Text-based Industry Classification by using Autoencoder

Kyounghun Bae[1], Daejin Kim[2], Rocku Oh[2*]

[1]Hanyang University

[2]Ulsan National Institute of Science and Technology

Outline

- ✓ **Introduction**
- ✓ **Problem statement**
- ✓ **Methodology**
  - Bag of Words representation
  - Dimensionality reduction using Autoencoder
  - Spherical clustering
- ✓ **Result**
- ✓ **Conclusion**

Introduction

- ✓ **A variety of approaches to group homogenous firms to analyze industry-based studies has been conducted.**
  - Corporate reorganization
  - Changes in financial and investment policy
  - Credit rating
- ✓ **Industry classification systems**
  - Standard Industrial Classification (SIC) uses information on selling end products and production process (Chan, Lakonishok, & Swaminathan, 2007)
  - Fama and French (1997) proposes 49-industry classification by merging several ranges of SIC codes.
  - North American Industry Classification System (NAICS)
  - Global Industry Classification System (GICS) is widely used by the investment analysts and portfolio managers. The system is based not only on operational characteristics of firms also on the investors' perceptions of what constitutes the firm's mainstream of their business (Kile & Phillips, 2009)

Problem Statement

- ✓ **To overcome the drawbacks of the previous industrial classification systems, *Hoberg and Phillips (2016)* propose a new text-based industry groups**
  - The system is based on a strong tendency of vocabulary usage among firms operating in the same market(industry) reported to the Securities and Exchange Commission (SEC)
  - They analyze text descriptions at the level of the word(vocabulary) from the annual report
  - They measure the pairwise cosine similarity of the word vectors extracted from the reported document of firms
  - The system is able to capture the changes in firms' business, namely diversification and pivoting as well.

Problem Statement

Limitation of previous approach

- ✓ The similarity measure can only represent firm-to-firm information, not firm-to-industry or industry-to-industry.
  - Cannot infer the overall map of industry closeness and relationship although they validate the across industry variation of their final clusters.

- ✓ Cosine similarity measure by high dimensional vectors cannot escape from the curse of dimensionality problem *(Skillicorn, 2012)*.
  - The dimension of word vectors used in their research is larger than 60,000
  - The word vectors are highly sparse as well.
  - The distance(similarity) measure may not even be qualitatively meaningful in high dimensional space *(Aggarwal, Hinneburg, & Keim, 2001)*.

Dimensionality reduction

Reducing the Dimensionality of Data with Neural Networks

(G.E Hinton and R. R. Salakhutdinov, 2006)*

- ✓ High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors.
  - Deep-Autoencoder networks works much better than principal components analysis as a tool to reduce the dimensionality of data

- ✓ They use “Bag-of-Words” representation as input vector
  - The Reuters Corpus Volume II contains 804,414 newswire stories
  - Each article is represented as a vector containing the counts of the most frequently used 2000 words in the training dataset.

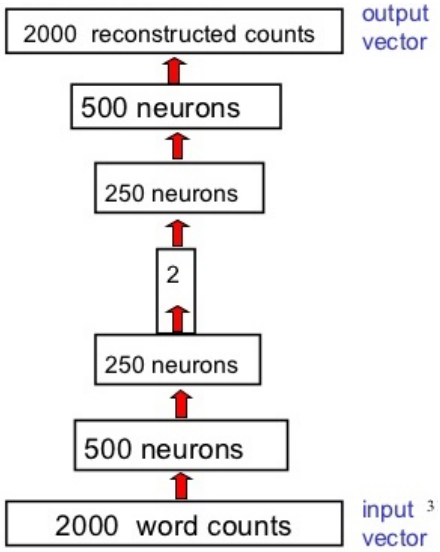
Dimensionality reduction

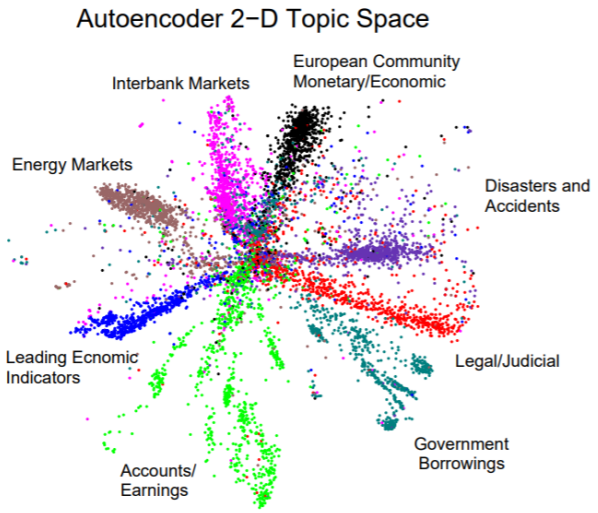
Reducing the Dimensionality of Data with Neural Networks

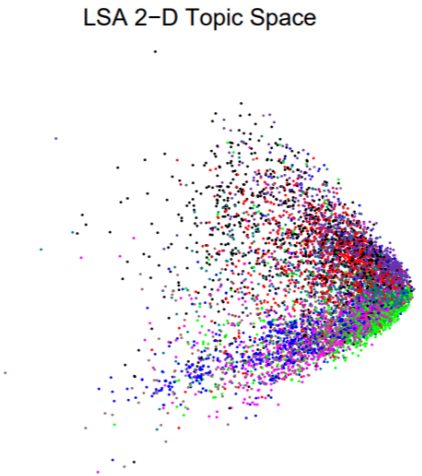
(G.E Hinton and R. R. Salakhutdinov, 2006)*

Instead of using codes to retrieve documents, we can use 2-D codes to visualize sets of documents.

- This works much better than 2-D PCA







(G.E Hinton* and R. R. Salakhutdinov, 2006)

Data

SEC 10-K Filings

- ✓ We collect 10-K annual reports filed by the Securities and Exchange Commission (SEC) from 2013 to 2016 using web crawling algorithm which results in the total number of 21,631 10-K reports.
- ✓ "Item 1. Business" part contains specified product description of firms

Firm 1: SANDISK CORP (SIC code: 3572)
Business: Flash Memory Storage Core words: memory(67), product(52), technology(44), storage(36), market(31), device(31), solution(28), NAND(26), flash(24), drive(20), manufacturer(19), design(19), corporation(18), venture(18), card(18), president(17), data(16), wafer(16), cost(15), year(15)
Firm 2: SCHEIN (HENRY) INC (SIC code: 5047)
Business: Healthcare Distribution Core words: health(102), product(89), care(65), service(62), state(47), customer(47), law(44), practice(41), president(40), business(40), distribution(37), sale(35), drug(33), act(30), vice(30), practitioner(26), officer(25), technology(24), order(23), management(23)
Firm 3: IMPAC MORTGAGE HOLDINGS INC (SIC code: 6162)
Business: Long-term Portfolio Core words: mortgage(174), loan(159), origination(53), portfolio(49), service(45), estate(34), operation(33), channel(33), Mae(30), interest(30), correspondent(29), lending(27), rate(27), credit(21), security(21), broker(20), sale(20), borrower(19), seller(18), act(17)

Methodology

Bag-of-Words representation

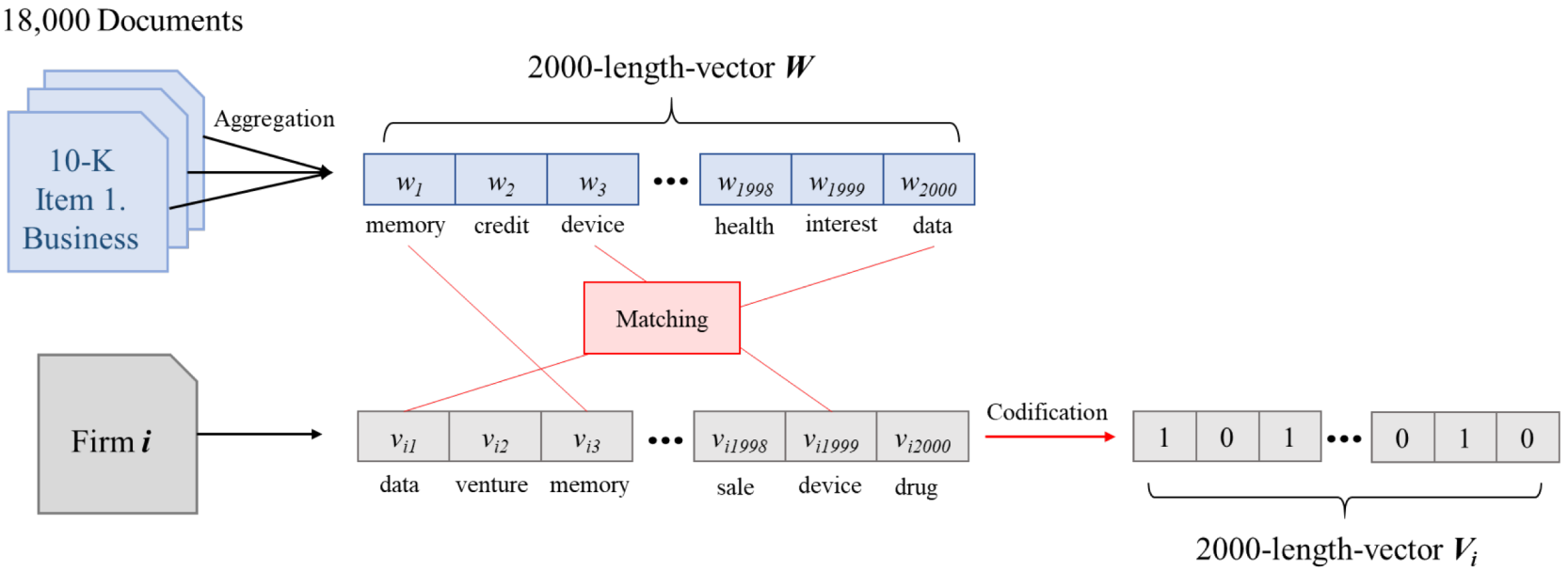
- ✓ The underlying hypothesis is based on the notion that firms classified in the same industry use more similar words to describe and offer their business and products than the firms classified in the different industries

Unique words out of 2000 words in the bag-of-words	Occurrences
Case 1 - Healthcare, Medical Equipment, and Drugs	
hospital, billing, physician, Medicaid, productivity, patient, submission, reimbursement, Medicare, referral	9 times out of 9 documents
beneficiary, methodology, recruitment, length, abuse, accountability, authorization, accreditation, CM, associate, prohibition, utilization, therapy, transition, employer, sanction, eligibility, safeguard, notification, fraud, worker, HIPPA(Health Insurance Portability and Accountability Act), spending, portability, admission, antikickback, update	8 times out of 9 documents
Case 2 - Oil, Gas, and Coal Extraction and Products	
crude, commodity, pipeline, hydrocarbon, petroleum, transport	8 times out of 8 documents
proximity, carrier, cleanup, liquid, pollution, barrel, index, discharge, mile, tank, basin, emergency, exploration, drilling, commerce, injection, FERC(Federal Energy Regulatory Commission), shale, formation, greenhouse, dioxide, emission, gathering, fuel	7 times out of 8 documents

Methodology

Bag-of-Words representation

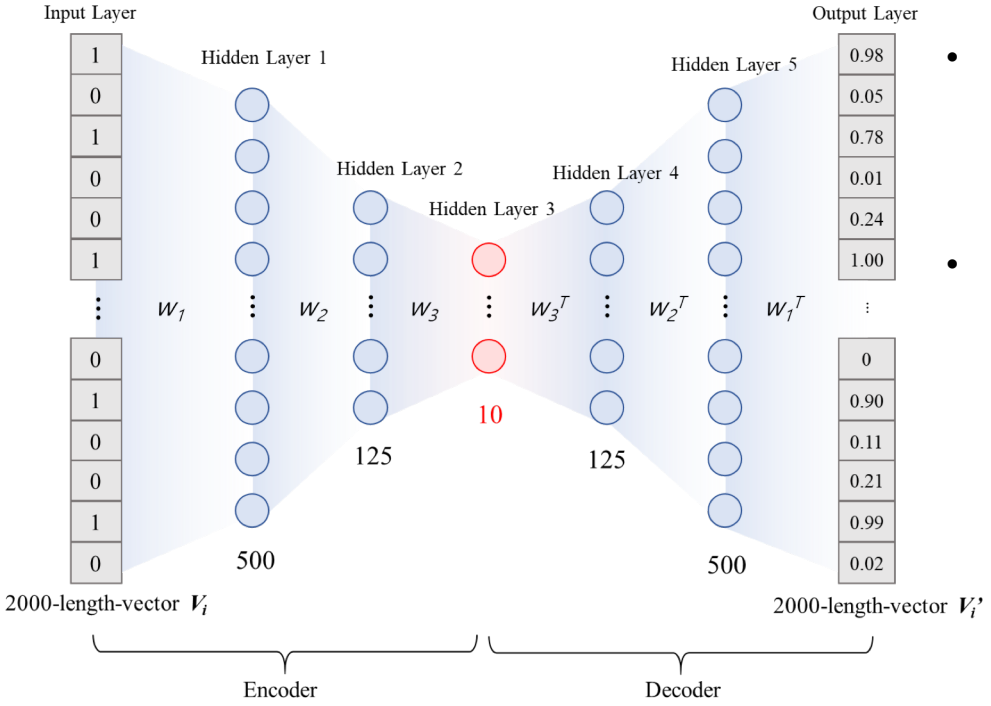
- ✓ We remove common words which appear more than 20% of documents during the preprocessing.
- ✓ We construct bag-of-words vector ***W*** which uses 2000-unique-word in order of frequent appearance among all the unique words.
- ✓ A business description of given firm *i* is converted to a 2000-dimensional binary (coded) vector $\boldsymbol{V}_i$



Methodology

Dimensionality reduction by using the autoencoder

- ✓ This high dimensional and sparse vector space arises an issue of the curse of dimensionality when computing the cosine similarity and applying it to the clustering method directly.
- ✓ We reduce the number of dimensions of feature while minimizing the cross entropy between the input vector and the reconstructed output vector



- $Reconstruction\ V_i' = Decoder_{W'}\left(Encoder_W(input\ V_i)\right)$
- $W^\star, W'^\star = \underset{W,W'}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} L(V_i, V_i')$

$$where\ L(V_i, V_i') = -\sum_{k=1}^{d}[V_{ik}\log(V_{ik}') + (1 - V_{ik})\log(1 - V_{ik}')]$$

Methodology

Industry classification by the Spherical K-means clustering

- ✓ The direction of a word vector is more important than the magnitude itself
- ✓ We use spherical K-means clustering algorithm which is a suitable for the vector space model
- ✓ The spherical k-means algorithm maximizes the average cosine similarity within the clusters.

Result

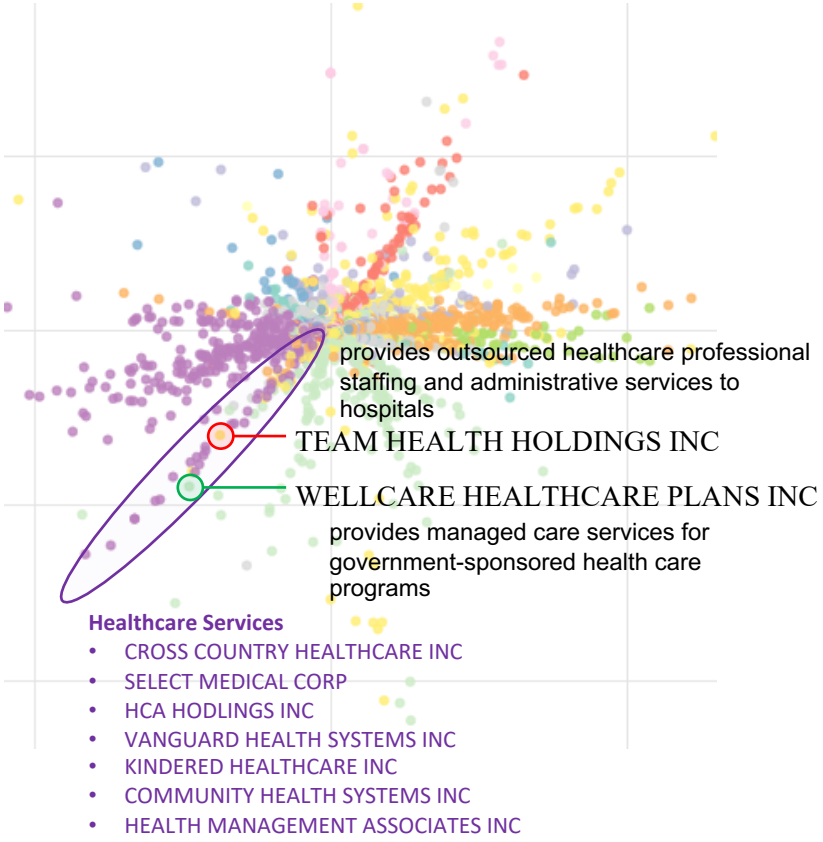
Two-dimensional representation of industry space

✓ 2D scatter plot for values from the last encoding layer after training

Result

Qualitative Analysis : Case 1 (Healthcare-related industry)

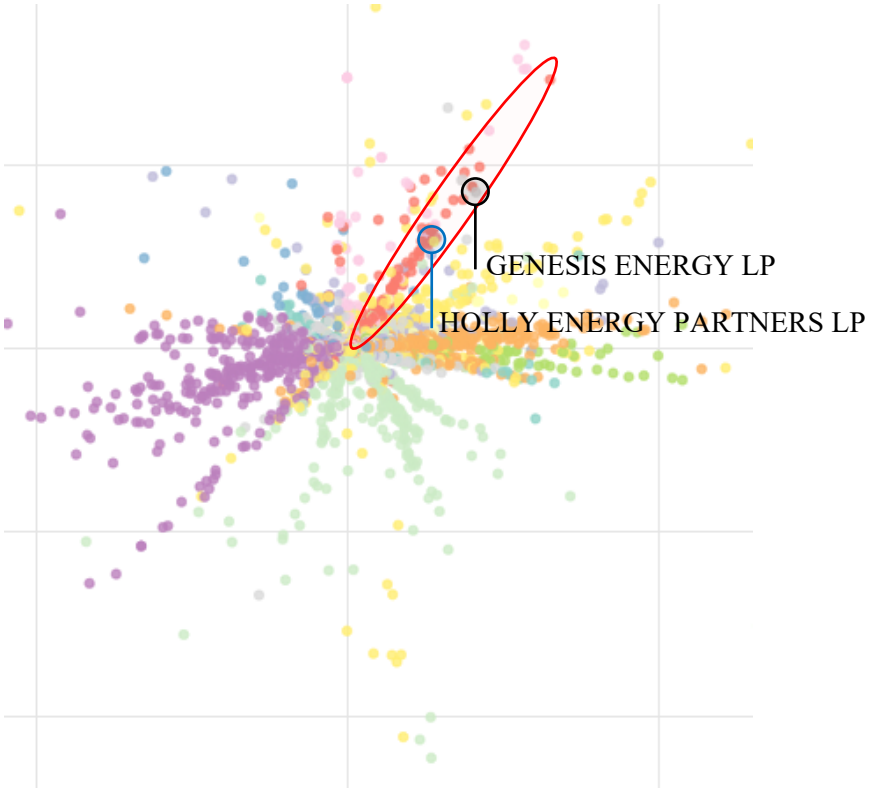
- ✓ Some firms related to healthcare products and process are within the purple ellipse along the similar direction (angle).
- ✓ The SIC code of firm “TEAM HEALTH HOLDINGS INC” is 7363 (Fama-French classification code 12; Others). “WELLCARE HEALTH PLANS INC” is coded as the SIC code of 6324 (Fama-French classification code 11; Money).
  - The two firms are clustered in the same industry based on the spherical K-means clustering method.
  - The other firms clustered with the two firms have SIC code range of 8000-8099 (Health Service).



Result

Qualitative Analysis : Case 2 (Energy-related industry)

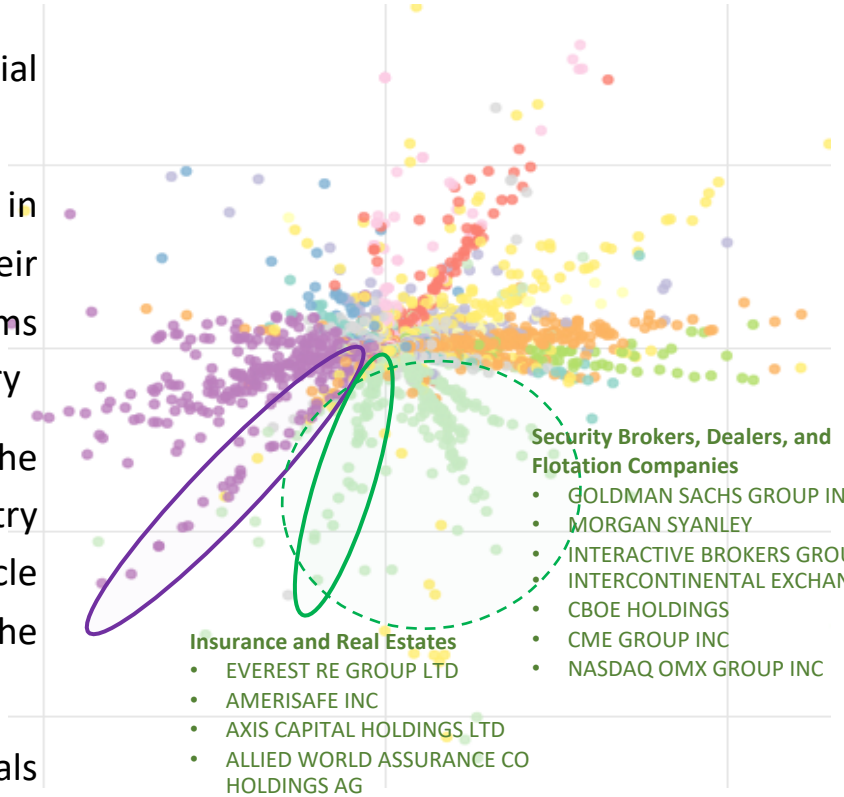
- ✓ Firms with in the red ellipse is related to the energy and mining industry including oil, gas, and materials.
- ✓ “GENESIS ENERGY LP” and “HOLLY ENERGY PARTNERS LP” belong to “Shops” in terms of the Fama-French classification code (9) and SIC code(5171)
    - The scatter point of the two firms are closer to the “Energy” firms



Result

Qualitative Analysis : Case 3 (Sub-industry issue)

- ✓ Four or five sub-groups are seen within financial industry colored as a green dashed-circle.
  - Each spike indicates different clusters in terms of the word presented in their business descriptions even though the firms are all related to money or financial industry
- ✓ The firms located in the green ellipse is one of the sub-groups of financial industry. The sub-industry is closer to the industry group in the purple circle (Case 1 group) than other sub-groups within the financial industry
  - Most of the firms in Case 1 groups are deals with the insurance related to healthcare.



Result

Quantitative Analysis

✓ **Within-industry variation**

- Compute standard deviation of individual groups by year→ Compute an industry-size-weighted average of the standard deviations
- *Smaller is the better*

✓ **Across-industry variation**

- Compute the firm-size-weighted average of individual groups by year → Compute a standard deviation of the averages
- *Larger is the better*

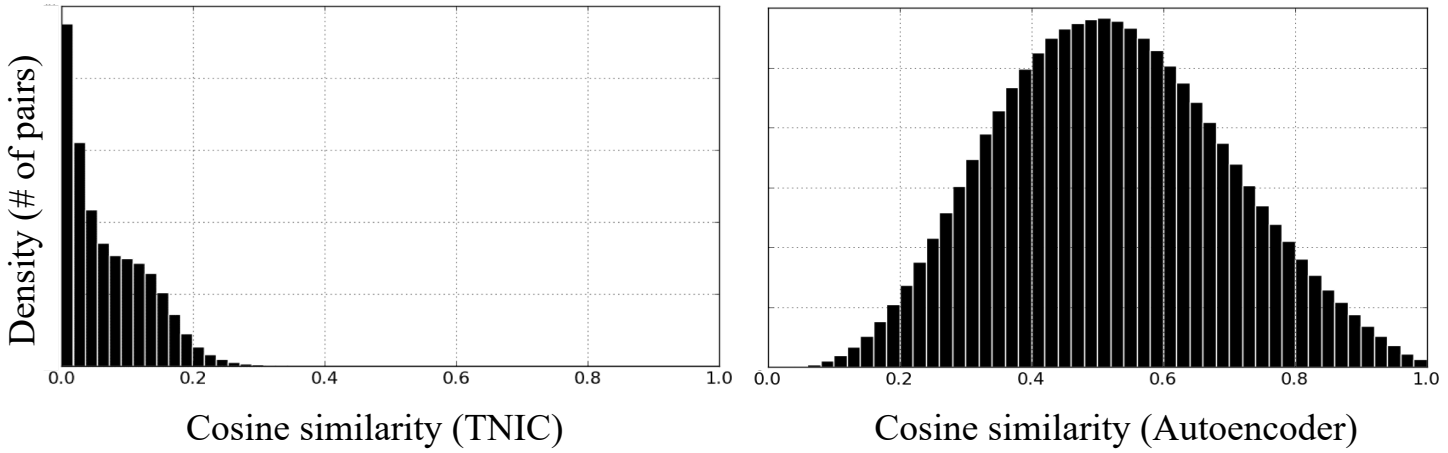
		Within-industry variations			Across-industry variations		
Classification method	N	Weighted OI/asset	Weighted OI/sales	Weighted Market β	Weighted OI/asset	Weighted OI/sales	Weighted Market β
SIC 3-digit	245	0.126	18.296	0.884	0.390	0.066	0.741
GICS_subind	157	0.143	13.823	**0.803**	3.455	0.136	0.619
TNIC 300 fixed code	300	0.130	10.243	0.980	4.493	0.139	0.809
Autoencoder + SKmeans	300	**0.113**	**5.857**	0.856	**10.819**	**0.150**	**0.924**
TNIC		**0.124**	8.655	1.055	**0.125**	19.081	0.678
Autoencoder + TNIC		0.132	**4.250**	**0.996**	0.115	**20.103**	**0.703**

Result

Quantitative Analysis

✓ **Cosine Similarity**

- TNIC similarity scores are highly skewed. The fact indicates computing a cosine similarity measure of high dimensional vectors directly is inappropriate, causing curse of dimensionality problem.



Classification method	Mean	Standard Deviation	Min	Max
TNIC	0.073	0.063	0.000	0.904
Autoencoder	0.521	0.173	0.011	0.988

Conclusion

- ✓ We collect 10-K annual reports from the Securities and Exchange Commission (SEC) using web crawling algorithm to extract business description text data of each firm.
- ✓ We use a deep learning method, which is called autoencoder, as a dimensionality reduction technique to reduce the dimension of original high dimension and sparse word vector to mitigate a curse of dimensionality problem in vector space.
- ✓ We clusters firms using the reduced features by spherical K-means clustering algorithm which is a suitable for the vector space model.

- ✓ **We are able to visualize similarity and closeness between industries as well as firms.**
- ✓ **We qualitatively shows several mis-classified firms by proposed method and visualization.**
- ✓ **We quantitatively validate the performance of proposed method by within and across the variations of clusters(industries)**

Thank you

감사합니다

Further research

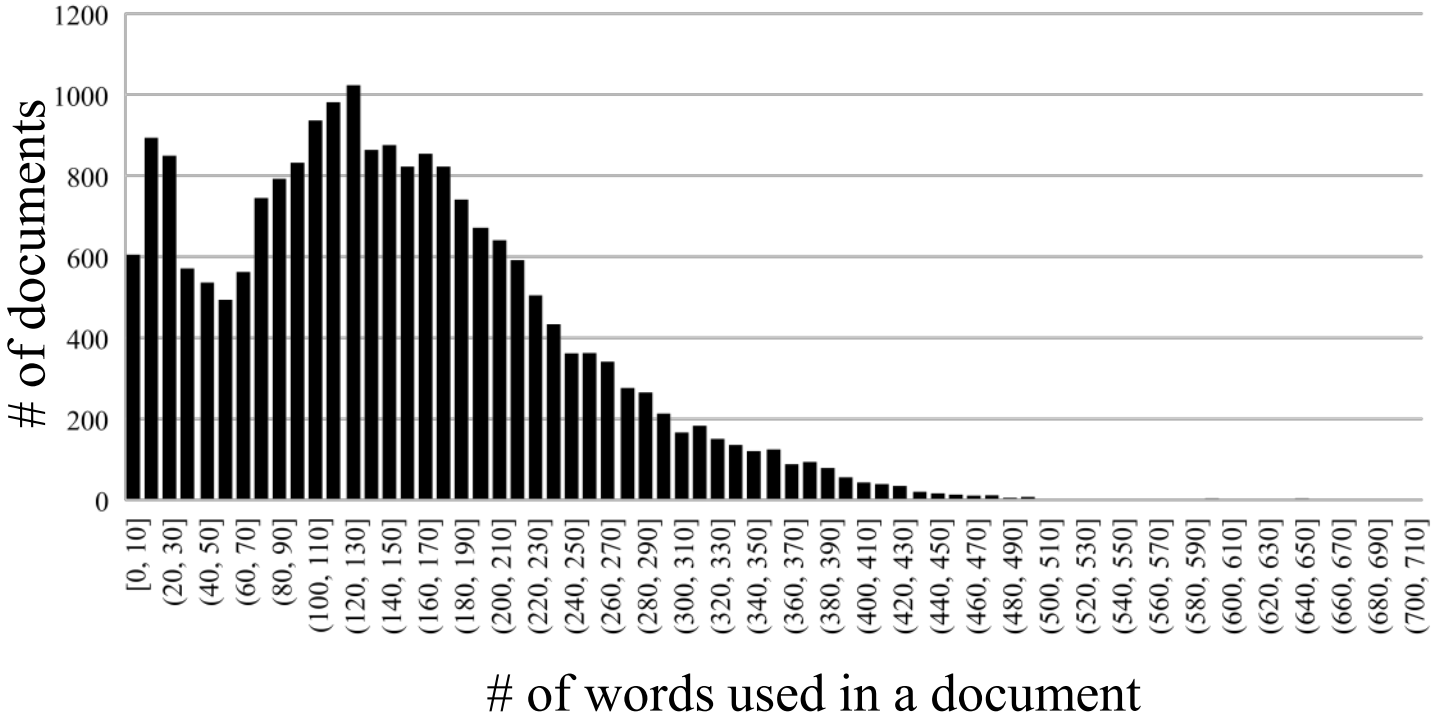
Clustering Method

- Optimal # of words for BOW
- Optimal # of clusters (industries)
- Optimal # of node(features) of encoder

Clustering result are based on NOUNS only

- Clustering result
  - focus on firms in same product and process chain
- Spanning the level of a token – Include verbs and several words as a token
  - *We produce corn chips from micro kernels that our customers sell by wholesale.*
  - *We sell by wholesale micro chips and kernels that our customers use to produce corn.*
- Novel technique to convert documents to vector notation
  - E.g.) Doc2Vec with Deep Convolutional Autoencoder

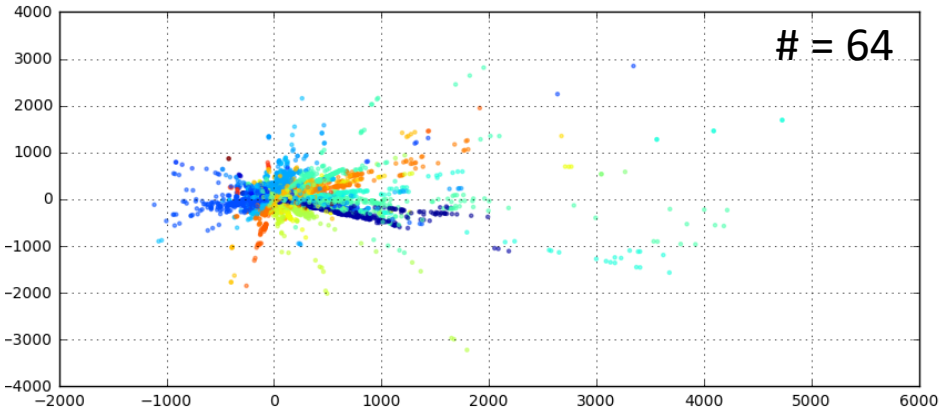
Appendix


1200
1000
800
600
400
200
0
# of documents
[0, 10]
(20, 30]
(40, 50]
(60, 70]
(80, 90]
(100, 110]
(120, 130]
(140, 150]
(160, 170]
(180, 190]
(200, 210]
(220, 230]
(240, 250]
(260, 270]
(280, 290]
(300, 310]
(320, 330]
(340, 350]
(360, 370]
(380, 390]
(400, 410]
(420, 430]
(440, 450]
(460, 470]
(480, 490]
(500, 510]
(520, 530]
(540, 550]
(560, 570]
(580, 590]
(600, 610]
(620, 630]
(640, 650]
(660, 670]
(680, 690]
(700, 710]
# of words used in a document

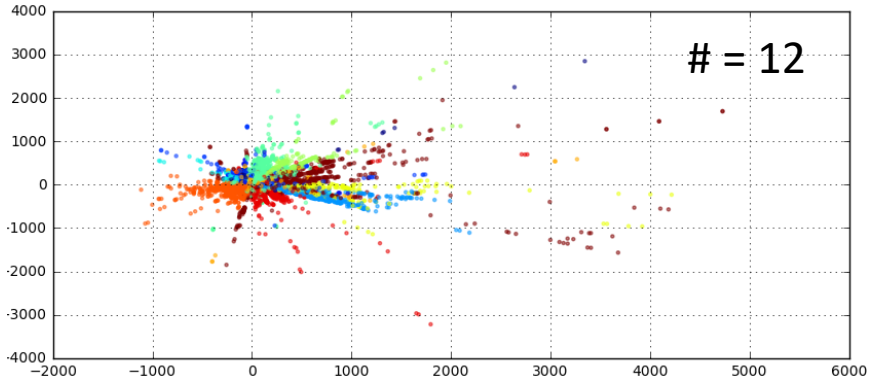
Appendix

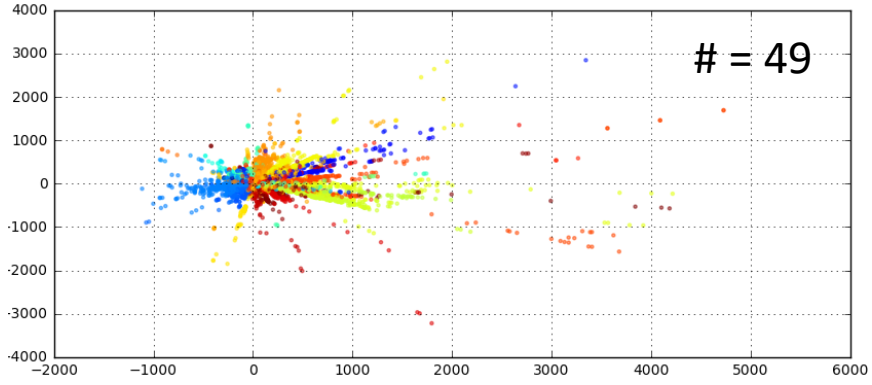
Comparison between classification systems

- 2-digit SIC code (# = 64)



- Fama-French 49 Industry classification code

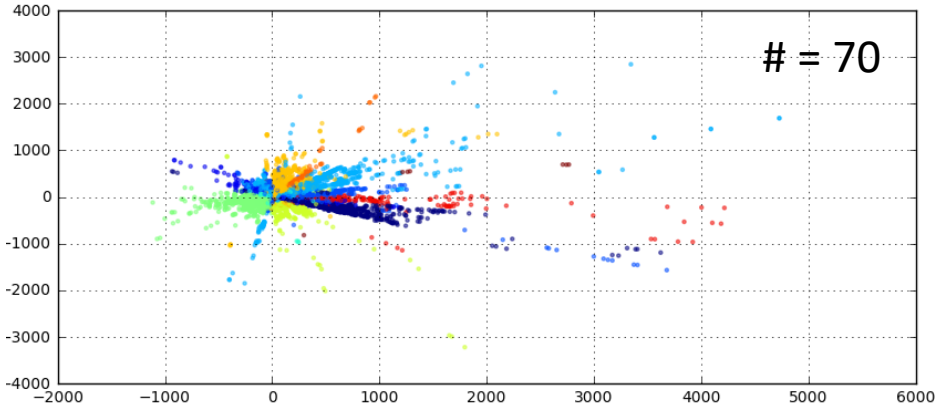




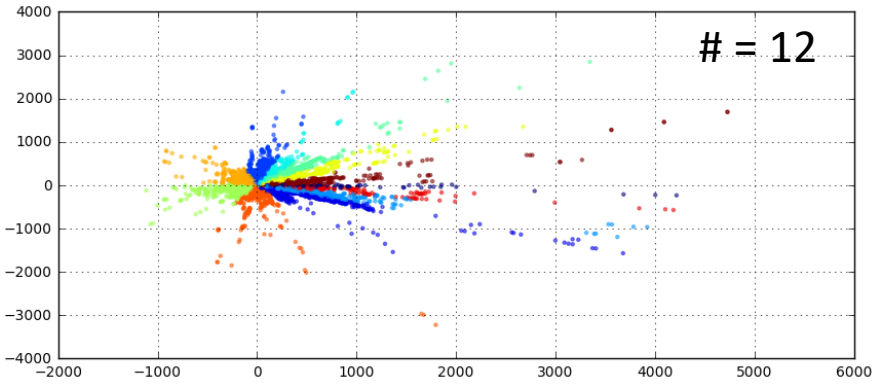
Appendix

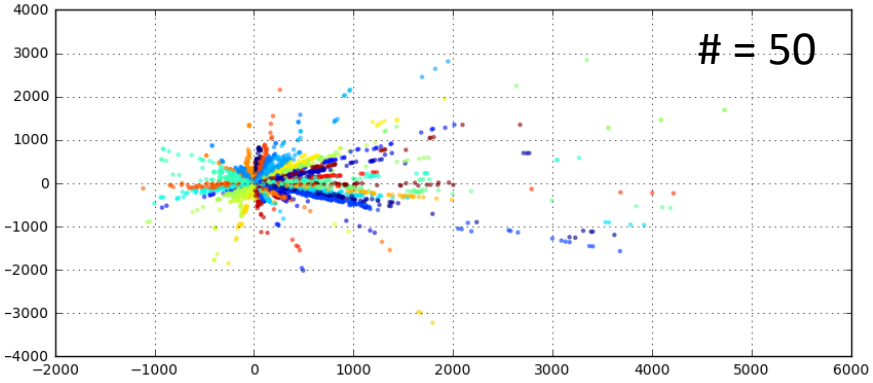
Comparison between classification systems

- GICS (# = 70)



- Spherical k-means clustering





Appendix

Table 1. Sample firms having different SIC code ranges but allocated to the same label of cluster.

Firm name	SIC code	Fama-French 12 Classification code		Clustered code
Case 1 - Healthcare, Medical Equipment, and Drugs				
TEAM HEALTH HOLDINGS INC	7363	12	Others	11
WELLCARE HEALTH PLANS INC	6324	11	Money	11
SELECT MEDICAL HOLDINGS CORP	8069	10	Hlth	11
SYMBION INC TN	8011	10	Hlth	11
LHC GROUP INC	8082	10	Hlth	11
LIFEPOINT HOSPITALS INC	8062	10	Hlth	11
TENET HEALTHCARE CORP	8062	10	Hlth	11
AMN HEALTHCARE SERVICES INC	8090	10	Hlth	11
HCA HOLDINGS INC	8062	10	Hlth	11
Case 2 - Oil, Gas, and Coal Extraction and Products				
GENESIS ENERGY LP	5171	9	Shops	4
CROSSTEX ENERGY LP	5172	9	Shops	4
HOLLY ENERGY PARTNERS LP	4613	12	Others	4
GULFPORT ENERGY CORP	1311	4	Energy	4
CONTINENTAL RESOURCES INC	1311	4	Energy	4
UNIT CORP	1311	4	Energy	4
MID CON ENERGY PARTNERS LP	1311	4	Energy	4
CHEVRON CORP	2911	4	Energy	4