2016.12.06

2016 IEEE Big NLP Workshop

Automatic Classification of Securities using Hierarchical Clustering of 10-Ks

Hoseong Yang, Hye Jin Lee, Eugene Cho, Sungzoon Cho

hoseong@dm.snu.ac.kr, hyejinlee@dm.snu.ac.kr, eugene.t.cho@gmail.com, zoon@snu.ac.kr

Department of Industrial Engineering
Seoul National University
Seoul, Korea

Contents

Introduction

What is an "Industry Classification Scheme" (ICS)?

- Methods to cluster financial entities into distinct bundles
  - **Intra-Industry Homogeneity**: Each bundle should contain entities that are "similar" types of business given their market activities
  - **Inter-Industry Homogeneity**: Each bundle's power of representation should be similar from one another
  - In equity market, these bundles represents industry-wise segments of the market with distinct financial characteristics

- ICS facilitate a broad range of cluster-level analysis [1]
  - Sector-wise identification of market competitors
  - Benchmarking company activities and performances
  - Measuring economic indicators
  - Setting up market share

- **Note!** Although the term "classification" is used as a convention, it actually is a ***clustering*** process!

Introduction

Most widely used industrial classification nowadays: SIC and ICB.

- SIC: Standard Industrial Classification [2]
  - Established by the U.S. government in 1937
  - Currently used by SEC and U.S. Government agencies
  - *VERY* outdated, while *rarely updated*
  - In transition to NAICS since 1997

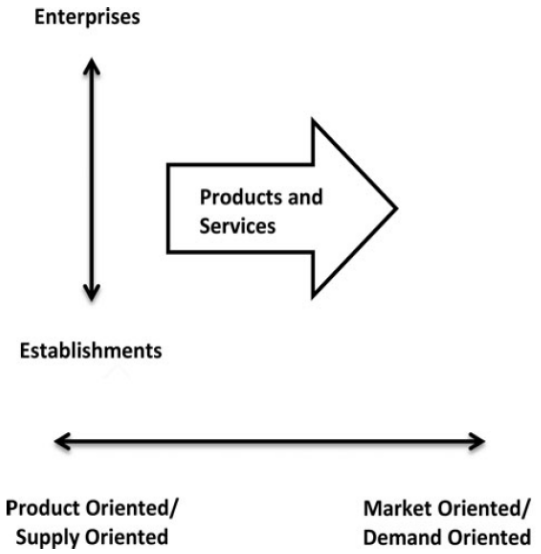
- GICS: Global Industry Classification Standard [3]
  - Developed by MSCI and S&P in 1999
  - One of the most frequently updated classification standards
  - 10 sectors, 24 industry groups, 67 industries and 256 sub-industries.
  - continuously updated by S&P Dow Jones Indices and MSCI.

Introduction

Industry classification scheme can be divided into two categories:
A product-oriented method, and; a market-oriented method.

Orientation refers to the underlying perspective used to aggregate and classify companies and their operations [1].

- Production-oriented classification
  - Classifying companies into industries by identifying similar processes used to produce goods or services.
  - Production processes and product output are key variables.
  - Example) SIC
- Market-oriented classification
  - Companies are classified by the source of revenues, earnings analysis, and market perception.
  - It focuses on consumers and markets a company serves.
  - Example) GICS

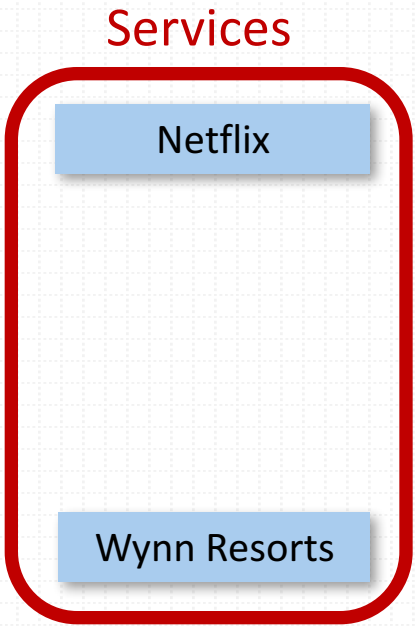


Introduction

Example 1: Netflix

Current classification system groups vastly different companies together

- SIC classifies Netflix into a “Services” sector, along with Wynn Resorts



Netflix (7841)	Wynn Resorts (7780)
Level 1	
Service	Service
Level 2	
Employment activities	Rental and leasing activities
Level 3	
Other human resources division	Other reservation service and related activities

Form 10-Ks, Netflix and Wynn Resorts, 2016

Introduction

Example 1: Netflix

Current classification system groups vastly different companies together

- SIC classifies Netflix into a “Services” sector, along with Wynn Resorts

Netflix

Netflix, Inc. (“Netflix”, “the Company”, “we”, or “us”) is the world’s leading Internet television network with over 75 million streaming members in over 190 countries enjoying more than 125 million hours of TV shows and movies per day, including original series, documentaries and feature films. Our members can watch as much as they want, anytime, anywhere, on nearly any Internet-connected screen. Members can play, pause and resume watching, all without commercials or commitments. Additionally, in the United States ("U.S."), our members can receive DVDs delivered quickly to their homes.
We are a pioneer in the Internet delivery of TV shows and movies, launching our streaming service in 2007. Since this launch, we have developed an ecosystem for Internet-connected screens and have added increasing amounts of content that enable consumers to enjoy TV shows and movies directly on their Internet-connected screens. As a result of these efforts, we have experienced growing consumer acceptance of, and interest in, the delivery of TV shows and movies directly over the Internet.

Wynn Resorts

Wynn Resorts, Limited, ("Wynn Resorts," or together with its subsidiaries, "we" or the "Company"), led by Chairman and Chief Executive Officer, Stephen A. Wynn, is a leading developer, owner and operator of destination casino resorts (integrated resorts) that integrate hotel accommodations and a wide range of amenities, including fine dining outlets, premium retail offerings, distinctive entertainment theaters and large meeting complexes.
Wynn Resorts currently owns 72% of Wynn Macau, Limited, which operates an integrated resort in the Macau Special Administrative Region of the People's Republic of China ("Macau"). Wynn Resorts also owns 100% of and operates an integrated resort in Las Vegas, Nevada.
We are currently constructing Wynn Palace, an integrated resort in the Cotai area of Macau, which we expect to open in the first half of 2016; however, potential construction delays could push the opening date into the second half of 2016. We have begun site preparation and pre-construction activities for the development and construction of an integrated resort in Everett, Massachusetts, adjacent to Boston.

Form 10-Ks, Netflix and Wynn Resorts, 2016

Introduction

Example 2: Amazon

The company structure of Amazon has changed over the course of years

- The “self-identity” of Amazon has evolved from a “bookstore” to multi-branch online retailer

1998

Amazon.com, Inc. ("Amazon.com" or the "Company") is the leading online retailer of books. Since opening for business as "Earth's Biggest Bookstore" in July 1995, Amazon.com has become one of the most widely known, used and cited commerce sites on the World Wide Web (the "Web"). Amazon.com strives to offer its customers compelling value through innovative use of technology, broad selection, high-quality content, a high level of customer service, competitive pricing and personalized services.

2016

We serve consumers through our retail websites and focus on selection, price, and convenience. We design our websites to enable millions of unique products to be sold by us and by third parties across dozens of product categories. Customers access our websites directly and through our mobile websites and apps. We also manufacture and sell electronic devices, including Kindle e-readers, Fire tablets, Fire TVs, and Echo. We strive to offer our customers the lowest prices possible through low everyday product pricing and shipping offers, and to improve our operating efficiencies so that we can continue to lower prices for our customers. We also provide easy-to-use functionality, fast and reliable fulfillment, and timely customer service.

Form 10-Ks, Amazon, 1998 and 2016

Introduction

Example 2: Amazon

The market structure in which Amazon competes has changed over the course of years

- The “self-identified” market competitors of Amazon now include a set of entities vastly different from the past

1998

The Company's current or potential competitors include (i) various online booksellers and vendors of other information-based products such as CDs and videotapes, including entrants into narrow specialty niches, (ii) a number of indirect competitors that specialize in online commerce or derive a substantial portion of their revenues from online commerce, through which retailers other than the Company may offer products and (iii) publishers, distributors and retail vendors of books, music and videotapes, including Barnes & Noble, Inc., Bertelsmann AG and other large specialty booksellers and integrated media corporations, many of which possess significant brand awareness, sales volume and customer bases.

2016

Our current and potential competitors include: (1) online, offline, and multichannel retailers, publishers, vendors, distributors, manufacturers, and producers of the products we offer and sell to consumers and businesses; (2) publishers, producers, and distributors of physical, digital, and interactive media of all types and all distribution channels; (3) web search engines, comparison shopping websites, social networks, web portals, and other online and app-based means of discovering, using, or acquiring goods and services, either directly or in collaboration with other retailers; (4) companies that provide e-commerce services, including website development, advertising, fulfillment, customer service, and payment processing; (5) companies that provide fulfillment and logistics services for themselves or for third parties, whether online or offline; (6) companies that provide information technology services or products, including on-premises or cloud-based infrastructure and other services; and (7) companies that design, manufacture, market, or sell consumer electronics, telecommunication, and electronic devices.

Form 10-Ks, Amazon, 1998 and 2016

Introduction

Limitations of previous Industry classification schemes

- Limitations

1) Terrible cases of between-class homogeneity, due to the emphasis on how firms do what they do, as opposed to the purpose for which they do it.

   (Example 1:Netflix)

2) Consequence of innovation and technological change that has resulted in products and services that are more complex, and means of production that vary over time.

   (Example 2:Amazon)

3) All existing classification system requires human input at some point, which can be very costly and time-consuming (especially for information updates).

Introduction

Previous limitations call for new approach

We aim to tackle them by exploiting the business section of the Form 10-K

- We address these issues by exploiting the business section of the Form 10-Ks and cluster securities whose content of the text appear similar
- The Form 10-K is a comprehensive summary of a company's business operations and market performances, filed annually as required by the U.S. Securities and Exchange Commission
- Its business section provides a finely detailed description of the company's business operations, organizational structure, risk factors, and market competitors
- We use define these textual information as the ***firm's self-identity*** and use it as the new standards for clustering securities

Data Source

Items typically reported in the Form 10-Ks

Form 10-K Example

Microsoft 10-K (2016)

Table of Contents

Introduction

Our newly developed ICS, Business Text Industry Classification (BTIC), is designed to outperform the existing classification system in four aspects

- By employing doc2vec and hierarchical clustering on the business section of the Form 10-Ks, we develop a new ICS, **Business Text Industry Classification** (BTIC)

- Given that it performs just as well as SIC or GICS, BTIC outperforms the existing schemes in four aspects:
  - **Process automation** – Once business section is extracted, the remaining steps are automated
  - **Objectivity** – Clustering is entirely data-driven
  - **Flexibility** – Organizational structure of BTIC is easily modified, fit to the user's research needs
  - **Result Interpretability** – BTIC provides a list of words that "represent" each cluster created, thus helping the user's understanding of the clustering process and results

Introduction

Proposed Method

1 Data collection: Extract Business section (item 1) of Form10-Ks using regular expressions


Table of Contents

AMAZON.COM, INC.

PART I

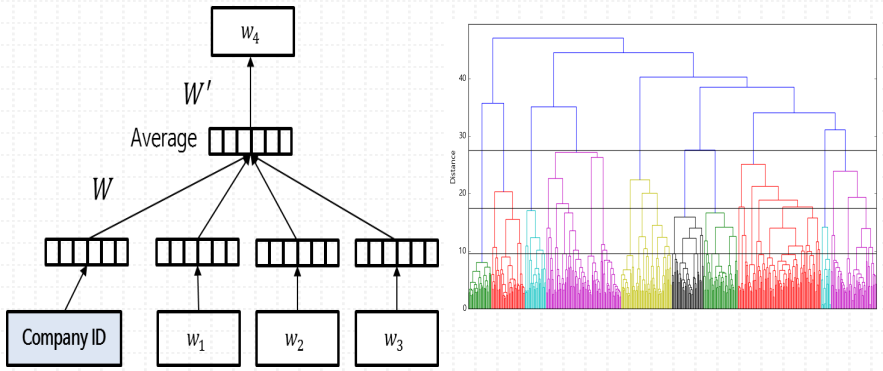
Item 1. ***Business***

This Annual Report on Form 10-K and the documents incorporated herein by reference contain forward-looking statements based on expectations, estimates, and projections as of the date of this filing. Actual results may differ materially from those expressed in forward-looking statements. See Item 1A of Part I—"Risk Factors."

Amazon.com, Inc. was incorporated in 1994 in the state of Washington and reincorporated in 1996 in the state of Delaware. Our principal corporate offices are located in Seattle, Washington. We completed our initial public offering in May 1997 and our common stock is listed on the NASDAQ Global Select Market under the symbol "AMZN."

As used herein, "Amazon.com," "we," "our," and similar terms include Amazon.com, Inc. and its subsidiaries, unless the context indicates otherwise.

2 Modeling: Document embedding, hierarchical Clustering



3 Clustering Exploration

Word	cos
electric/JJ	0.5795
generation/NN	0.5167
electricity/NN	0.5009
transmission/NN	0.4570
power/NN	0.4520
energy/NN	0.4362
utility/NN	0.4289
renewable/JJ	0.4176
utilities/NNS	0.4123
generating/VBG	0.4097
generating/NN	0.3811
plants/NNS	0.3776
load/NN	0.3573
fuel/NN	0.3341
fossil/NN	0.3338

Word	cos
oil/NN	0.5303
exploration/NN	0.5232
drilling/NN	0.4710
gas/NN	0.4642
natural/JJ	0.4576
wells/NNS	0.4507
crude/JJ	0.4212
crude/NN	0.4105
drilling/VBG	0.3794
production/NN	0.3773
offshore/RB	0.3761
pipeline/NN	0.3700
liquids/NNS	0.3677
reserves/NNS	0.3643
gathering/NN	0.3565

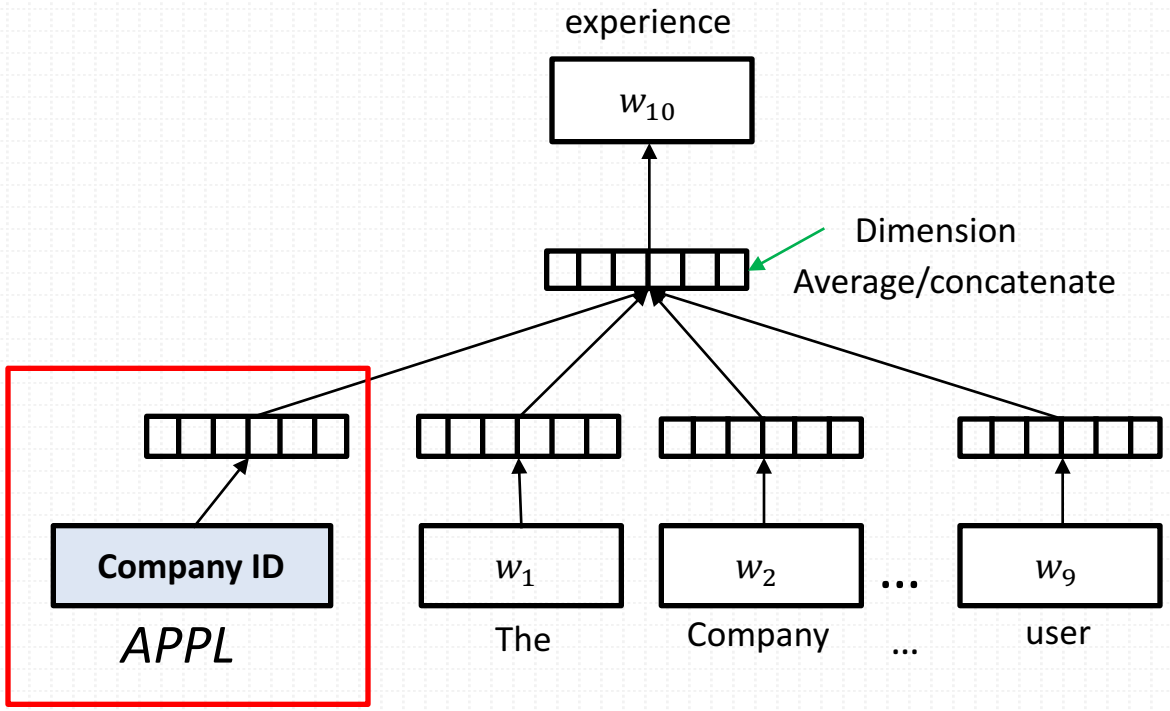
word	cos
banking/NN	0.6384
institutions/NNS	0.5216
banks/NNS	0.5029
bank/NN	0.4983
institution/NN	0.4732
lending/NN	0.4319
nonbank/JJ	0.4196
depository/NN	0.3856
holding/VBG	0.3770
deposit/NN	0.3716
deposits/NNS	0.3557
risk/NN	0.3414
banking/VBG	0.3335
insured/JJ	0.3264
regulators/NNS	0.3248

4 Evaluation: computing R^2 of 12 financial variables

$$var_{i,t} = \alpha + \beta \cdot var_{ind,t} + \varepsilon_{i,t}$$

Proposed Methods

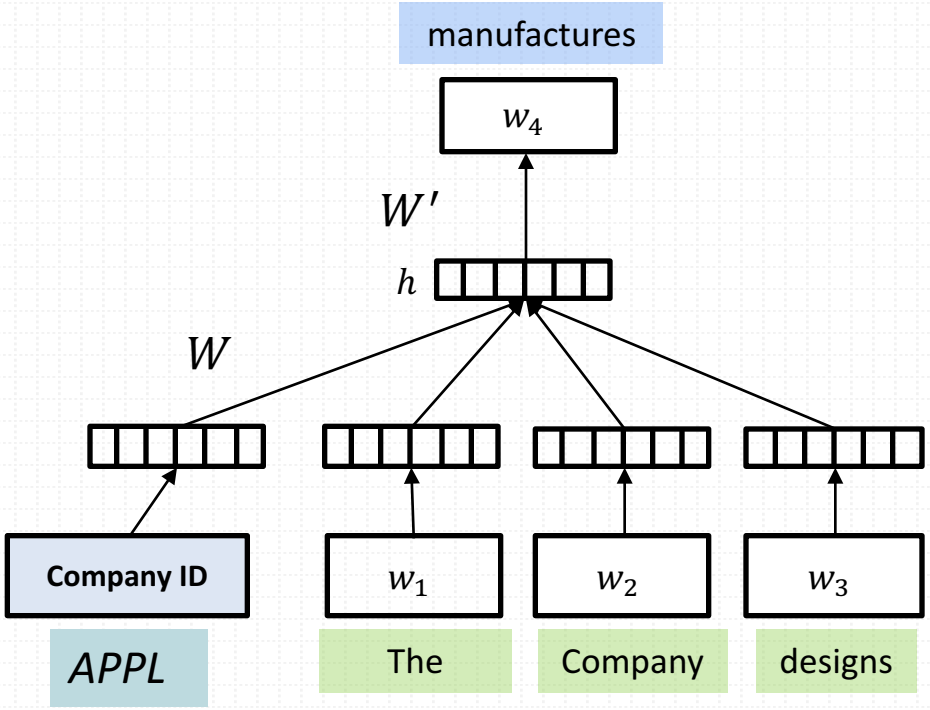
We employ Doc2vec [4] modelling,
Embedding Form 10-K documents and words appearing in the reports on to a continuous space



- Word2vec captures semantic, and syntactic relationships between words, sentences, and documents
- Doc2vec is very similar to word2vec. Document vector is trained with word vectors.
- Compared to Bag-of-Words(BoW) approach, we can consider the word orders, with low-dimensional vectors, in our analysis
- Performs well in document sentiment classification
- Can also use for information retrieval

Methods

We employ Doc2vec [4] modelling,
The process of training



- Objective function
  (Max likelihood = Min negative likelihood)

$$L = -\sum_{i=1}^{V} \log p(w_i|w_{context}) - \sum_{j=1}^{D}\sum_{i=1}^{V} \log p(w_i|d_j)$$

w_i: word vector, d_j : company vector

- Training w_i, d_j using backpropagation

$$w_i^{(new)} = w_i^{(old)} - \eta \cdot e_i \cdot h$$

$$d_{APPL}^{(new)} = d_{APPL}^{(old)} - \eta \cdot e_i \cdot h$$

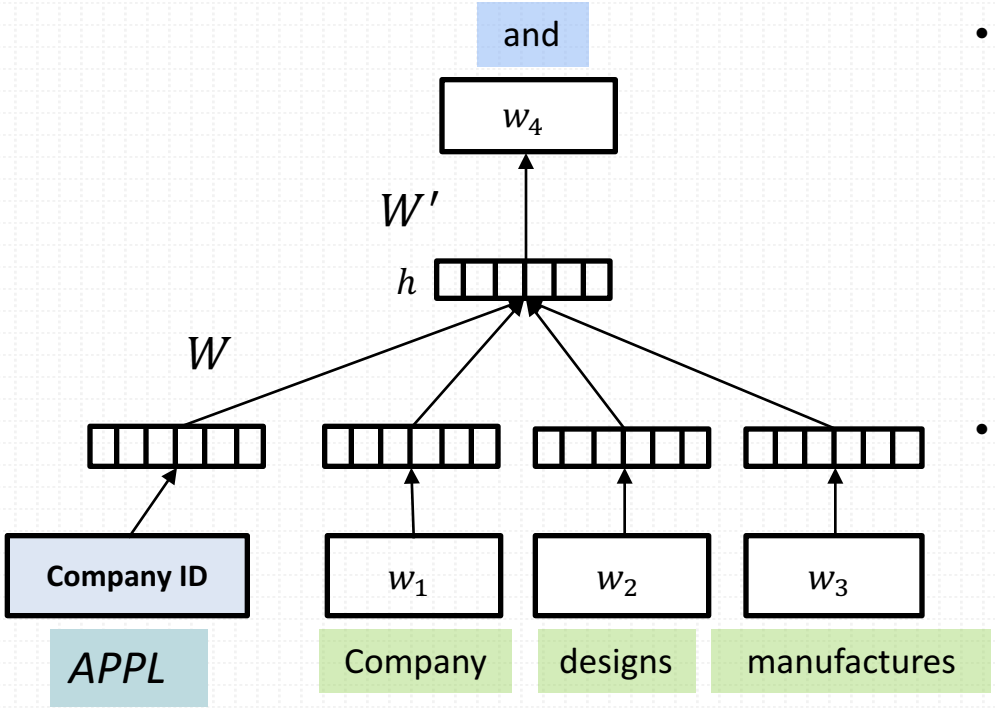
η:learning rage, $e_i = \frac{\partial L}{\partial w_i}$, $h = \frac{1}{4}(w_1 + w_2 + w_3 + d_{APPL})$

APPL

The Company designs, manufactures and markets mobile communication and media devices, personal computers and portable digital music players, and sells a variety of related software,
⋮

Methods

We employ Doc2vec [4] modelling,
The process of training



- Objective function
  (Max likelihood = Min negative likelihood)

$$L = -\sum_{i=1}^{V} \log p(w_i|w_{context}) - \sum_{j=1}^{D}\sum_{i=1}^{V} \log p(w_i|d_j)$$

w_i: word vector, d_j : company vector

- Training w_i, d_j using backpropagation

$$w_i^{(new)} = w_i^{(old)} - \eta \cdot e_i \cdot h$$

$$d_{APPL}^{(new)} = d_{APPL}^{(old)} - \eta \cdot e_i \cdot h$$

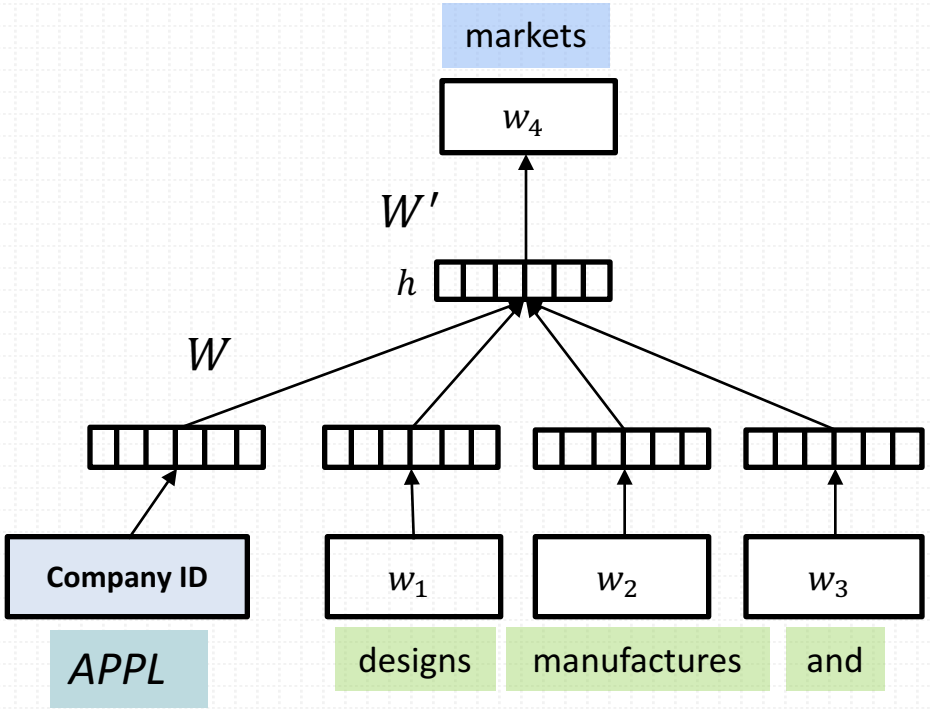
η:learning rage, $e_i = \frac{\partial L}{\partial w_i}$, $h = \frac{1}{4}(w_1 + w_2 + w_3 + d_{APPL})$

APPL

The Company designs, manufactures and markets mobile communication and media devices, personal computers and portable digital music players, and sells a variety of related software,
⋮

Methods

We employ Doc2vec [4] modelling,
The process of training



- Objective function
  (Max likelihood = Min negative likelihood)

$$L = -\sum_{i=1}^{V} \log p(w_i|w_{context}) - \sum_{j=1}^{D}\sum_{i=1}^{V} \log p(w_i|d_j)$$

w_i: word vector, d_j : company vector

- Training w_i, d_j using backpropagation

$$w_i^{(new)} = w_i^{(old)} - \eta \cdot e_i \cdot h$$

$$d_{APPL}^{(new)} = d_{APPL}^{(old)} - \eta \cdot e_i \cdot h$$

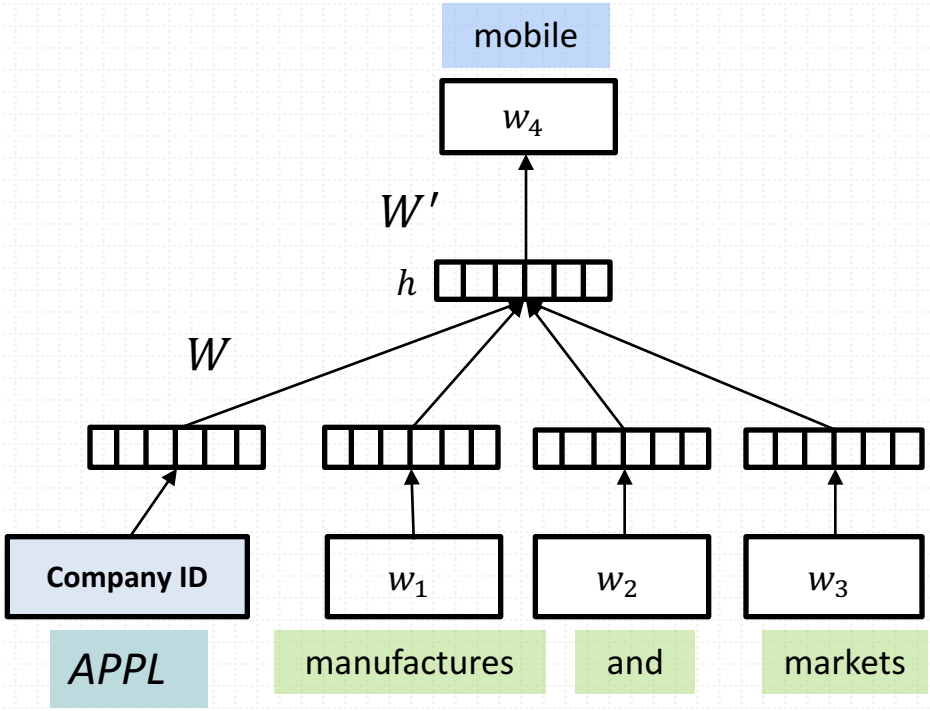
η:learning rage, $e_i = \frac{\partial L}{\partial w_i}$, $h = \frac{1}{4}(w_1 + w_2 + w_3 + d_{APPL})$

APPL

The Company designs, manufactures and markets mobile communication and media devices, personal computers and portable digital music players, and sells a variety of related software,
⋮

Methods

We employ Doc2vec [4] modelling,
The process of training



- Objective function
  (Max likelihood = Min negative likelihood)

$$L = -\sum_{i=1}^{V} \log p(w_i|w_{context}) - \sum_{j=1}^{D}\sum_{i=1}^{V} \log p(w_i|d_j)$$

w_i: word vector, d_j : company vector

- Training w_i, d_j using backpropagation

$$w_i^{(new)} = w_i^{(old)} - \eta \cdot e_i \cdot h$$

$$d_{APPL}^{(new)} = d_{APPL}^{(old)} - \eta \cdot e_i \cdot h$$

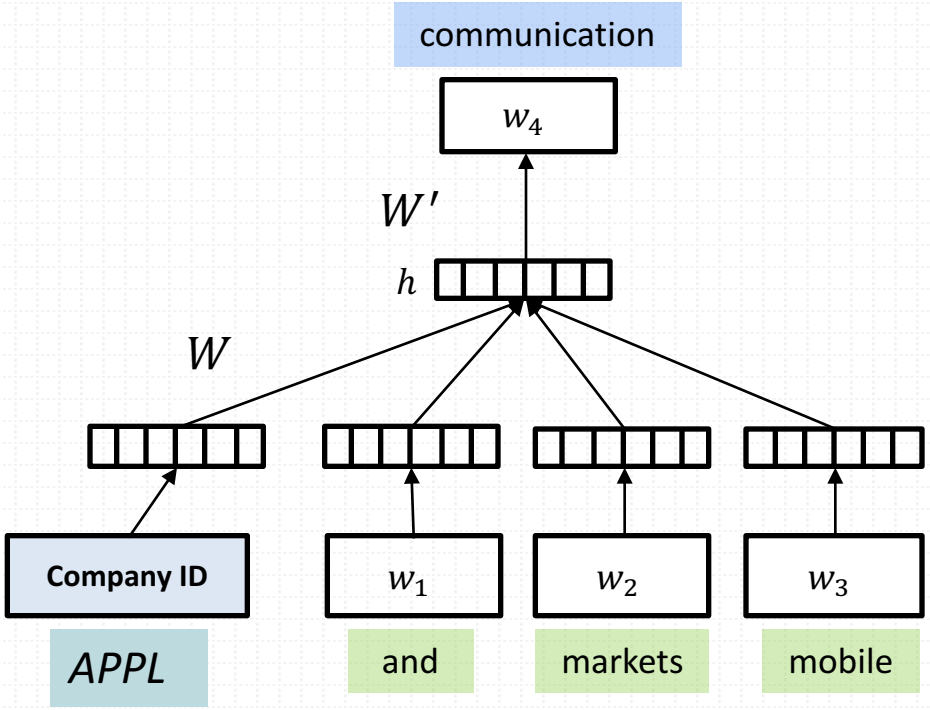
η:learning rage, $e_i = \frac{\partial L}{\partial w_i}$, $h = \frac{1}{4}(w_1 + w_2 + w_3 + d_{APPL})$

APPL

The Company designs, manufactures and markets mobile communication and media devices, personal computers and portable digital music players, and sells a variety of related software,
⋮

Methods

We employ Doc2vec [4] modelling,
The process of training



- Objective function
  (Max likelihood = Min negative likelihood)

$$L = -\sum_{i=1}^{V} \log p(w_i|w_{context}) - \sum_{j=1}^{D}\sum_{i=1}^{V} \log p(w_i|d_j)$$

w_i: word vector, d_j : company vector

- Training w_i, d_j using backpropagation

$$w_i^{(new)} = w_i^{(old)} - \eta \cdot e_i \cdot h$$

$$d_{APPL}^{(new)} = d_{APPL}^{(old)} - \eta \cdot e_i \cdot h$$

η:learning rage, $e_i = \frac{\partial L}{\partial w_i}$, $h = \frac{1}{4}(w_1 + w_2 + w_3 + d_{APPL})$

APPL

The Company designs, manufactures and markets mobile communication and media devices, personal computers and portable digital music players, and sells a variety of related software,
⋮

Proposed Methods

We employ Ward's hierarchical clustering [5] to resemble the organizational structure of the existing classification schemes

- Industry classification schemes group entities in a hierarchical manner
- BTIC employs a Ward's hierarchical clustering [5] method to resemble the organizational structure of the existing ICS
  - At each iteration, Ward's method finds a pair of clusters that leads to the minimum increase in total within cluster variance after merging the clusters
  - Uses Euclidean distance to cluster entities

$$\Delta(A,B) = \sum_{i\in A\cup B}\|\vec{x}_i-\vec{m}_{A\cup B}\|^2-\sum_{i\in A}\|\vec{x}_i-\vec{m}_A\|^2-\sum_{i\in B}\|\vec{x}_i-\vec{m}_B\|^2 = \frac{n_An_B}{n_A+n_B}\|\vec{m}_A-\vec{m}_B\|^2$$

$\overrightarrow{m_j}$: center of cluster j

n_j : number of points in cluster j

- Studies have shown that the distributed vectors are correlated with the semantic similarity [6, 7]

Experiment

Data Summary

✓ Form 10-Ks

- Used web crawling to collect Form 10-Ks of 504 S&P500 companies, published around January 1, 2016.
- Broadcom Corp.(BRCM), Coca-Cola Enterprises(CCE) ACE Limited and Cablevision Systems Corp. are discarded since their 10-Ks have not been reported.
- Ultimately, our experiment data includes 500 S&P500 companies.

✓ Financial market ratios

- Matched collected companies to 12 market ratios from January 1, 2015, to December 31, 2015
- Raw data was retrieved from the Center for Research in Security Prices (CRSP) and Compustat databases, which were later used to calculate representative market ratios

Experiment

Doc2vec modeling

1) Doc2vec modeling

- Vector size : 50 dimension
- Window size : 2
- Number of training epochs : 10

2) Hierarchical clustering using Ward's method

- Number of level 1 categories: 10
- Number of level 2 categories: 24
- Number of level 3 categories: 68

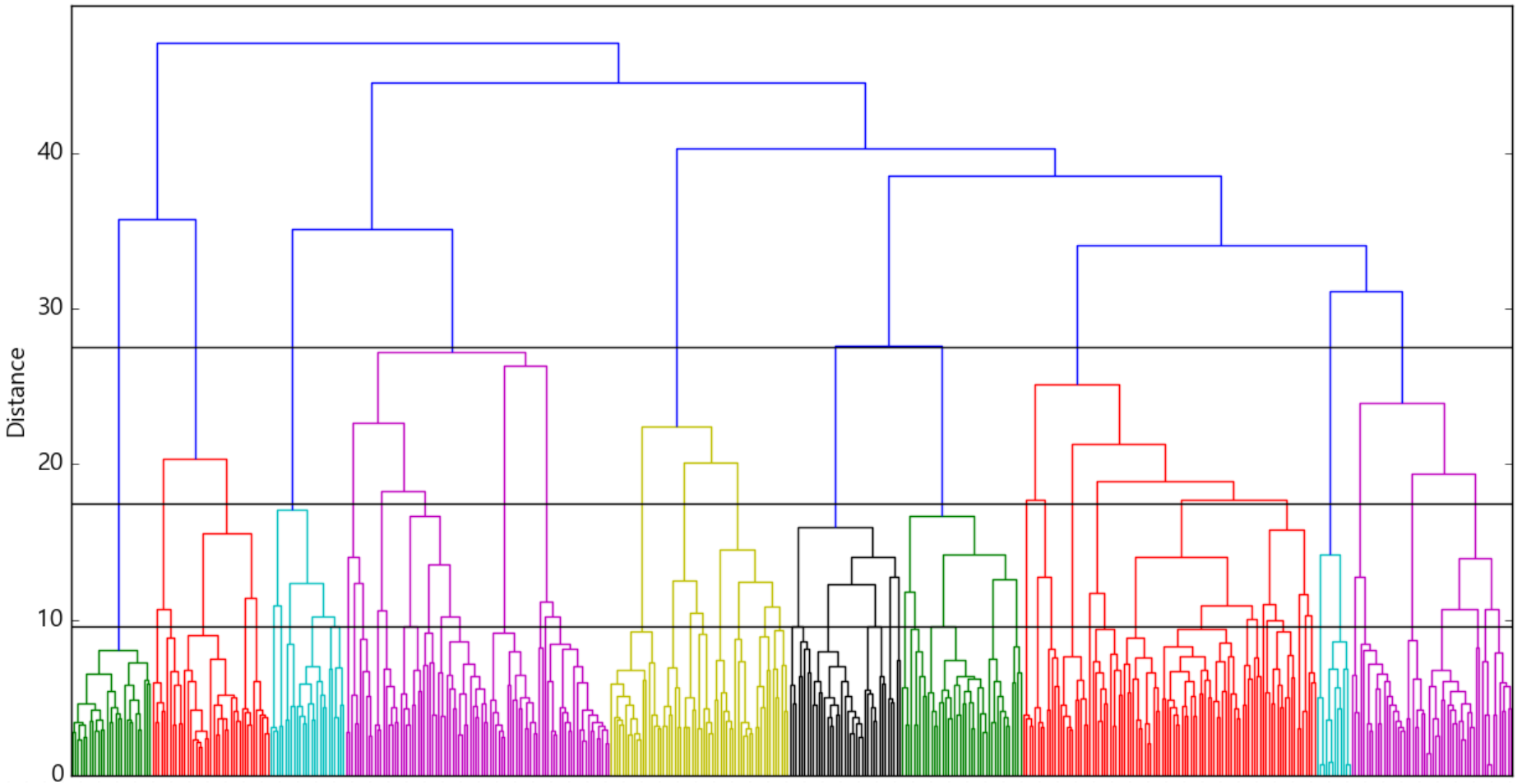
Table I: Summary Statistics

	Level	# official categories	Mean # of firms per industry
SIC	Level 1 (broadest)	10	464
	Level 2	71	15
	Level 3	264	7
GICS	Level 1 (broadest)	10	228
	Level 2	24	103
	Level 3 (narrowest)	68	52
TF-IDF	Level 1 (broadest)	10	281
	Level 2	24	117
	Level 3 (narrowest)	68	42
BTIC	Level 1 (broadest)	10	283
	Level 2	24	101
	Level 3 (narrowest)	68	50

Notes – Mean # of firms per industry records the average number of securities per category in the corresponding division level of the subject classification scheme in our experiment data.

Experiment

Clustering Result: Represent each company as a vector, then run hierarchical clustering
Set different threshold values to create various groupings



Experiment

Clustering Result: Cluster Investigation (Top 10 Market Cap Firms)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Public Serv. Enterprise Inc.	Chevron Corp.	Simon Property Group Inc.	Morgan Stanley	Johnson & Johnson
NiSource Inc.	Marathon Petroleum	Prologis	Wells Fargo	Gilead Sciences
Southern Co.	Marathon Oil Corp.	Plum Creek Timber Co.	JPMorgan Chase & Co.	Allergan, Plc
Duke Energy	ConocoPhillips	Public Storage	Principal Financial Group	Amgen Inc.
SCANA Corp.	Pioneer Natural Resources	American Tower Corp.	Bank of America Corp.	Bristol-Myers Squibb
Dominion Resources	EOG Resources	Equity Residential	Citigroup Inc.	United Health Group Inc.
American Electric Power	Valero Energy	Crown Castle Int'l Corp.	Omnicom Group	CVS Health Corp.
Pepco Holdings Inc.	Kinder Morgan	AvalonBay Communities	Visa Inc.	Medtronic Inc.
Exelon Corp.	Halliburton Co.	Welltower Inc	Mastercard Inc.	AbbVie Inc.
NRG Energy	Anadarko Petroleum Corp.	General Growth Properties	Marsh & McLennan	Celgene Corp.

Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
PepsiCo Inc.	Wal-Mart Stores	Berkshire Hathaway	The Walt Disney Co.	Apple Inc.
The Coca Cola Co.	Home Depot	Exxon Mobil Corp.	Time Warner Inc.	Amazon.com Inc.
Monster Beverage	Costco Co.	General Electric	Scripps Networks Int.	Alphabet Inc. Class C
McDonald's Corp.	Lowe's Cos.	Boeing Co.	Twenty-First Century Fox	Facebook, Inc.
Kraft Heinz Co	O'Reilly Automotive	Raytheon Co.	Twenty-First Century Fox Inc Class B	Alphabet Inc. Class A
J.M. Smucker Co.	TJX Companies Inc.	United Technologies	CBS Corp.	AT&T Inc.
Mondelez Int'l	Target Corp.	Honeywell Int'l Inc.	Viacom Inc.	Verizon Communications
Colgate-Palmolive	Kroger Co.	United Parcel Service	News Corporation	Intel Corp.
DuPont	Activision Blizzard Inc	Union Pacific	Discovery Com Class C	Cisco Systems
Philip Morris Int'l	L Brands Inc.	Lockheed Martin Corp.	News Corp Class B	Comcast Class A Comm.

Experiment

Clustering Result: Cluster Investigation (Top 5 Market Cap Firms)

Table II: Clustering result by BTIC

Cluster	Top 5 Securities
C1	Public Serv. Enterprise Inc., NiSource Inc., Southern Co. Duke Energy, SCANA Corp.
C2	Chevron Corp., Marathon Petroleum, Marathon Oil Corp. ConocoPhillips, Pioneer Natural Resources
C3	Simon Property Group Inc., Prologis, Plum Creek Timber Public Storage, American Tower Corp.
C4	Morgan Stanley, Wells Fargo, JPMorgan Chase & Co. Principal Financial Group, Bank of America Corp.
C5	Johnson & Johnson, Gilead Sciences, Allergan, Plc. Amgen Inc., Bristol-Myers Squibb
C6	PepsiCo Inc., The Coca Cola Co., Monster Beverage McDonald's Corp., Kraft Heinz Co.
C7	Wal-Mart Stores, Home Depot, Costco Co. Lowe's Cos., O'Reilly Automotive
C8	Berkshire Hathaway, Exxon Mobil Corp., General Electric Boeing Co., Raytheon Co.
C9	The Walt Disney Co., Time Warner Inc., Scripps Networks 21st Century Fox, 21st Century Fox Inc Class B
C10	Apple Inc., Amazon.com Inc., Alphabet Inc. Class C Facebook Inc., Alaphabet Inc. Class A (Netflix)

For C10, we add Netflix in a parenthesis; Netflix does not make the top 5 market cap firms in the group, but it serves as an important instance which we talk about in Section IV.

- BTIC automatically learned to form a "real estate" group apart from other "finance and banking" entities
- GICS just announced the inclusion of real estate sector to its system
- This shows that BTIC can effectively detect fundamental differences across firms and form groups that are reasonable to human understanding

Evaluation

Qualitative Evaluation: Comparison to GICS

- In SIC, Netflix belongs to the “Services” sector with “Wynn Resort”.
- Netflix is ultimately a television network streamed over internet.
- BTIC captures this leading feature well.

Name	Ticker	BTIC_10	BTIC_24	BTIC_68	GICS_Sector	GICS_Industry	SIC
Alphabet Inc. Class C	GOOG	10	24	67	Information Technology	Software & Services	Services
Facebook, Inc.	FB	10	24	67	Information Technology	Software & Services	Services
Alphabet Inc. Class A	GOOGL	10	24	67	Information Technology	Software & Services	Services
Yahoo Inc.	YHOO	10	24	67	Information Technology	Software & Services	Services
Amazon.com Inc.	AMZN	10	24	68	Consumer Discretionary	Retailing	Retail Trade
PayPal Holdings Inc.	PYPL	10	24	68	Information Technology	Software & Services	Services
Adobe Systems Inc.	ADBE	10	24	68	Information Technology	Software & Services	Services
eBay Inc.	EBAY	10	24	68	Information Technology	Software & Services	Services
Intuit Inc.	INTU	10	24	68	Information Technology	Software & Services	Services
Netflix Inc.	NFLX	10	24	68	Consumer Discretionary	Retailing	Services

Evaluation

Qualitative Evaluation: Comparison to GICS

- In GICS, Amazon belongs to the “Consumer Discretionary” sector.
- Amazon has been classified as “Retail Trade” in SIC system, and as “Consumer Discretionary” in GICS scheme, along with CostCo and WalMart
- BTIC places Amazon.com in the same group with IT companies such as Google and Facebook.

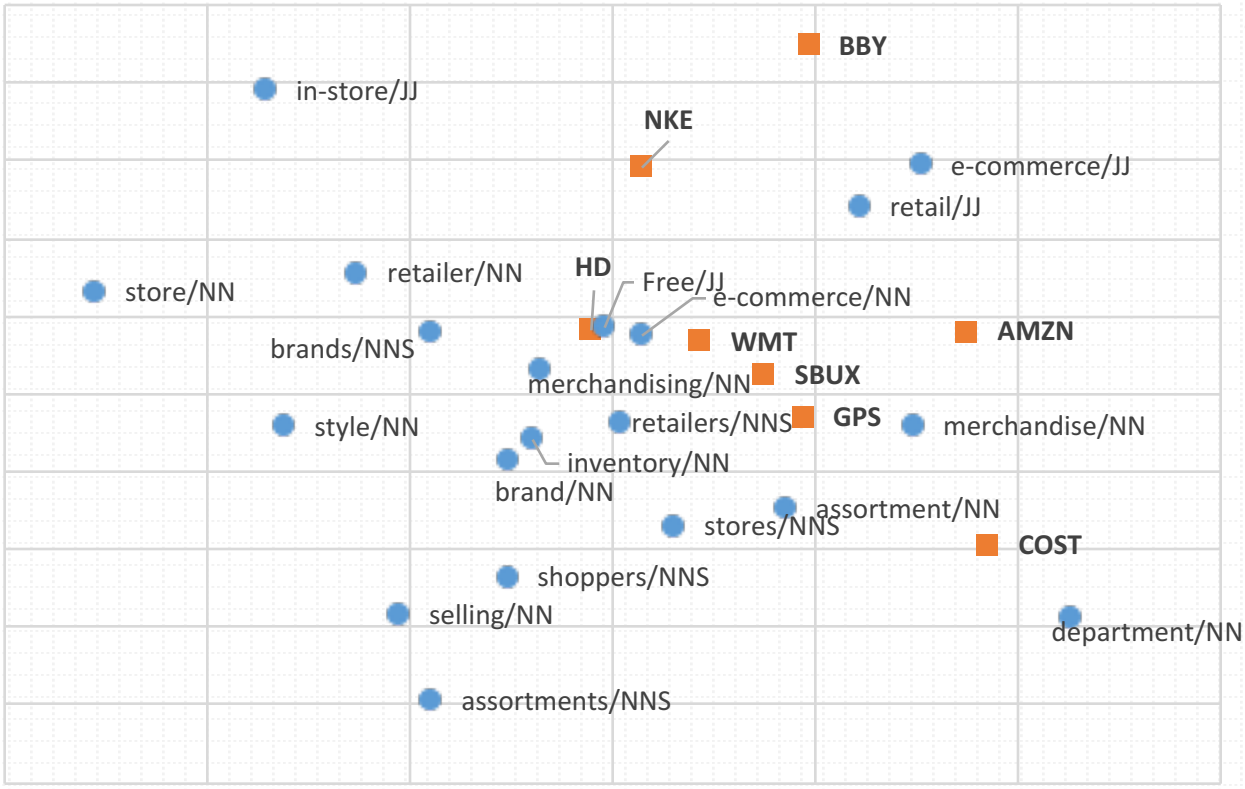
Name	Ticker	BTIC_10	BTIC_24	BTIC_68	GICS_Sector	GICS_Industry	SIC
Alphabet Inc. Class C	GOOG	10	24	67	Information Technology	Software & Services	Services
Facebook, Inc.	FB	10	24	67	Information Technology	Software & Services	Services
Alphabet Inc. Class A	GOOGL	10	24	67	Information Technology	Software & Services	Services
Yahoo Inc.	YHOO	10	24	67	Information Technology	Software & Services	Services
Amazon.com Inc.	AMZN	10	24	68	Consumer Discretionary	Retailing	Retail Trade
PayPal Holdings Inc.	PYPL	10	24	68	Information Technology	Software & Services	Services
Adobe Systems Inc.	ADBE	10	24	68	Information Technology	Software & Services	Services
eBay Inc.	EBAY	10	24	68	Information Technology	Software & Services	Services
Intuit Inc.	INTU	10	24	68	Information Technology	Software & Services	Services
Netflix Inc.	NFLX	10	24	68	Consumer Discretionary	Retailing	Services

Evaluation

Qualitative Evaluation Interpretability (20 groups)

We can get interpretability by calculating similarities between securities and words

Visualization of Security vectors and word vectors using t-SNE



- Calculate the cosine similarities between cluster vectors and word vectors

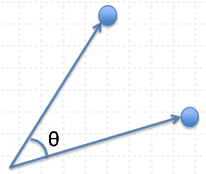
$$C_i = \frac{1}{N_i}\sum_{j=1}^{N_i} d_{i,j}$$

C_i : cluster vector

N_i : number of securities in cluster i

$d_{i,j}$: j-th securities in cluster i

$$sim(C, W) = \frac{C \cdot W}{\|C\|\|W\|}$$



Evaluation

Qualitative Evaluation: Interpretability (10 groups) by Word Cloud

The size of words is proportional to cosine similarity.

Cluster 1
Public Serv. Enterprise Inc.
NiSource Inc.
Southern Co.
Duke Energy
SCANA Corp.

Cluster 2
Chevron Corp.
Marathon Petroleum
Marathon Oil Corp.
ConocoPhillips
Pioneer Natural Resources

Cluster 3
Simon Property Group Inc.
Prologis
Plum Creek Timber Co.
Public Storage
American Tower Corp.

Evaluation

Qualitative Evaluation: Interpretability (10 groups) by Word Cloud

The size of words is proportional to cosine similarity.

Cluster 4
Morgan Stanley
Wells Fargo
JPMorgan Chase & Co.
Principal Financial Group
Bank of America Corp.

Cluster 5
Johnson & Johnson
Gilead Sciences
Allergan, Plc
Amgen Inc.
Bristol-Myers Squibb

Cluster 6
PepsiCo Inc.
The Coca Cola Co.
Monster Beverage
McDonald's Corp.
Kraft Heinz Co

Evaluation

Qualitative Evaluation: Interpretability (10 groups) by Word Cloud

The size of words is proportional to cosine similarity.

Cluster 7	Cluster 8	Cluster 9	Cluster 10
Wal-Mart Stores	Berkshire Hathaway	The Walt Disney Co.	Apple Inc.
Home Depot	Exxon Mobil Corp.	Time Warner Inc.	Amazon.com Inc.
Costco Co.	General Electric	Scripps Networks Int.	Alphabet Inc. Class C
Lowe's Cos.	Boeing Co.	Twenty-First Century Fox	Facebook, Inc.
O'Reilly Automotive	Raytheon Co.	Twenty-First Century Fox Inc Class B	Alphabet Inc. Class A

Evaluation

Quantitative Evaluation: Inter- and Intra-Industry Tests of Homogeneity

We follow Bhojraj et al. (2003) as the benchmark [8]

Table IV: Financial Variables used in evaluation

G	Variable	Calculation
1	Calendar month-end returns (RET)	Returns from month-end to month-end
2	Year end price-to-book (P/B)	Market capitalization / total common equity
	Enterprise value-to-sales (EVS)	(Market capitalization + debt)/ net sales
	Price-to-earnings (P/E)	Market capitalization / income before extraordinary items
3	Return-on-assets (ROA)	Income before extraordinary items / total assets
	Return-on-equity (ROE)	Income before extraordinary items / total common equity
	Profit margin (PM)	Income before extraordinary items / total common equity
	Leverage (LEV)	Total liabilities / total common equity
	Asset turnover (AT)	Total assets / net sales
	Current ratio (CR)	Total current assets / total current liabilities
4	One-quarter-ahead sales growth (SGR)	(1 quarter ahead - current net sales) / current year net sales
	R&D	Research and development expense / net sales

Notes – Group 1: Ecomonic relatedness, Group 2: Accounting measures, Group 3: Firm-level ratios, Group 4: Financial Information

- Following Bhojraj et al. (2003) as the benchmark, we propose to evaluate BTIC quantitatively across 12 variables commonly used in capital market research

Evaluation

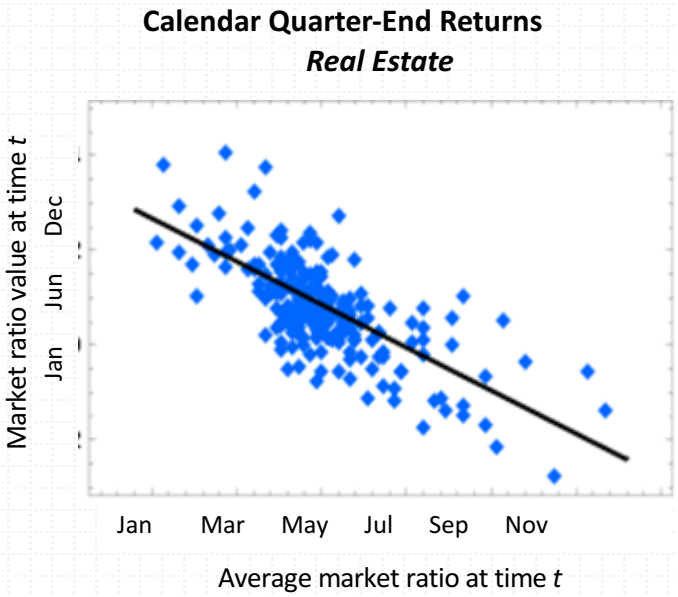
Quantitative Evaluation: Inter- and Intra-Industry Tests of Homogeneity
We follow Bhojraj et al. (2003) as the benchmark [8]

- As a measure of homogeneity for each industry cluster, we look at R^2 value of the univariate regression:

$$var_{i,t} = \alpha + \beta \cdot var_{k,t} + \varepsilon_{i,t}$$

where $var_{i,t}$ = market ratio variable for security *i* within a particular industry group *k* at time t, and $var_{k,t}$ = the industry *k*'s average at time *t*

- Then, R^2 would represent the portion of variations in the subject market ratio, explained by the variations in the market ratio of the corresponding industry group on average



Evaluation

Quantitative Evaluation: Inter- and Intra-Industry Tests of Homogeneity

We follow Bhojraj et al. (2003) as the benchmark [8]

Table V: Intra-Industry tests of homogeneity

	SIC		GICS		TF-IDF		BTIC	
	(A)	(I)	(A)	(I)	(A)	(I)	(A)	(I)
Level 1	0.11		0.13		0.10		0.13	
Level 2	0.49	0.38	0.20	0.07	0.16	0.06	0.19	0.06
Level 3	0.72	0.23	0.37	0.17	0.25	0.09	0.32	0.13

Table VI: Inter-industry tests of homogeneity

SIC	(R)	GICS	(R)	TF-IDF	(R)	BTIC	(R)
Lv. 1	0.11	Lv. 1	0.13	Lv. 1	0.10	Lv. 1	0.13
–	–	Lv. 2	0.20	Lv. 2	0.16	Lv. 2	0.19
Lv. 2	0.49	Lv. 3	0.37	Lv. 3	0.25	Lv. 3	0.32
TF-IDF vs SIC		TF-IDF vs GICS		BTIC vs SIC		BTIC vs GICS	
-0.01		-0.03		0		0	
–		-0.04		–		-0.01	
-0.23		-0.12		-0.17		-0.05	

Notes – Lev. # denotes the #-th division level of the subject classification scheme.

- **Intra-Industry Homogeneity**: Each bundle should contain entities that are "similar" types of business given their market activities
- **Inter-Industry Homogeneity**: Each bundle's power of representation should be similar from one another
- BTIC performs just as well as SIC and GICS in terms of intra- and inter-industry homogeneity

Discussion

Contribution : Four main contributions

---

1) Objectivity

- We cluster companies based on their self-identification as disclosed on the Form 10-Ks
- Data-driven, and requires no human effort in the process

2) Interpretability

- Clustering based on simple market numerics (stock prices, etc.) offer not much interpretability
- Our clusters offer “stories” from which one can tell “why” select companies are clustered together

3) Flexibility

- Our clustering methodology allows users to change the number of clusters flexibly according to their needs, simply by controlling threshold values

4) Automation

- We collect and process hundreds and thousands of corporate documents automatically and can cluster companies on real-time basis

Discussion

Future Work & Limitations & Expected Work

1. Expand text corpus to include greater number of firms in the analysis

2. Include market conception index when calculating similarities between the securities

3. Employ other means of clustering techniques and compare the results
   - Ensemble clustering
   - Multiple membership
   - Trajectory analysis

Reference

[1] Phillips, R. L., and R. Ormsby. "Industry classification schemes: An analysis and review." Journal of Business & Finance Librarianship. Vol. 21, No. 1. (2016) pp:1-25.

[2] United States Department of Labor. “SIC Divions Structure.” Occupational Safety and Health Administration. http://www.osha.gov/pls/imis/sic_manual.html (2016)

[3] MSCI Inc. “Global Industry Classification Standard (GICS).” (2016)

[4] Le, Q. V., and T. Mikolov. “Distributed representations of sentences and documents.” ICML. Vol. 14. (2013) pp: 1188 – 2296.

[5] Ward Jr., J. H. “Hierarchical Grouping To Optimize an Objective Function.” Journal of the American Statistical Association. Vol. 58, No. 301. (1963) pp: 236-244.

[6] Collobert, R., and J. Weston. “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning.” Proceedings of the 25th International Conference on Machine Learning. (2008) pp: 160-167.

[7] Das, R., M. Zaheer, and C. Dyer/ “Gaussian LDA for Topic Models with Word Embeddings.” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. (2015)

[8] Bhojraj, S., C. Lee, and D. K. Oler. "What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research." Journal of Accounting Research. Vol. 41, No. 5. (2003) pp: 745-774.