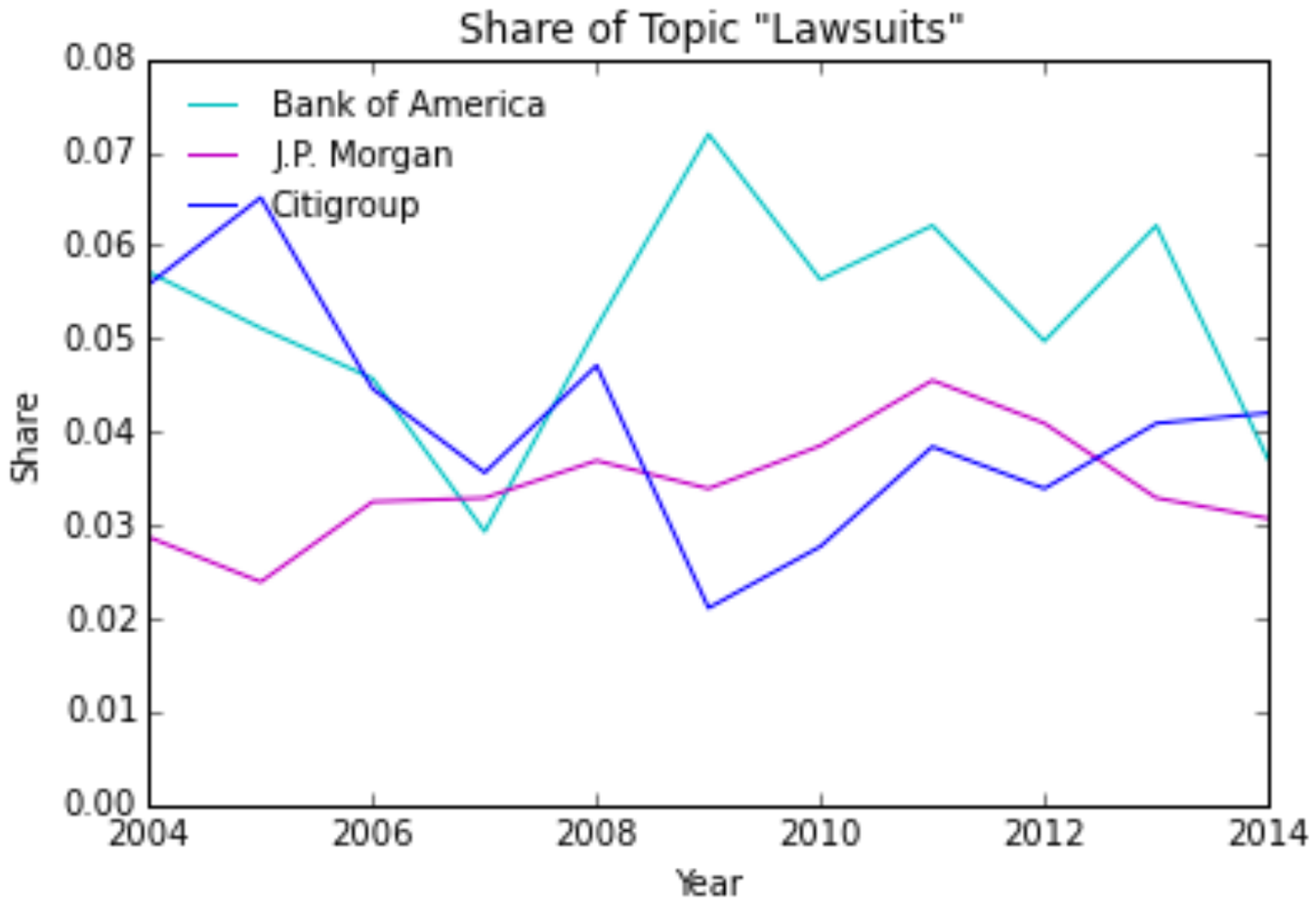


Source: "Probabilistic Topic Models" (Article) by David Blei (2012). http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

What topics did I find?

“Mortgage-backed securities”	“Risks”	“Growth & Revenues”	“Expected credit losses”	“Derivatives”
mortgage	could	million	allowance	fair
repurchase	risks	increased	consumer	debt
servicing	businesses	banking	portfolio	trading
securitization	operational	revenue	mortgage	liabilities
claims	regulatory	higher	card	derivative
representations	significant	services	impaired	level
warranties	operations	expense	past	instruments
corporation	affect	investment	home	derivatives
securitizations	economic	offset	days	rate
loss	results	primarily	principal	mortgage

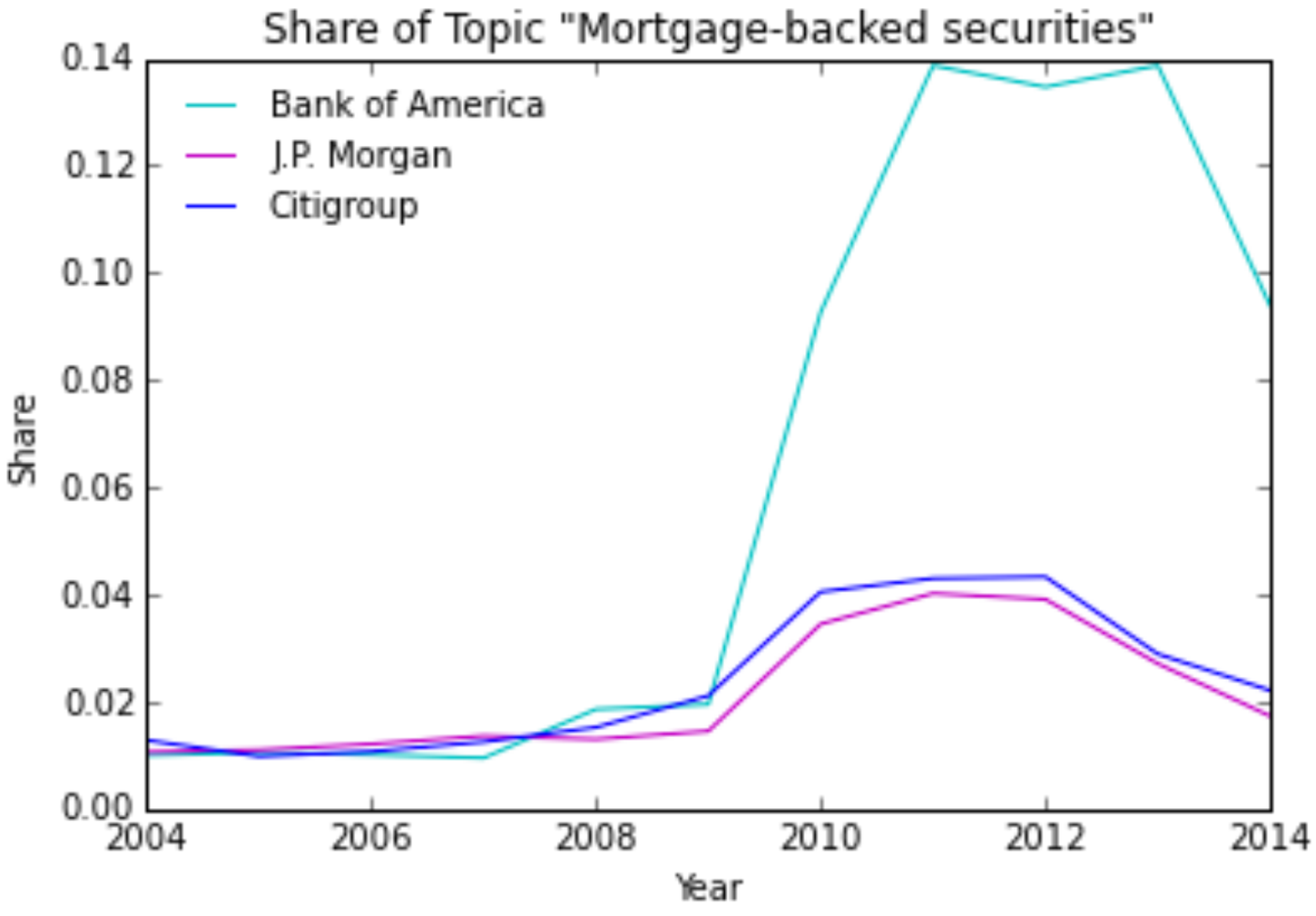
How topics evolve over time?



“Lawsuits”

court
district
claims
actions
filed
action
corporation
defendants
class
plaintiffs

How topics evolve?



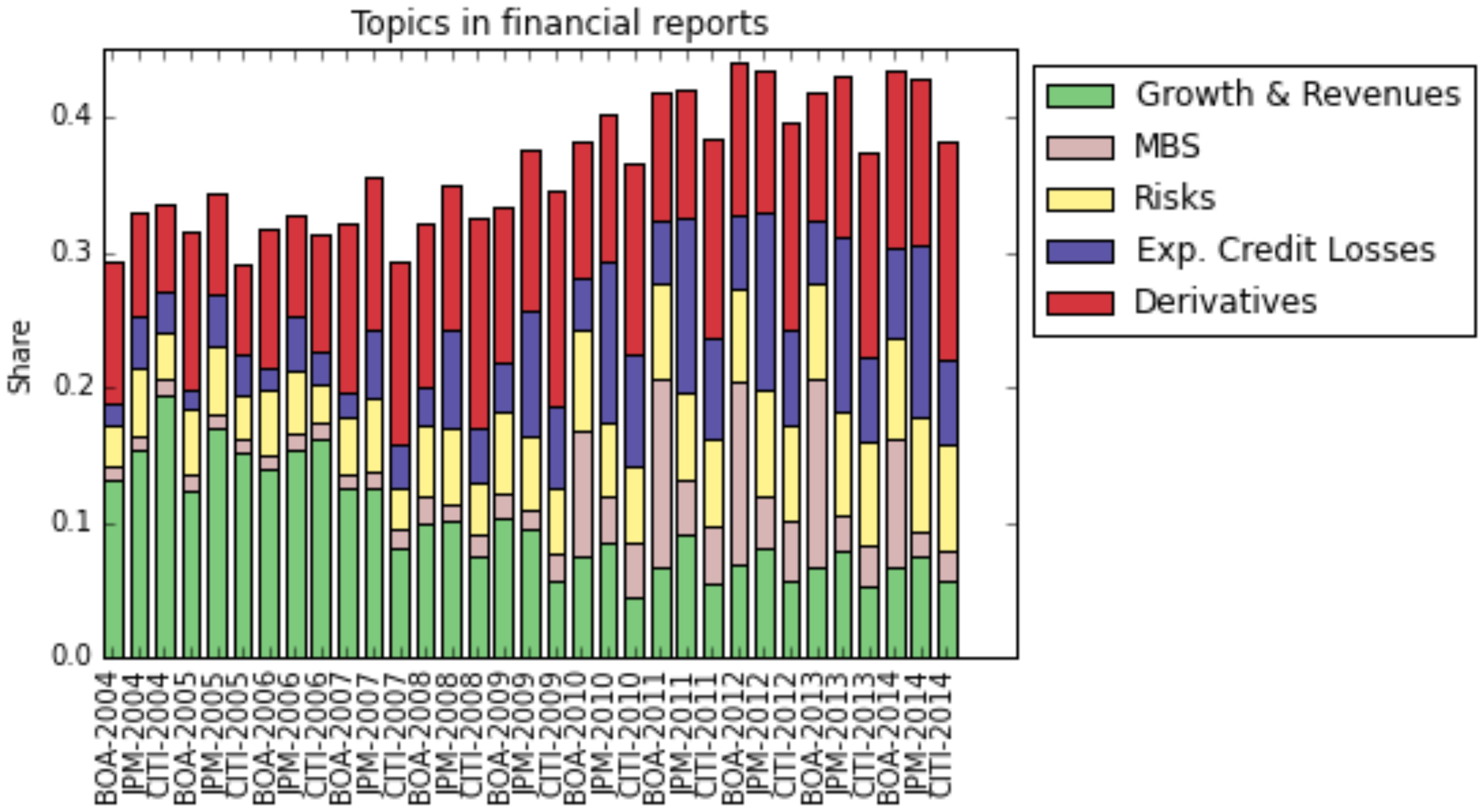
“Mortgage-backed securities”

mortgage
repurchase
servicing
securitization
claims
representations
warranties
corporation
securitizations
loss

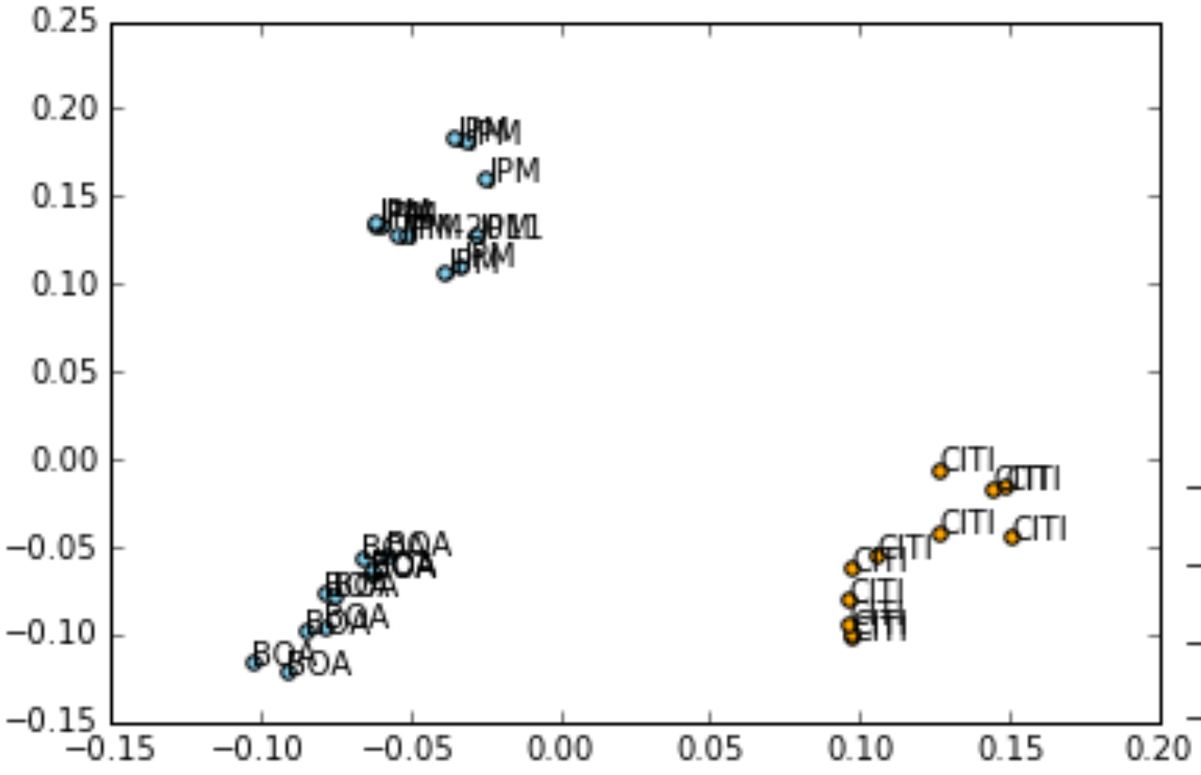
The case of Bank of America

- **Jan. 11, 2008: BoA buys Contrywide Financial**
- Jun. 7, 2008: $108 MM settlement in Contrywide fee complaint
- **Sept. 14, 2008: BoA buys Merril Lynch**
- Feb. 22, 2010: $150 MM settlement with SEC on Merril deal
- Aug. 2, 2010: $600 MM Settlement of Class-Action Suits against Countrywide
- Jun. 11, 2011: $2.5 Bn Buyback in Mortgage deal with Fannie Mae
- Apr. 15, 2011: $1.6 Bn insurance settlement
- Jun. 28, 2011: $8.5 Bn Deal in Investor Suit on Mortgage Debt
- Feb. 8, 2012: $11.8 Bn Settlement on Foreclosure Abuses
- Sept. 28, 2012: $2.4 Bn Shareholder Settlement Over Merrill
- Oct. 23, 2013: Jury Finds Bank Liable of Having Sold Defective Mortgages
- Mar. 26, 2014: $6.3 Bn Settlement of F.H.F.A.'s Mortgage Lawsuit
- Ju. 30, 2014: $1.3 Bn Penalty in Federal Mortgage Case
- **Aug. 21, 2014: $16.7 Bn Mortgage Settlement with Justice Dep.**

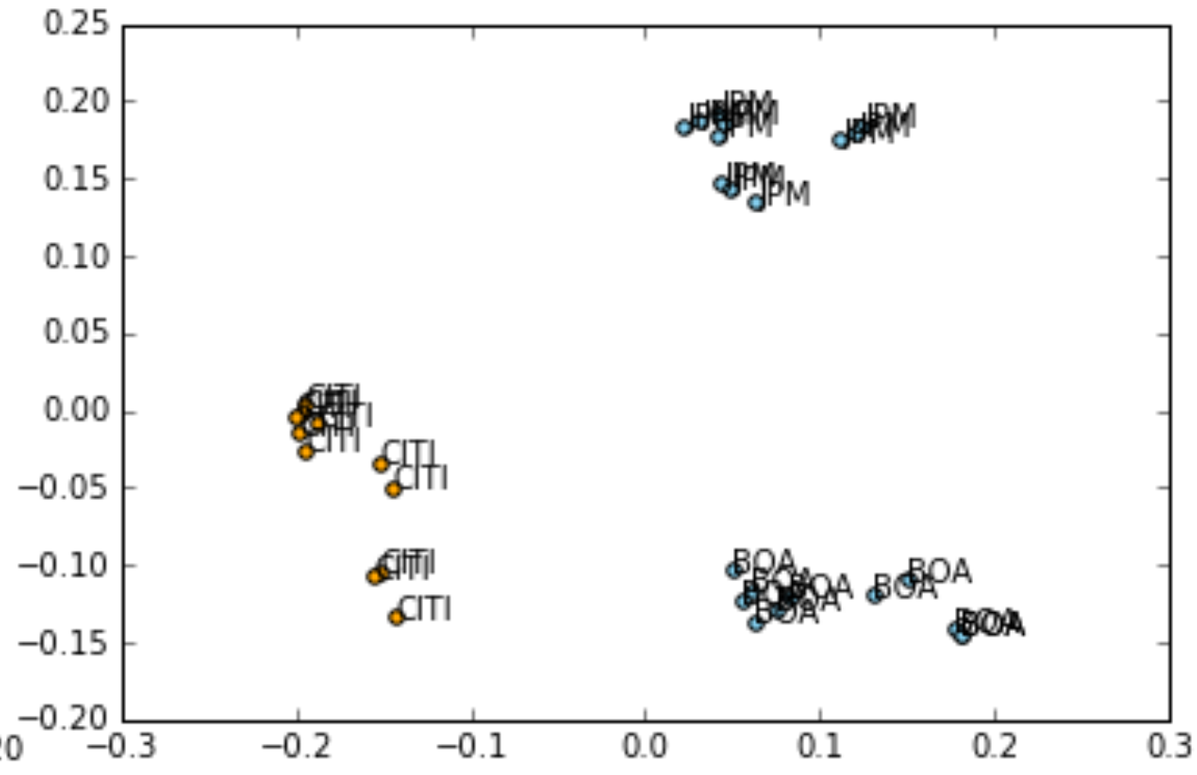
How topics evolve over time?



How similar are the financial reports ?



Distance calculated using word frequencies



Distance calculated using topic shares

Conclusions

1. Topic modeling (TM) could be used to identify themes affecting an industry or a particular company

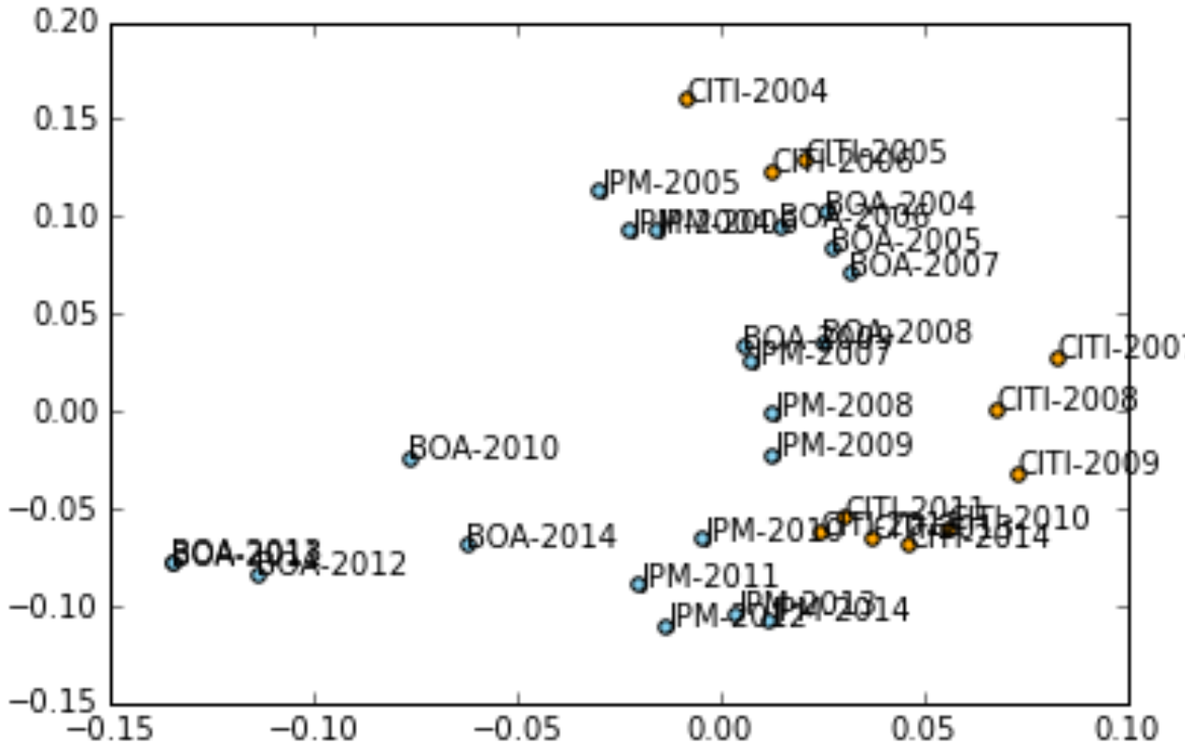
2. TM could be used for dimensionality reduction

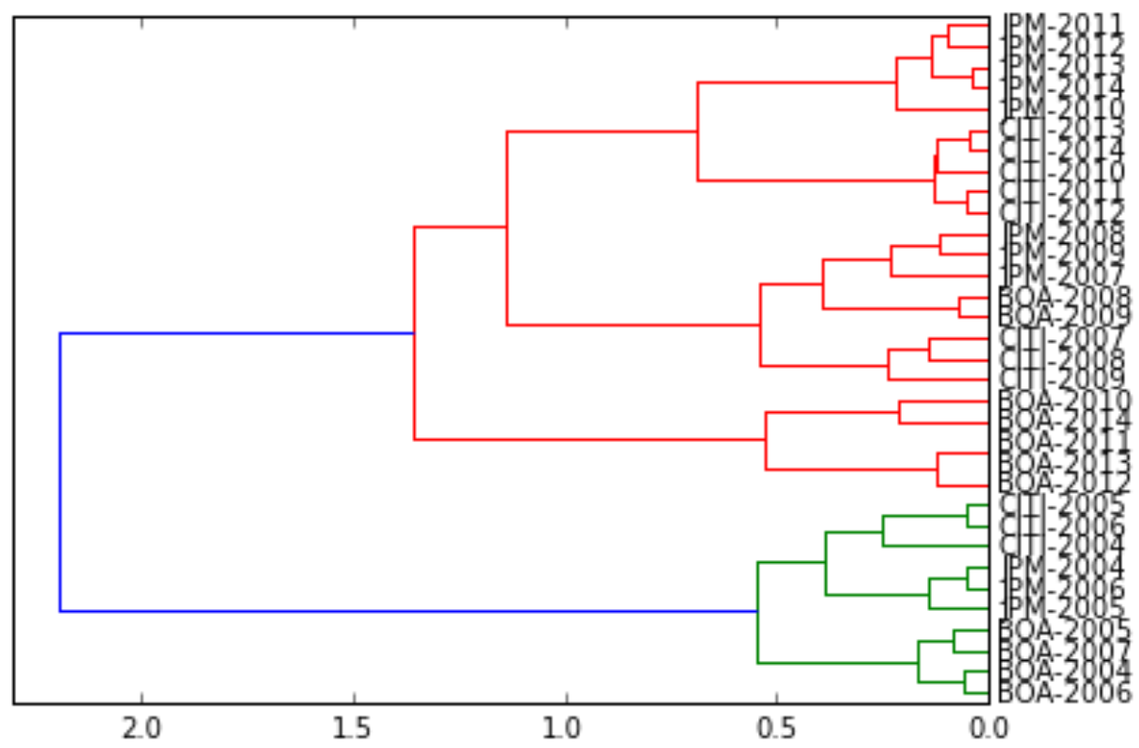
3. Domain knowledge is important to interpret results

4. Certain topics could capture sentiment (Future research)

Thanks!

What happens if we remove the most distinctive topics?

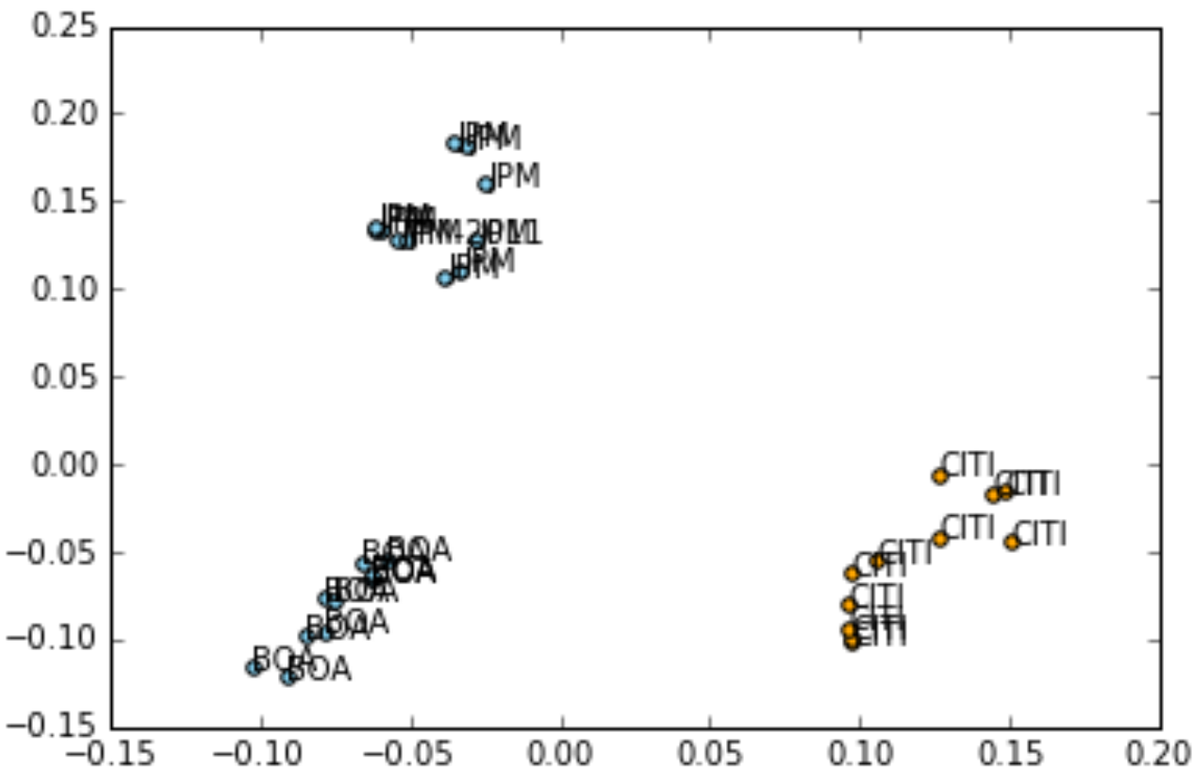




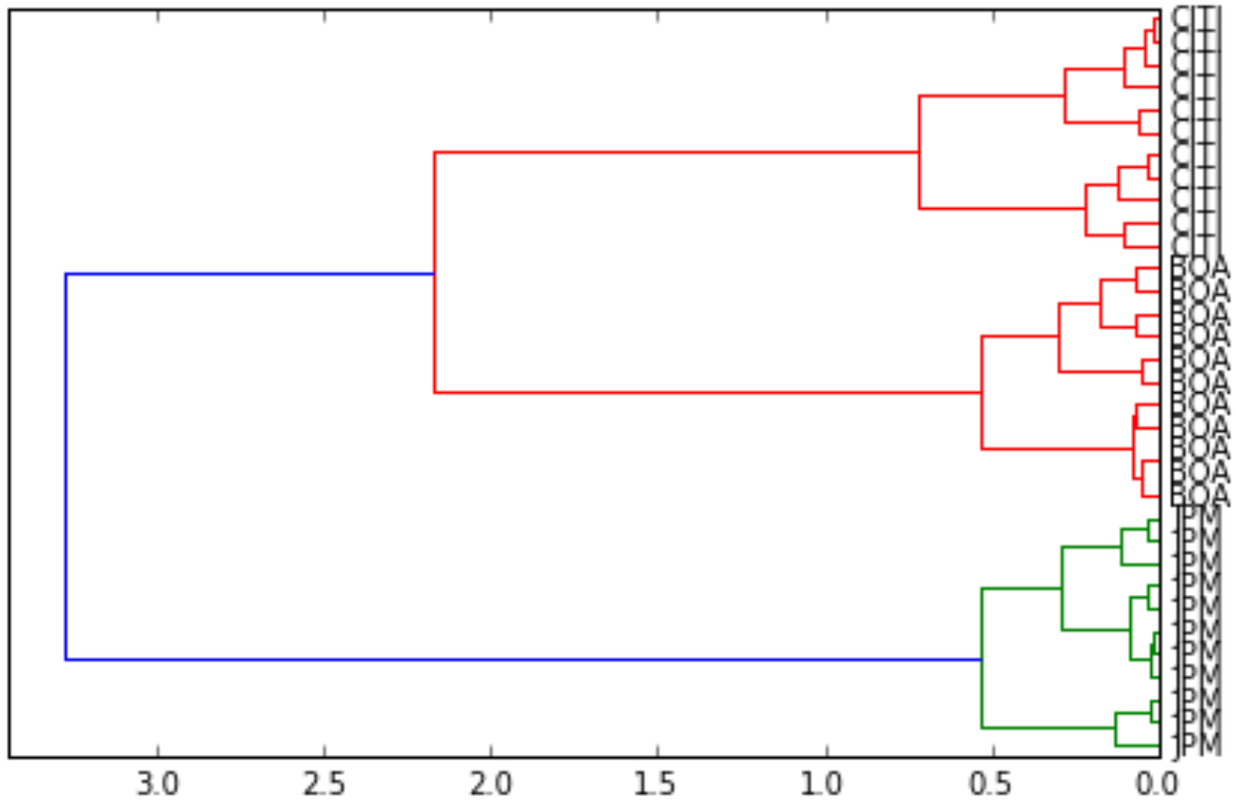
Mutidimensional Scaling **Ward's Clustering**

Distance calculated using topic shares

Are financial reports similar?



Mutidimensional Scaling



Ward's Clustering

What are the topics?

Topic #0 (“Accounting”): tax accounting sfas fair company goodwill earnings reporting impairment statements
Topic #1 (“J.P. Morgan”): firm chase jpmorgan report annual note internal statements pages reporting
Topic #2 (“Mortgage backed securities”): mortgage repurchase servicing securitization claims representations warranties corporation securitizations loss
Topic #3 (“Lawsuits”): court district claims actions filed action corporation defendants class plaintiffs
Topic #4 (“Pension Plans”): plans stock plan benefit pension million awards options expense rate
Topic #5 (“Regulation”): bank subsidiaries federal act banking capital company holding companies fdic
Topic #6 (“Bank of America”): capital corporation bank america stock modifications home mortgage new term
Topic #7 (“Stress testing”): average rate trading var rates basis deposits liabilities portfolio changes
Topic #8 (“Risks”): could risks businesses operational regulatory significant operations affect economic results
Topic #9 (“Bank of America: Countrywide”): commercial percent consumer portfolio million leases allowance real nonperforming estate
Topic #10 (“Collateral”): commitments amount company guarantees agreements transactions collateral corporation obligations respectively
Topic #11 (“Credit risk”): exposure liquidity funding ratings derivative collateral derivatives exposures rating portfolio
Topic #12 (“Growth & Revenues”): million increased banking revenue higher services expense investment offset primarily
Topic #13 (“J.P.Morgan: Washington Mutual”): year revenue prior expense ended annual charge report allowance million
Topic #14 (“Citigroup”): citigroup citi company operations year approximately loss dollars million consumer
Topic #15 (“Company filings”): incorporated reference exhibit form stock registrant dated company bank report
Topic #16 (“Expected credit losses”): allowance consumer portfolio mortgage card impaired past home days principal
Topic #17 (“Regulation”): capital tier basel common regulatory ratio bank iii requirements rules
Topic #18 (“Stock issuance”): stock corporation preferred debt trust company term issued common notes
Topic #19 (“Derivatives”): fair debt trading liabilities derivative level instruments derivatives rate mortgage