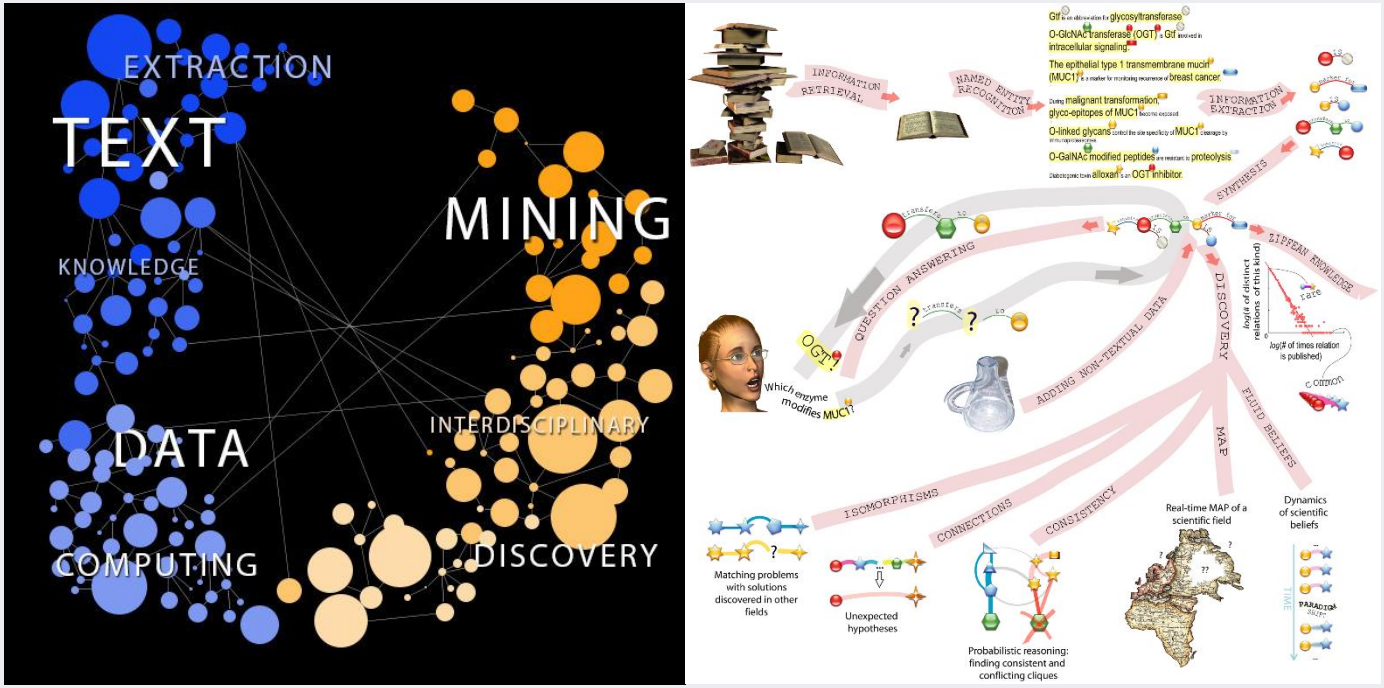


Lecture 1: Introduction to Text Mining

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Text Mining Overview

02 TM Process 1: Collection & Preprocessing

03 TM Process 2: Transformation

04 TM Process 3: Dimensionality Reduction

05 TM Process 4: Learning & Evaluation

Text Mining: Background

- Motivation
  - ✓ Approximately 80% of the world's data is help in unstructured formats
  - ✓ Simple document retrieval is not enough, but knowledge discovery is required!

Topic: Storage

Follow via:

Within two years, 80% of all medical data will be unstructured

Summary: *The data storage load for medical organizations is going to skyrocket.*

By Denise Amrich for ZDNet Health | April 9, 2013 -- 01:09 GMT (18:09 PDT)

Follow @deniseamrich

We are all aware of the growth in medical health records because of changes in regulation. You would think, then, that most of the storage requirements would be for managing and storing those health records.

But the following IBM video states that 80 percent of health data stored in 2015 will be unstructured data, data that doesn't fit into nice rows and columns. Since nearly all medical records subject to the Affordable Care Act are, essentially, form data, that means that a huge amount of the storage required by the medical world isn't a result of that particular set of regulations.

In fact, IBM says that the body of medical knowledge doubles every five years. As imaging improves and M2M grows, more mobile data is captured. More intelligent sensors are deployed, both to home and hospital-bound patients. The data storage load for medical organizations is simply going to skyrocket.

This, of course, leads to another huge problem: Confidentiality and security. The Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health Act (HITECH) both require record-keeping safeguards. As medical data explodes in volume, that data protection challenge gets greater and greater.

The following video from IBM explores some of the scope of the problem. It's a marketing video, so keep that in mind, but it gives you a good understanding of the sorts of challenges and solutions that we'll be looking at for medical data storage.

http://www.zdnet.com/within-two-years-80-percent-of-medical-data-will-be-unstructured-7000013707/

How to manage unstructured data for business benefit

Stephen Pritchard, Contributor

Share 16 +1 0 Tweet 56

Businesses across all industries are gathering and storing more and more data on a daily basis. But when it comes to assessing the benefits and challenges of big data, sometimes it is easy to overlook one key point: Most of the business information in use today does not reside in a standard relational database.

An often-cited statistic is that 80% of business data is unstructured, be it in word processor, spreadsheet and PowerPoint files, audio, video, sensor and log data, or external data such as social media feeds.

Exploiting unstructured data

This raises some issues for organisations that need to process and exploit unstructured data. Some big data tools, primarily those based on Hadoop, are designed from the ground up to manage and analyse unstructured information. Other, more conventional business intelligence (BI) and data warehousing technologies may not be.

Business intelligence (BI) and data warehousing suppliers have been adding support for unstructured data management to their tool sets, and some IT organisations have built their own platforms for converting unstructured data into structured records, for example, through knowledge management systems. But that can be a time-consuming and expensive process.

http://www.computerweekly.com/feature/How-to-manage-unstructured-data-for-business-benefit

Text Mining: Background

- Kensho
  - ✓ Computational knowledge engine for the financial industry (https://www.kensho.com)

Rise of the Machines

"The worry that AI could do to white-collar jobs what steam power did to blue-collar ones during the Industrial Revolution is therefore worth taking seriously. Examples, such as Narrative Science's digital financial journalist and Kensho's quant, abound. Kensho's system is designed to interpret natural-language search queries such as, "What happens to car firms' share prices if oil drops by $5 a barrel?" It will then scour financial reports, company filings, historical market data and the like, and return replies, also in natural language, in seconds. Kensho supplies the software to big banks and sophisticated traders."

The Economist
May 9th, 2015
View Article

FT
FINANCIAL TIMES

Goldman Sachs leads $15m financing of data service for investors

Goldman Sachs has emerged as the largest investor in a financial analytics start-up that enables institutions to mine a wealth of big data, underscoring Wall Street's drive to tap new technology.

FINANCIAL TIMES
November 23, 2014
View Article

Forbes
November 24, 2014

"Goldman Sachs Leads $15 Million Investment In Tech Start Up Kensho"

The New York Times
June 9, 2015

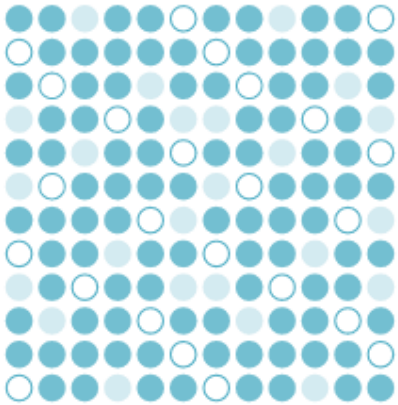
"Wall St. Courts Start-Ups It Once May Have Ignored"

On the first day of a hackathon at the Manhattan headquarters of Goldman Sachs, participants from the Wall Street bank showed up in suits. The programmers from Kensho, a start-up that had recently gotten money from Goldman, were there in jeans and ripped shirts.

Text Mining: Background

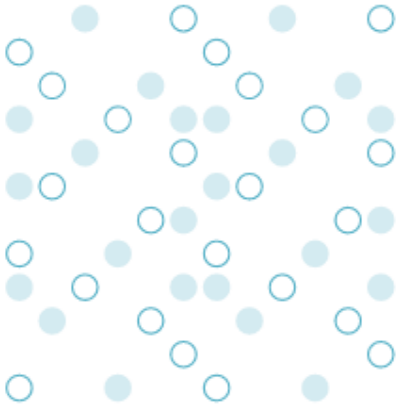
Methodology: How are the indices constructed?

The Kensho New Economy indices are constructed using an entirely systematic, rules-based methodology. Leveraging Kensho's proprietary natural language processing (NLP) platform, millions of pages of regulatory filings and other public information are analyzed to identify the companies involved in each of the New Economies. Proprietary algorithms determine the extent to which each company is involved in a given New Economy; those companies for which a New Economy is a principal components of its business strategy will be overweight in the index.



1 Scan

UNIVERSE OF PUBLIC COMPANIES



2 Filter

COMPANIES MATCHING THE NEW ECONOMY

3 Organize

COMPANIES ORGANIZED BY IMPACT

Text Mining: Background

Kensho Autonomous Vehicles℠ Index

New Economies
KENSHO 21ST CENTURY SECTORS

KCARS
234.78 ↓-0.2%
as of 03/07/2017

Index Live Date:
06/27/2016

↓ Methodology
↓ Factsheet
↓ Factsheet (Price Return)

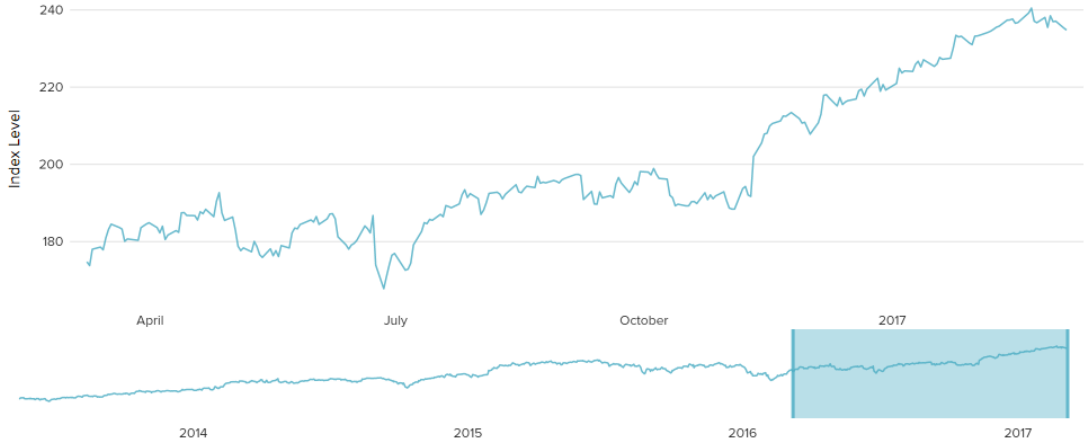
Transportation is a key enabler of the economy, and the global automotive industry itself generates US$5tn in annual output.[1] Companies involved in transporting people and goods are increasingly focused on introducing autonomous alternatives, those which can operate without human input using sensors, radars, cameras etc. Although fully self-driving vehicles are not yet in the mainstream, partially autonomous cars and trucks are already a reality today. Some of the main companies are well-known players in the automotive industry, but aspiring entrants continue to make significant inroads. The "connected cars" market could exceed $40bn in size by 2025,[2] and its growth will have far-reaching economic consequences, as it will transform how human and physical resources move about, and will impact productivity, location decisions, health, and even personal finances and vehicle ownership.

The Kensho® Autonomous Vehicles℠ Index uses an entirely rules-based methodology to objectively uncover companies leading the way forward toward autonomous transportation, whether directly through the the development of autonomous vehicles themselves or as part of the ecosystem supporting these initiatives. Only those companies with a market capitalization of $100mm and a 3 month average daily traded value (ADTV) of $1mm as of the most recent index selection date are eligible to be included in the index.

LEVEL

1W | 1M | 1Y | 3Y | Max

↓ Timeseries Download (CSV)



Any information presented prior to the index live date is hypothetical and back-tested, and is not a guarantee of actual performance. No assurances are given with respect to the future performance of any Kensho index. Hypothetical back-tested calculations are prepared with the benefit of hindsight, and may involve the use of proxy or substitute index components or index methodology adjustments. The actual future performance of any Kensho index may be higher or lower than the historical or hypothetical back-tested performance of any Kensho index.

Example: The BioTech Industry

Access to information is a serious problem

- 80% of biological knowledge is **only** in reasearch papers
- finding the information you need is prohibitively expensive

Humans do not scale well

- if you read 60 research papers/week...
- ...and 10% of those are interesting...
- ...a scientist manages 6/week, or 300/year

This is not good enough

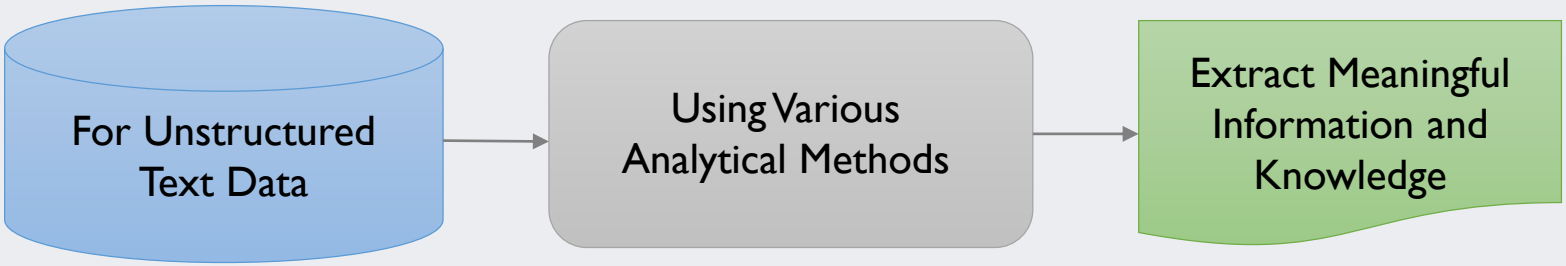
- MedLine adds more than 10 000 abstracts each *month*!
- Chemical Abstracts Registry (CAS) registers 4000 entities **each day**, 2.5 million in 2004 alone

[cf. Talk by Robin McEntire of GlaxoSmithKline at KBB'05]

Text Mining: Definition

Text Mining (Def. *Wikipedia*)

Text mining, also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value.

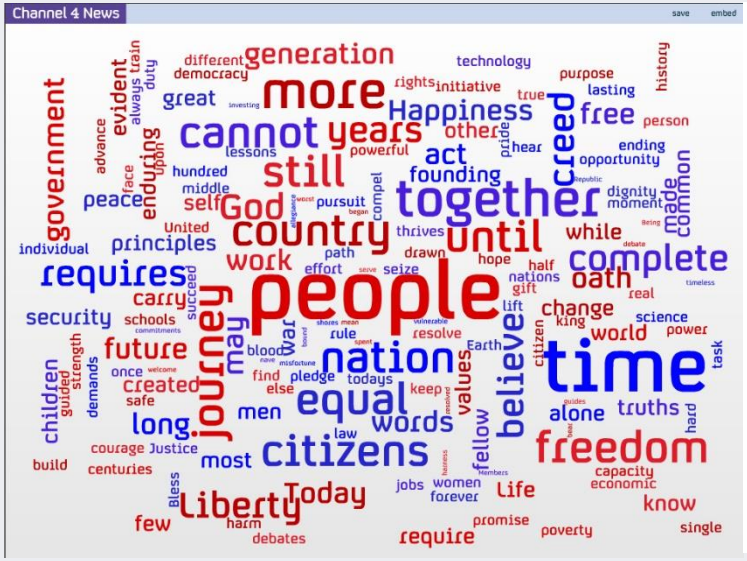


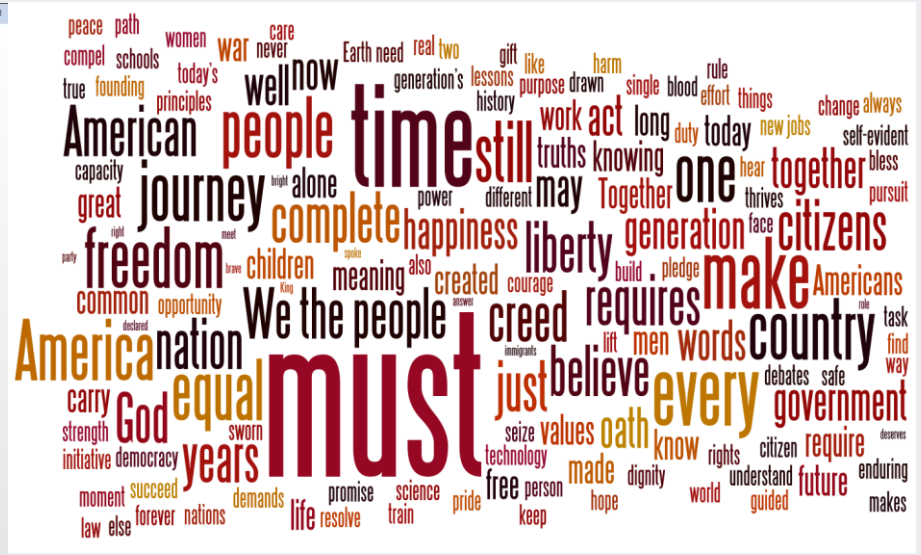
Text Mining: Applications

- Information Abstraction/Summarization/Visualization

Text Mining: Applications

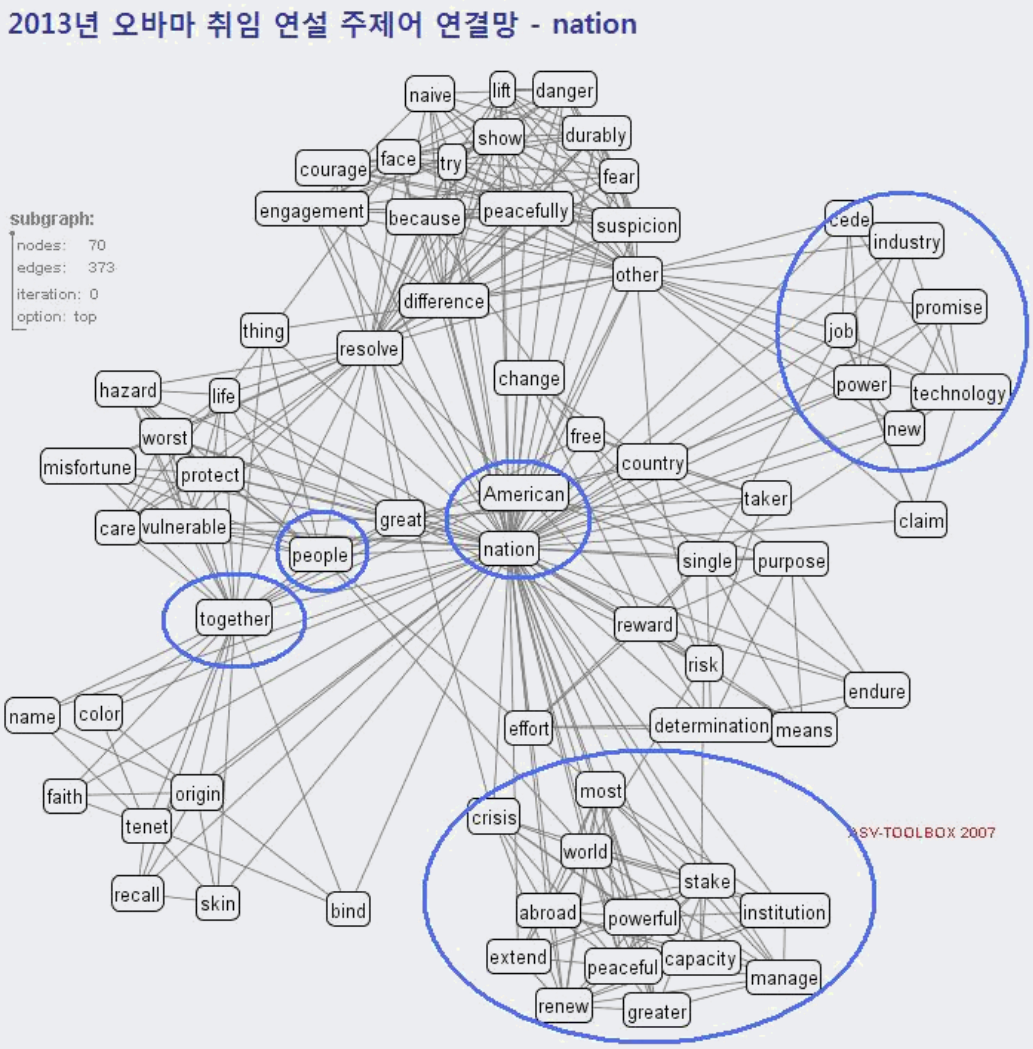
- Information Abstraction/Summarization/Visualization





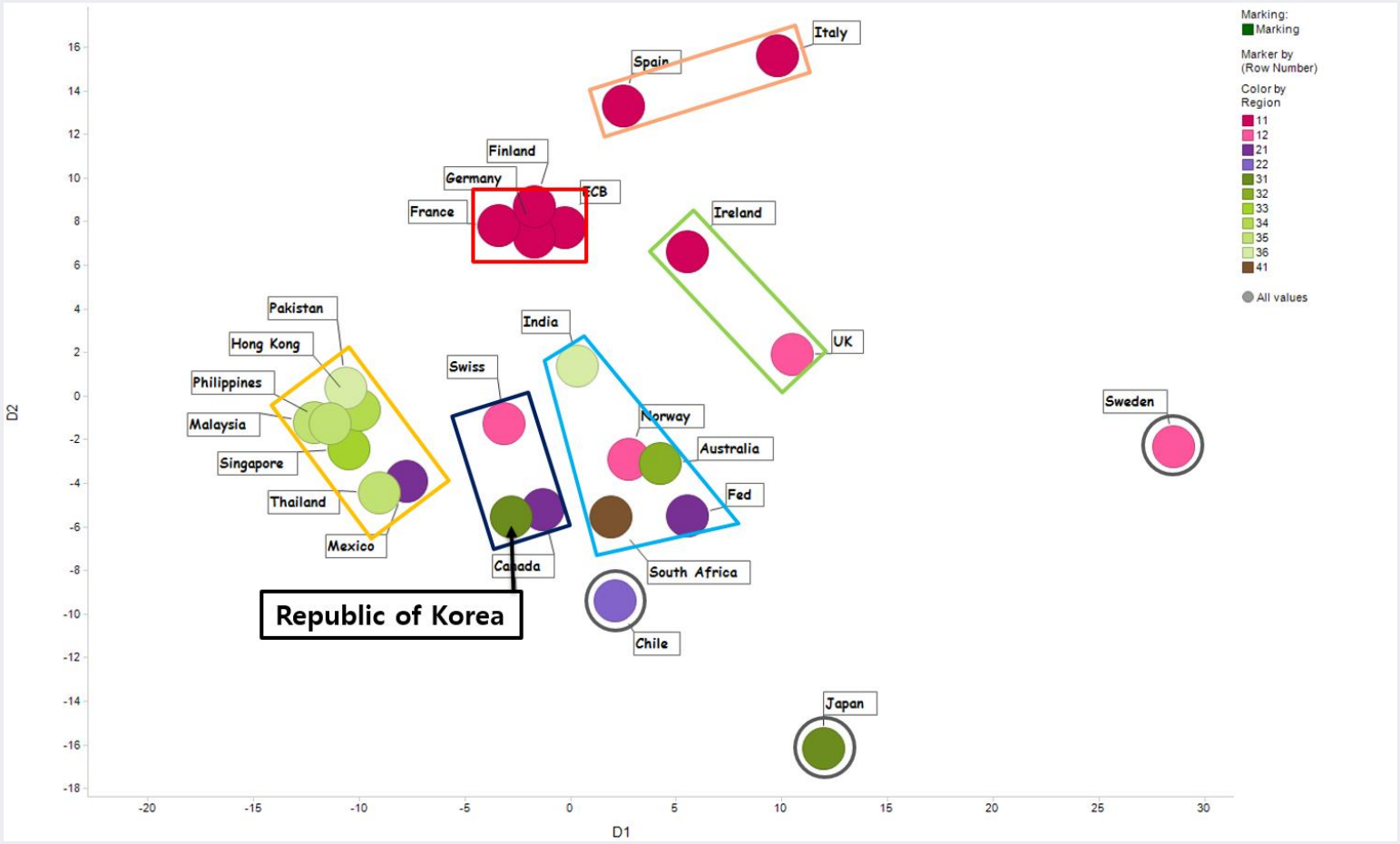
Text Mining: Applications

- Information Abstraction/Summarization/Visualization



Text Mining: Applications

- Information Abstraction/Summarization/Visualization
  - ✓ Central Bank speech analysis: Similarities between the central banks in the world



Text Mining: Applications

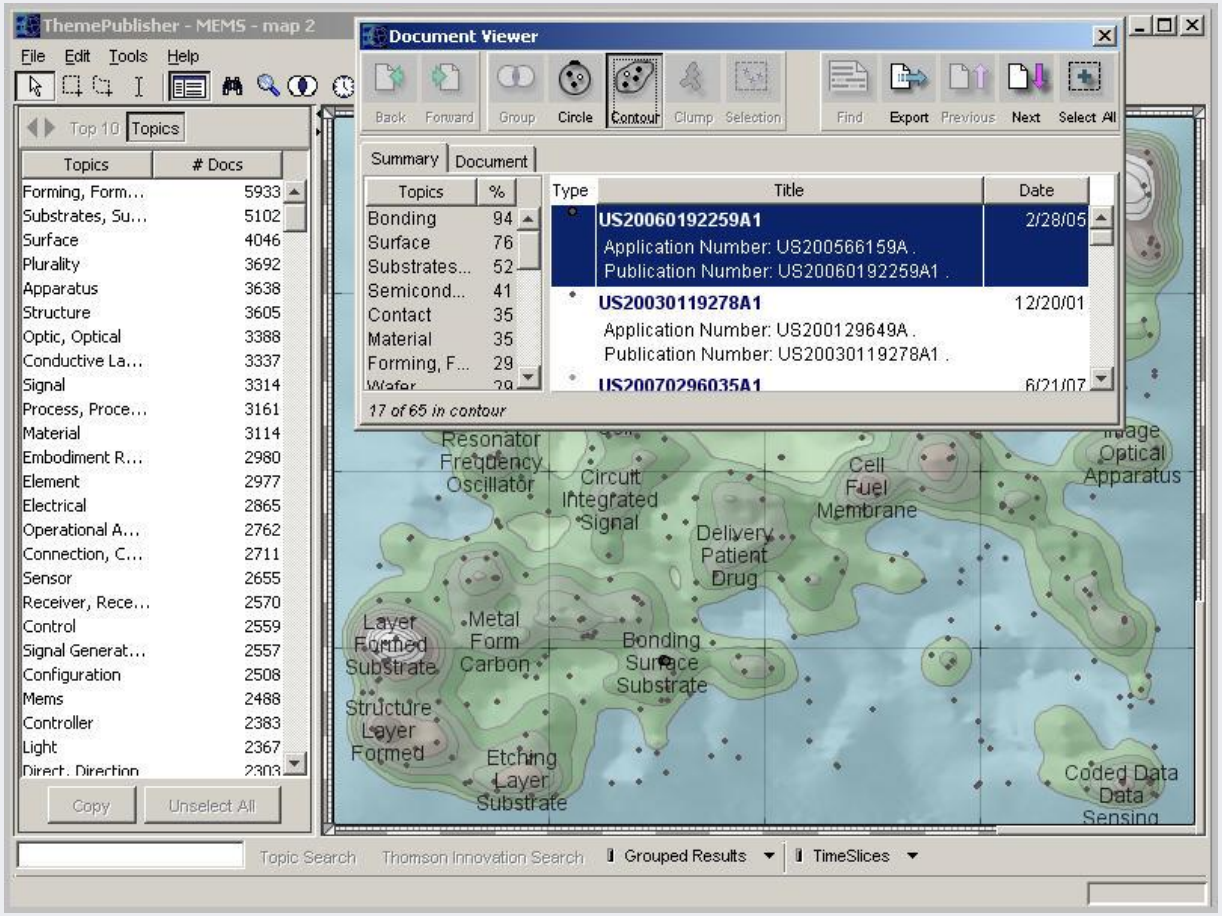
- Information Abstraction/Summarization/Visualization

✓ Central Bank speech analysis: Similarities between the central banks in the world

Year	Common Concern	Federal Reserve System	European Central Bank	Deutsche Bundesbank	Bank of England	Bank of Japan
2004	Sustainability Credibility	Expansion Imports	Parliaments Cooperation	Retirement Ages Working Hours	Household-Spending	QE Deflation
2005	China Inflation	Deficits Competitive	Financial Integration	Global Imbalance	Households Future Inflation	QE Recession
2006	Competitive Global Imbalance	Incentives Risk Taking	Administered Price Indirect Taxes	Inflation	China / India Commodities	Domestic and-External Demand
2007	Subprime-Mortgage	Subprime-Mortgage	Price Stability Turmoil	Banking Supervision Disclosure	Credit	Subprime-Mortgage
2008	Financial Turmoil Commodity Prices	Financial Turmoil Funding Markets	Financial Turmoil Liquidity	Financial Turmoil Subprime	Commodity Prices Housing Market	Securitized Product
2009	Financial Crisis Lehman Brothers	Financial Crisis ABS	Non-standard-Measure	Financial Crisis Rescue	Asset Purchase Recovery	Credit Bubble Financial Crisis
2010	Recovery Reform	Recovery Recession	ESRB/ FSB Deficits	Microprudential Macroprudential	Recovery/ QE VAT/ TAX	Deflation
2011	Sovereign Debt Basel III	Dodd-Franc Act Recovery	Sovereign Debt EFSF	Debt Crisis Basel III	Commodity Prices Basel III	Asset Purchase ETFs / REITs
2012	Europe Deleveraging	Recovery (has been) Labor Market	OMT/ ESM/ SSM Fragmentation	Banking Union Taxpayer	Investment-Banking	European Debt Deleveraging
2013	Real Economy Price Stability	(At least as long as) Unemployment	SRM/ SSM/ OMT	SRM / SSM Banking Union	Prudential-Regulation	Price Stability QE

Text Mining: Applications

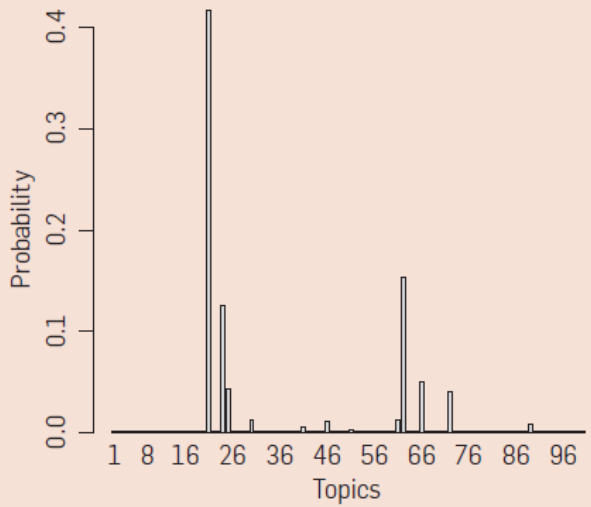
- Document Clustering
  - ✓ Cluster documents and extract representative keywords for each cluster



Text Mining: Applications

- Topic Extraction
  - ✓ Analyze documents and extract latent topics in the corpus

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Text Mining: Applications

Kim et al. (2016)

- Topic Extraction
  - ✓ 30 Topics discovered by LDA

Fault detection with DBN	Convolutional neural network	Network Learning	Representation learning	Face Recognition	Speech Recognition	Acoustic Modeling	Extreme Learning	Deep learning architecture	Image Segmentation
deep	neural	layer	feature	face	speaker	speech	deep	deep	image
belief	convolutional	input	level	recognition	speech	recognition	learn	architecture	scene
network	pool	output	extract	estimation	noise	acoustic	algorithm	neural	scale
dbn	convolution	unit	learn	facial	adaptation	hmm	structure	standard	segmentation
fault	convnet	hide	extraction	shape	source	neural	extreme	explore	pixel
		function							

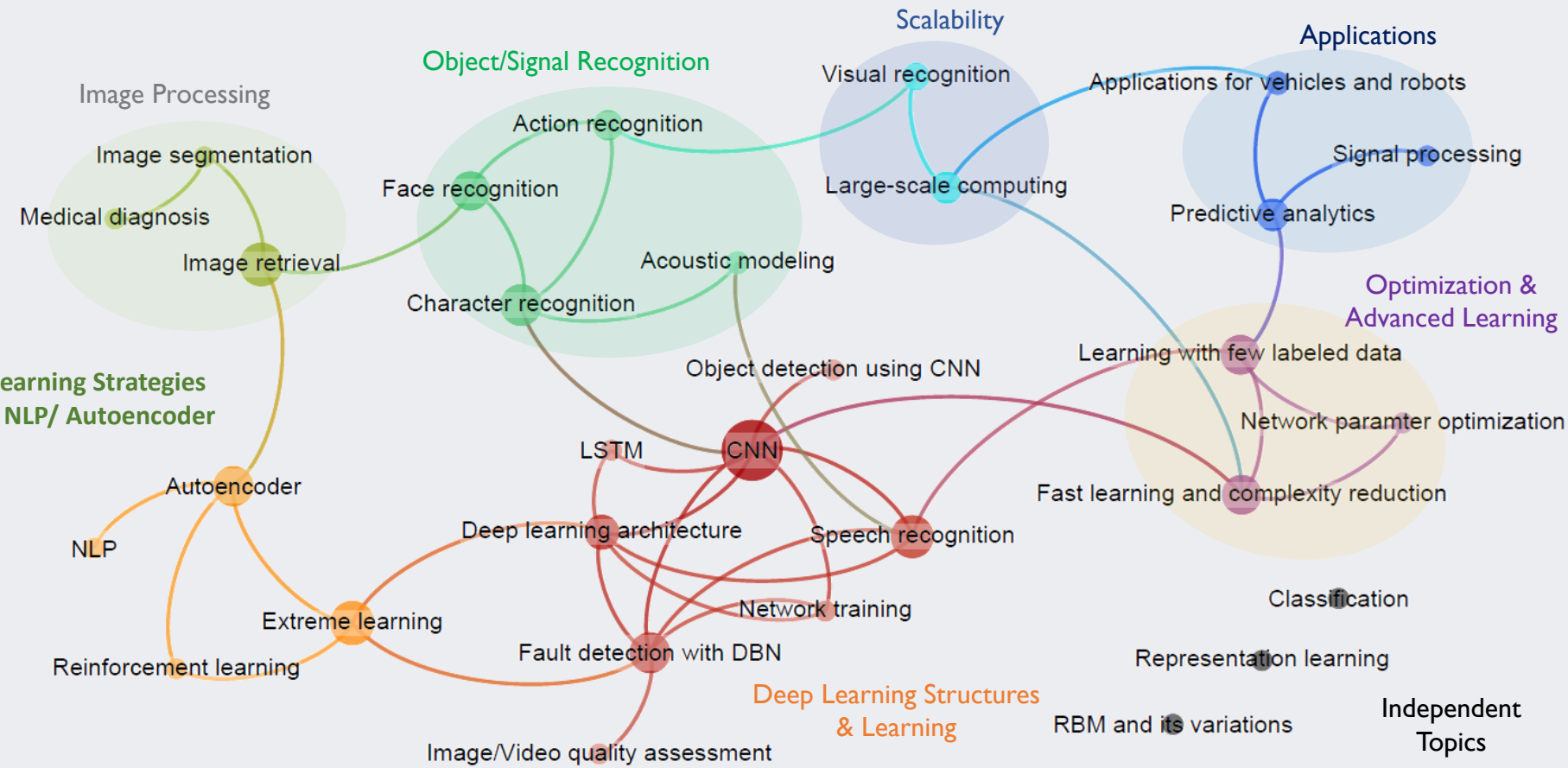
Long-short term memory	Predictive analytics	Signal processing	Classification models	Large-scale computing	Image quality assessment	Visual recognition	NLP	Detection using CNN	Action recognition
term	data	analysis	classification	application	domain	pattern	word	cnn	video
recurrent	prediction	filter	classifier	implementation	state	process	text	detection	human
long	technique	signal	class	efficient	quality	compute	language	convolutional	temporal
lstm	information	component	vector	process	resolution	visual	representation	neural	action
network	research	audio	support	power	relationship	field	semantic	detect	track

Image retrieval	Medical image di agnosis	Reinforcement learning	Parameter optimization	Auto encoder	RBM and variations	Learning with few labeled data	Fast learning complexity reduction	Applications for vehicles & robots	Character recognition
image	image	learn	train	representation	machine	train	fast	time	recognition
visual	segmentation	question	algorithm	learn	boltzmann	data	reduce	real	system
retrieval	disease	state	gradient	sparse	rbm	label	parameter	application	character
descriptor	cell	answer	sample	encode	restrict	few	weight	drive	network
attribute	medical	reinforcement	optimization	stack	distribution	transfer	complexity	Vehicle	neural

Text Mining: Applications

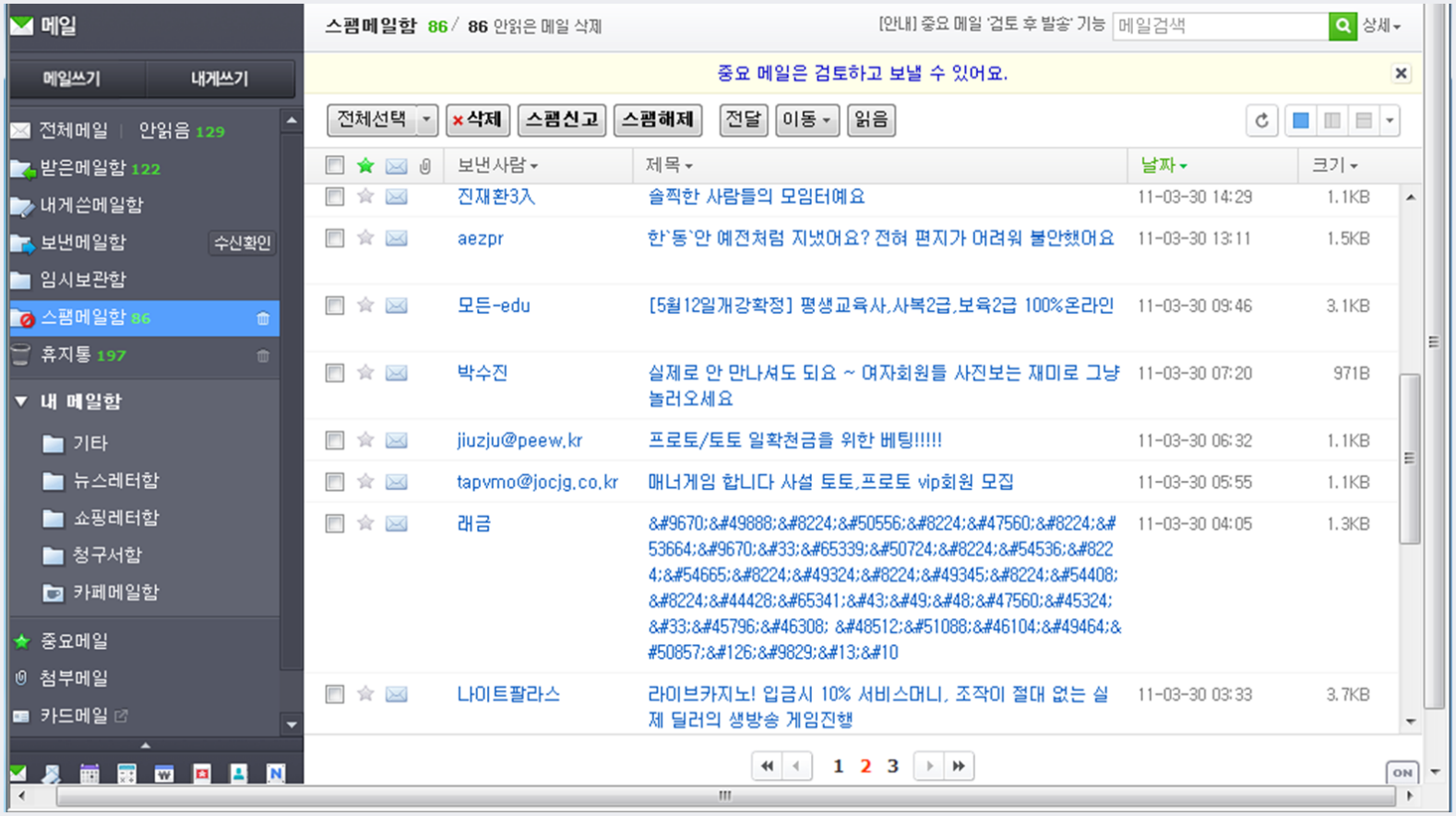
Kim et al. (2016)

- Topic Extraction
  - ✓ Relations between topics



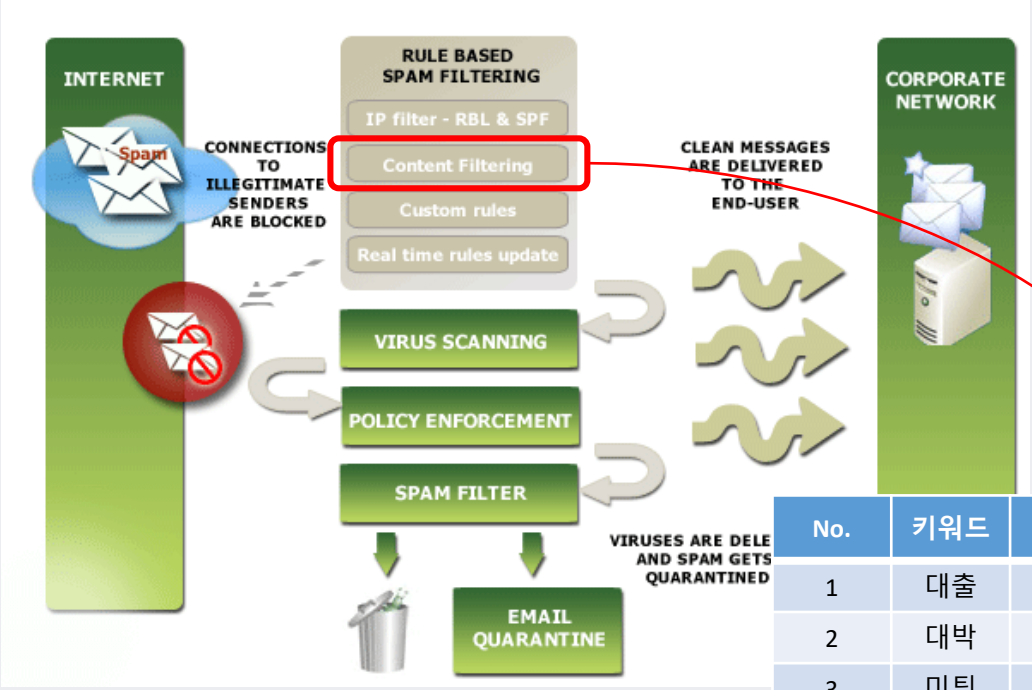
Text Mining: Applications

- Document Categorization/Classification
  - ✓ Spam mail filtering



Text Mining: Applications

- Document Categorization/Classification
  - ✓ Spam mail filtering

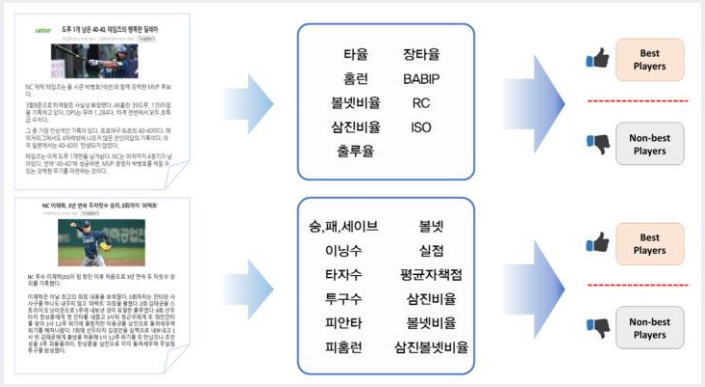


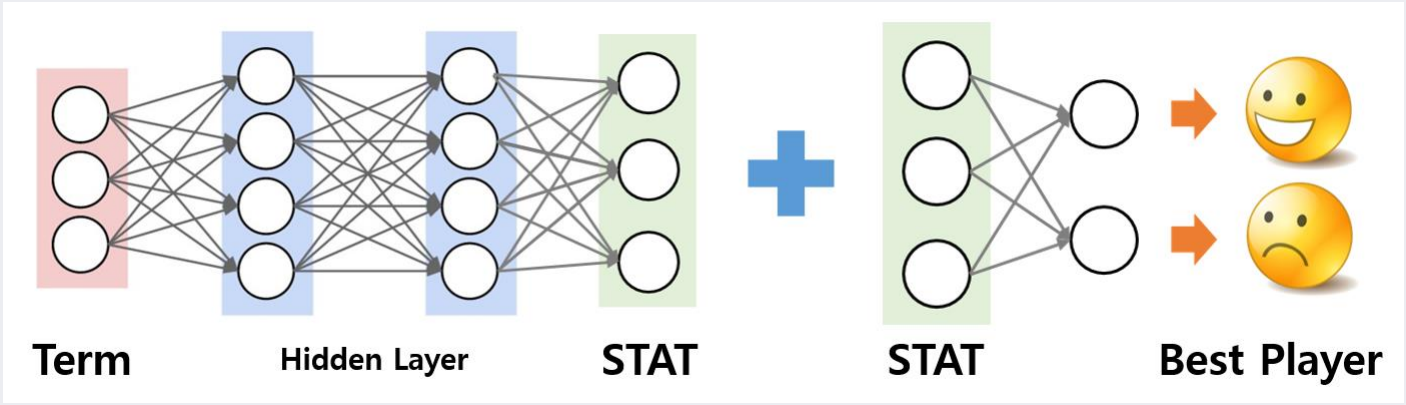
No.	키워드	Mail 1	Mail 2	Mail 3	...	Mail N
1	대출	0	2	0	...	0
2	대박	0	0	0	...	0
3	미팅	0	0	2	...	0
4	이상형	0	0	2	...	0
5	머니	0	2	0	...	0
6	외로	0	0	3	...	1
스팸 여부		N	Y	Y	...	N

Text Mining: Applications

Kim et al. (2016)

- Document Categorization/Classification
  - ✓ Sport player evaluation





Text Mining: Applications

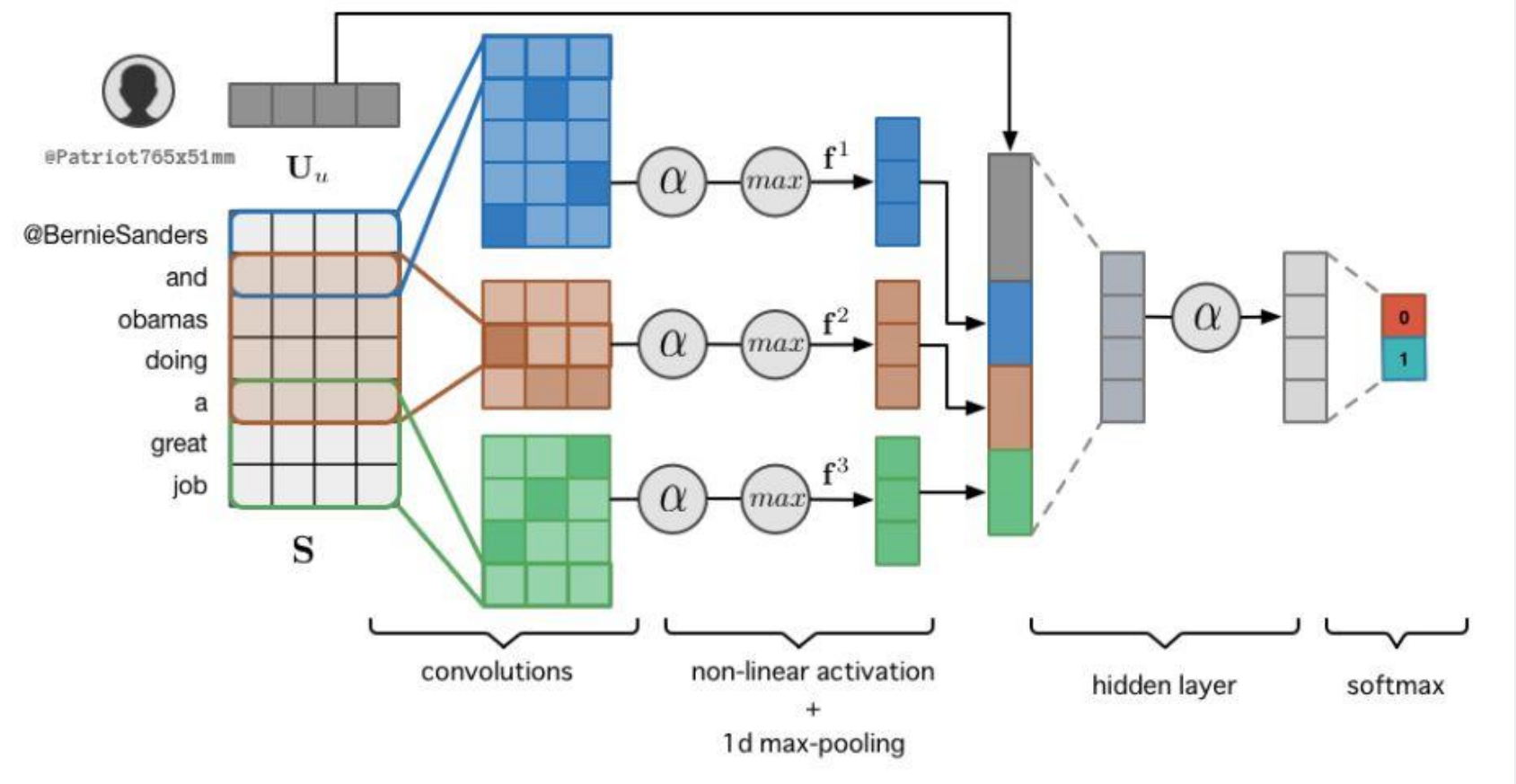
Lee et al. (2017)

- Document Categorization/Classification
  - ✓ Sentiment Analysis

Player	Line	Score	승	패배	피안타	피홈런	볼넷	실점	평균 자책점	이닝당 삼진	이닝당 볼넷	삼진 /볼넷	방어율
웨버	NC웨버는 7이닝 동안 1안타 3볼넷 10탈삼진으로 무실점 호투했다.	0.991	1	0	0	0	1	0	0	1	1	1	0
이재학	이재학 '10승', 3년만에 달라진 위상 증명하다	0.986	1	0	0	1	0	1	0	0	0	1	0
이태양	'이태양 완벽투' NC, LG 상대로 창단 첫 스윕	0.966	1	0	0	0	0	0	0	0	0	1	0
에릭	한편, 이날 7이닝 3실점으로 잘 던지며 7경기 만에 국내 첫 승을 거둔 에릭은 "굉장히 흥분된다.	0.803	1	0	1	0	0	0	0	0	0	0	0
찰리	어이없는 실책 2개가 나오자 찰리는 흔들리기 시작했다.	0.506	0	1	1	1	0	1	0	1	0	1	1
찰리	찰리가 대량 실점을 했지만 그 과정에서 3루수 모창민과 1루수 테임즈의 실책이 동반되면서 자책은 1점 밖에 되지 않았다.	0.506	0	1	1	1	0	1	0	1	0	1	1
원종현	NC 원종현이 패전투수가 됐다.	0.293	0	1	0	0	0	1	1	1	0	1	1
원종현	세 번째 투수로 등판한 원종현 역시 1사 후 8번 김성현에게 솔로포를 맞았다.	0.102	0	0	0	1	1	0	1	1	1	0	0
이재학	NC 선발 이재학은 8이닝 8피안타(2홈런) 5탈삼진 3볼넷 2실점을 기록했다.	0.079	0	0	1	1	0	1	1	0	0	0	1
이혜천	이혜천이 올라온 뒤 한꺼번에 5점을 내주면서 맥빠진 경기가 되고 말았다.	0.066	0	0	1	1	0	1	0	0	1	0	0

Text Mining: Applications

- Document Categorization/Classification
  - ✓ Sentiment Analysis

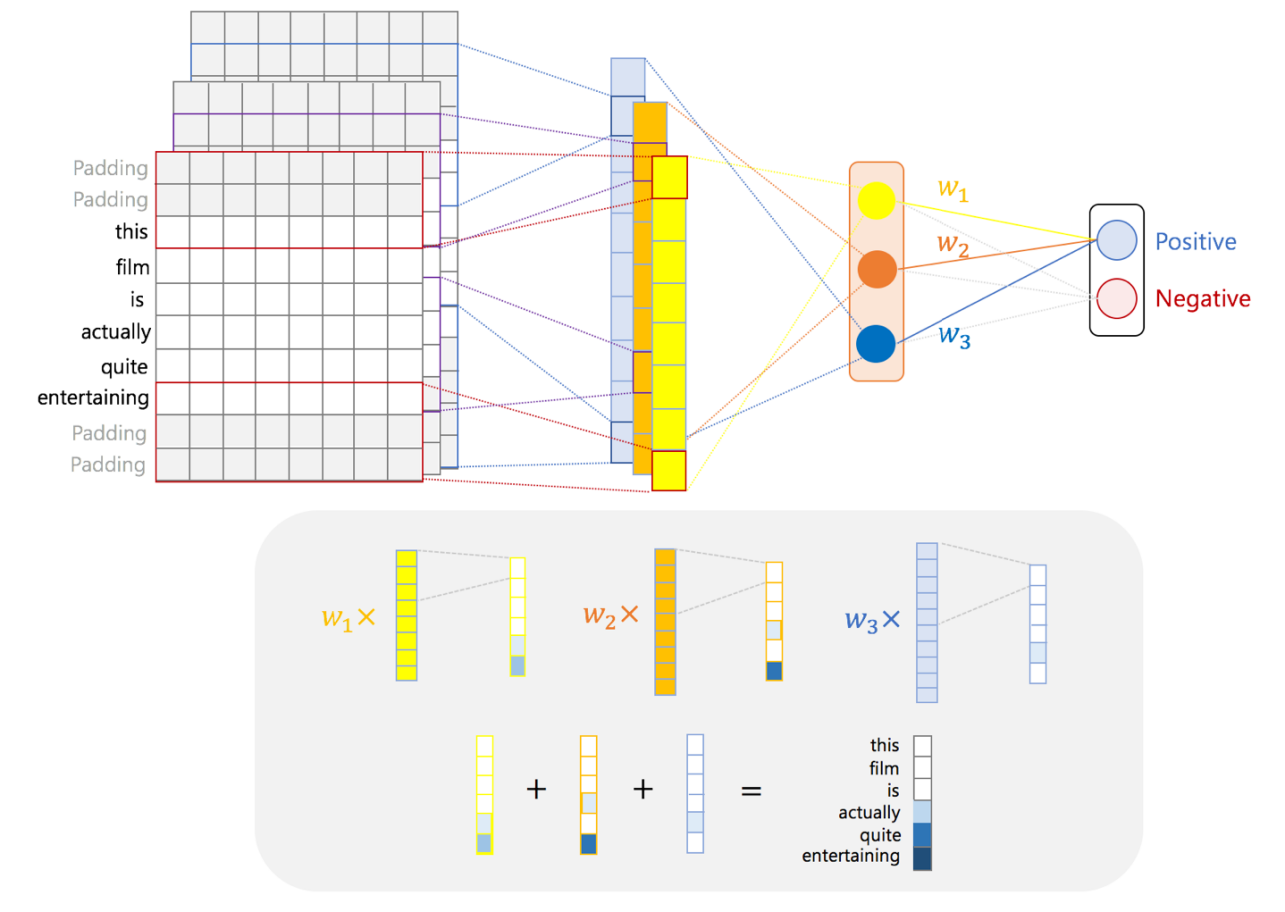


https://techxplore.com/news/2016-08-deep-neural-network-approach-sarcasm.html

Text Mining: Applications

Lee et al. (2017)

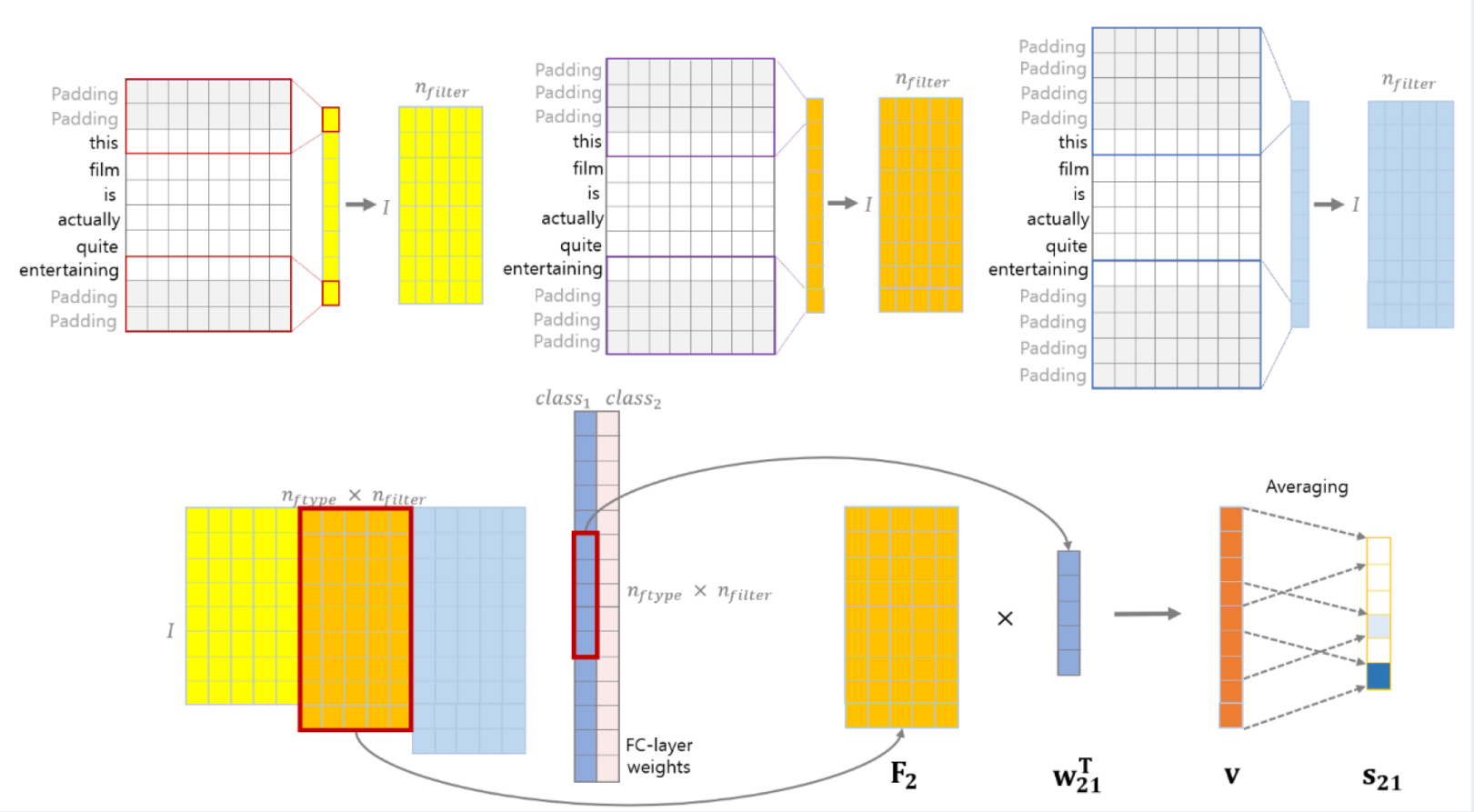
- Document Categorization/Classification
  - ✓ Sentiment Analysis



Text Mining: Applications

Lee et al. (2017)

- Document Categorization/Classification
  - ✓ Sentiment Analysis



Text Mining: Applications

Lee et al. (2017)

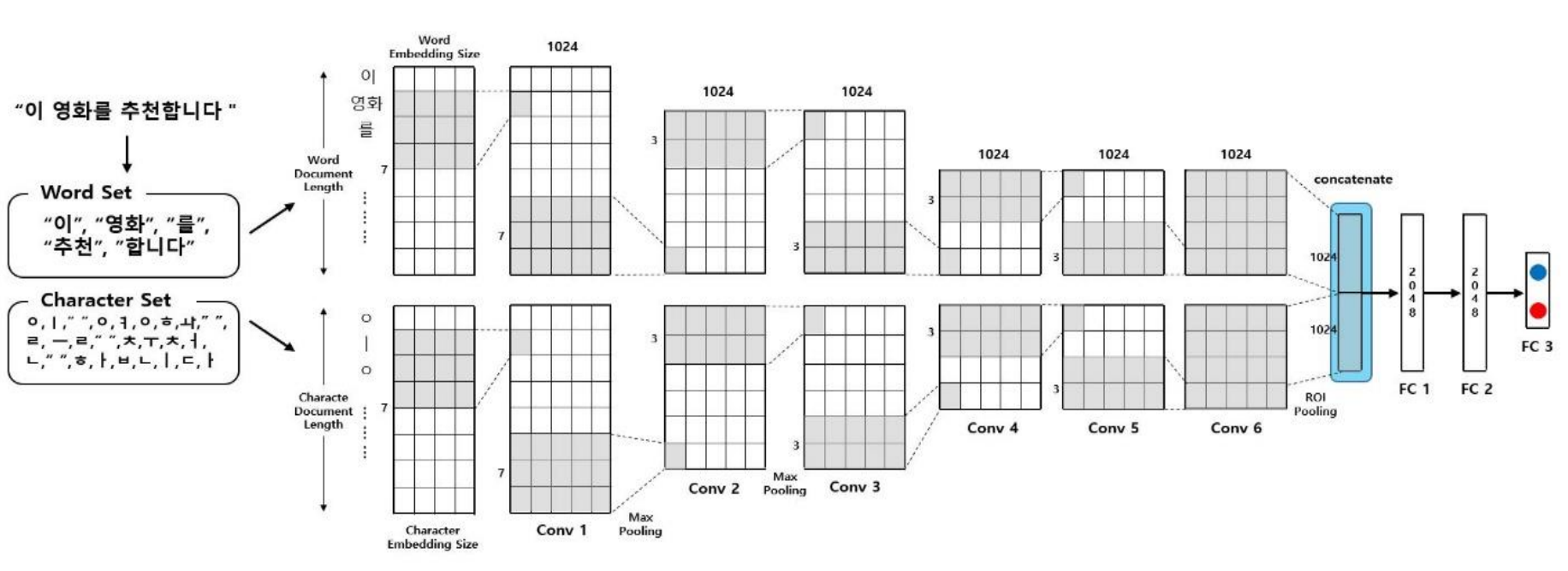
- Document Categorization/Classification
  - ✓ Sentiment Analysis

Methodology	Sentence
Raw text	I hate this movie. It is a horrid movie. Sean Young's character is completely unsympathetic. Her performance is wooden at best. The storyline is completely predictable and completely uninteresting. I would never recommend this film to anyone. It is one of the worst movies I have ever had the misfortune to see. *(1 / 10 points)*
CAM²-Rand	I hate this movie It is a horrid movie Sean Young s character is completely unsympathetic He r performance is wooden atbest The storyline is completely **predictable and completely uni nteresting** I would never recommend this film to anyone It is one of the **worst** movies I have ever had the misfortune to see **Negative**
CAM²-Static	I hate this movie It is a horrid movie Sean Young s character is completely unsympathetic He r performance is wooden atbest The storyline is **completely predictable and completely uni nteresting** I would never recommend this film to anyone It is one of the worst movies I have ever had the misfortune to see **Negative**
CAM²-Non-Static	I hate this movie It is a horrid movie Sean Young s character is completely unsympathetic He r performance is wooden atbest The storyline is completely **predictable and completely uni nteresting** I would never recommend this film to anyone It is one of the **worst** movies I have ever had the misfortune to see **Negative**
CAM²-2channel	I hate this movie It is a horrid movie Sean Young s character is completely unsympathetic He r performance is woodenat best The storyline is completely **predictable and completely uni nteresting** I would never recommend this film to anyone It is one of the **worst** movies I have ever had the misfortune to see **Negative**
CAM²-4channel	I hate this movie It is a horrid movie Sean Young s character is completely unsympathetic He r performance is woodenat best The storyline is completely **predictable and completely uni nteresting** I would never recommend this film to anyone It is one of the **worst** movies I have ever had the misfortune to see **Negative**
HAN	I **hate** this movie It is a **horrid** movie Sean Youngs character **is** completely unsympathetic He rperformance **is** wooden at best The storyline is completely predictable and completely **unin teresting** I would never recommend **this** film to anyone It is one of the **worst** movies I have e ver had the misfortune to see **Negative**

Text Mining: Applications

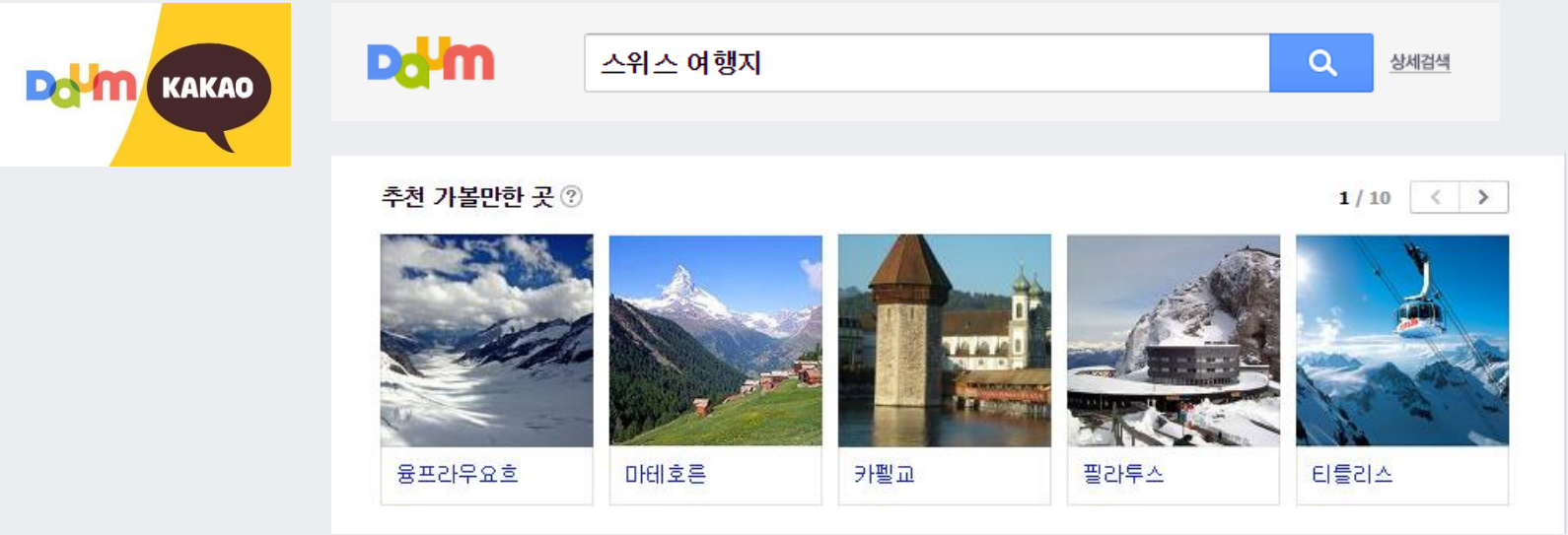
Mo et al. (2017)

- Document Categorization/Classification
  - ✓ Sentiment Analysis



Text Mining: Applications

- Recommendation
  - ✓ Analyze texts in daum café, blogs, and SNS contents
  - ✓ Named entity recognition/extraction (NEE/NER) technique in natural language processing is used
  - ✓ For 60,000 keywords



Text Mining: Applications

- Recommendation
  - ✓ Dining code: restaurant recommendation service
    - Analyze restaurant review from top 3 blog services (naver, daum, tistory)
    - Assign higher weights to opinion leaders' posts
    - Filter advertising blog posts by analyzing the comments on a post

Developed by HS Shin, KKU http://www.diningcode.com/

Text Mining: Applications

Kim et al. (2015)

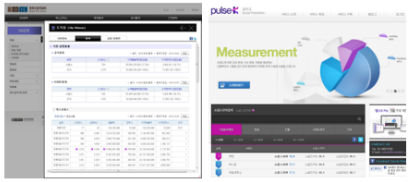
- Improve forecasting accuracy combined with structured data
  - ✓ Forecasting the box office scores based on the polarity of SNS posts

1 Movie Selection & Data Collection

Filtering rules
- Not too general title (ex: Mother)
- Total number of audiences ≥ 100,000

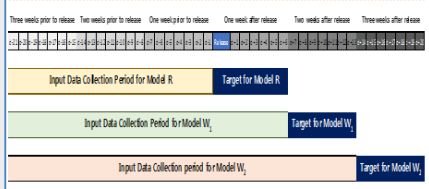
Screening data
- KOFIC
- www.kobis.or.kr

SNS data
- pulseK
- www.pulsek.com



2 Model Configuration & Variable Definition

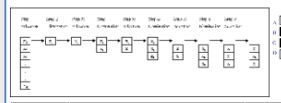
Model configuration
- Three forecasting models at different forecasting time and data collection periods



Variable definition
- Model **R** : 25 variables (1 screening, 24 SNS)
- Model $\mathbf{W_1}$: 32 variables (4 screening, 28 SNS)
- Model $\mathbf{W_2}$: 38 variables (6 screening, 32 SNS)

3 Variable Selection

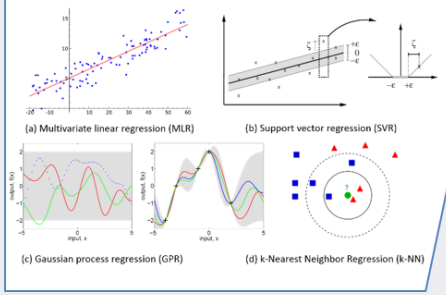
Variable selection
- Perform stepwise selection for forecasting models with screening data only
- Perform genetic algorithm for forecasting models with screening & SNS data



		w/o SNS	with SNS			
		Baseline	MLR	SVR	GPR	k-NN
Model R	Screening	1	1	1	1	1
	SNS		7	7	3	5
Model W_1	Screening	2	2	2	4	2
	SNS		2	6	0	8
Model W_2	Screening	4	3	6	5	6
	SNS		7	6	10	9

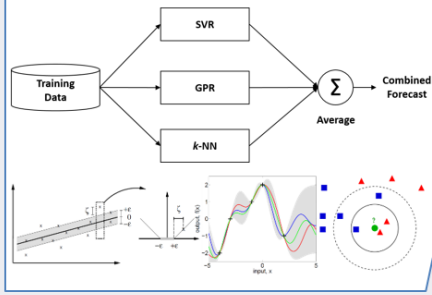
4 Single Forecasting Model Development & Evaluation

Single forecasting models
- 1 linear algorithm
- 3 machine learning-based algorithms



5 Combining Forecasts & Evaluation

Forecasting Combination
- Individual models: machine learning-based algorithms
- Combining rule: Equally weighted average



Text Mining: Applications

Kim et al. (2015)

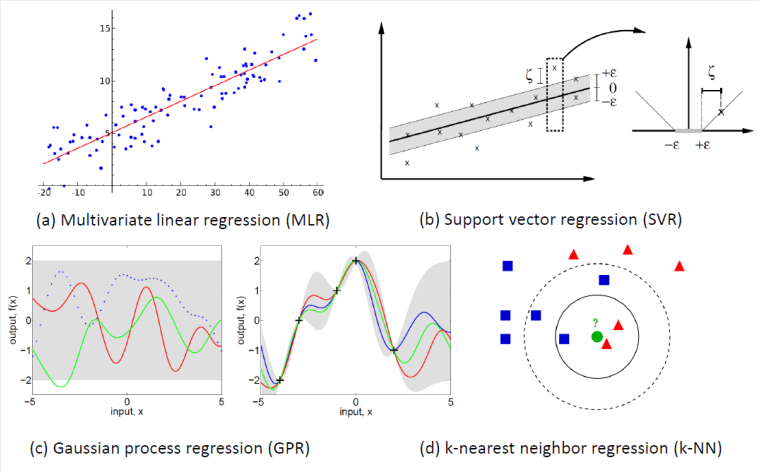
- Improve forecasting accuracy combined with structured data
  - ✓ Forecasting the box office scores based on the polarity of SNS posts

Category	Attribute	Model R	Model W_1	Model W_2
Target 1		$\text{Log}_{10}(\text{Box\_office}^1)$	$\text{Log}_{10}(\text{Box\_office}^2)$	$\text{Log}_{10}(\text{Box\_office}^3)$
Target 2		$\text{Log}_{10}(\text{Box\_office}^1)$	$\text{Log}_{10}(\sum_{i=1}^{2}\text{Box\_office}^i)$	$\text{Log}_{10}(\sum_{i=1}^{3}\text{Box\_office}^i)$
Screening	Original	N_seat_1	Box_office^1 N_seat_2 N_seat^1	$\text{Box\_office}^1+\text{Box\_office}^2$ N_seat_3 N_seat^2
	Derived		Weekly_seat_increase	$\text{N\_seat}^1+\text{N\_seat}^2$ Weekly_seat_increase Weekly_cumulative_seat_increase
SNS	Original	N_mention^p N_emotional^p N_positive^p N_negative^p $p \in \{-3, -2, -1\}$	N_mention^p N_emotional^p N_positive^p N_negative^p $p \in \{-3, -2, -1, 1\}$	N_mention^p N_emotional^p N_positive^p N_negative^p $p \in \{-3, -2, -1, 1, 2\}$
	Derived	Cumulative_mention Cumulative_emotional Cumulative_positive Cumulative_negative Avg_mention_increase Weekly_mention_increase Avg_emotional_increase Weekly_emotional_increase Avg_positive_increase Weekly_positive_increase Avg_negative_increase Weekly_negative_increase	Cumulative_mention Cumulative_emotional Cumulative_positive Cumulative_negative Avg_mention_increase Weekly_mention_increase Avg_emotional_increase Weekly_emotional_increase Avg_positive_increase Weekly_positive_increase Avg_negative_increase Weekly_negative_increase	Cumulative_mention Cumulative_emotional Cumulative_positive Cumulative_negative Avg_mention_increase Weekly_mention_increase Avg_emotional_increase Weekly_emotional_increase Avg_positive_increase Weekly_positive_increase Avg_negative_increase Weekly_negative_increase

Text Mining: Applications

Kim et al. (2015)

- Improve forecasting accuracy combined with structured data
  - ✓ Forecasting the box office scores based on the polarity of SNS posts



Model	Linear Algorithm		Machine Learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.8626	0.5958 (30.93%)	0.4278 (50.40%)* (28.19%)$^{+}$	0.5087 (41.02%)* (14.62%)$^{+}$	0.4605 (46.61%)* (22.71%)$^{++}$	0.4657 (46.01%) (21.84%)	0.4496 (47.88%)* (24.54%)$^{++}$ (3.45%)
Model W_1	0.1664	0.1535 (7.76%)	0.1251 (24.83%) (18.51%)	0.1206 (27.52%) (21.43%)	0.1288 (22.62%) (16.11%)	0.1248 (24.99%) (18.68%)	0.1216 (26.93%) (20.78%) (2.58%)
Model W_2	0.0624	0.0587 (5.90%)	0.0473 (24.16%)* (19.41%)$^{+}$	0.0541 (13.29%) (7.86%)	0.0486 (22.15%) (17.27%)	0.0500 (19.87%) (14.84%)	0.0453 (27.38%)* (22.83%)$^{+}$ (9.38%)

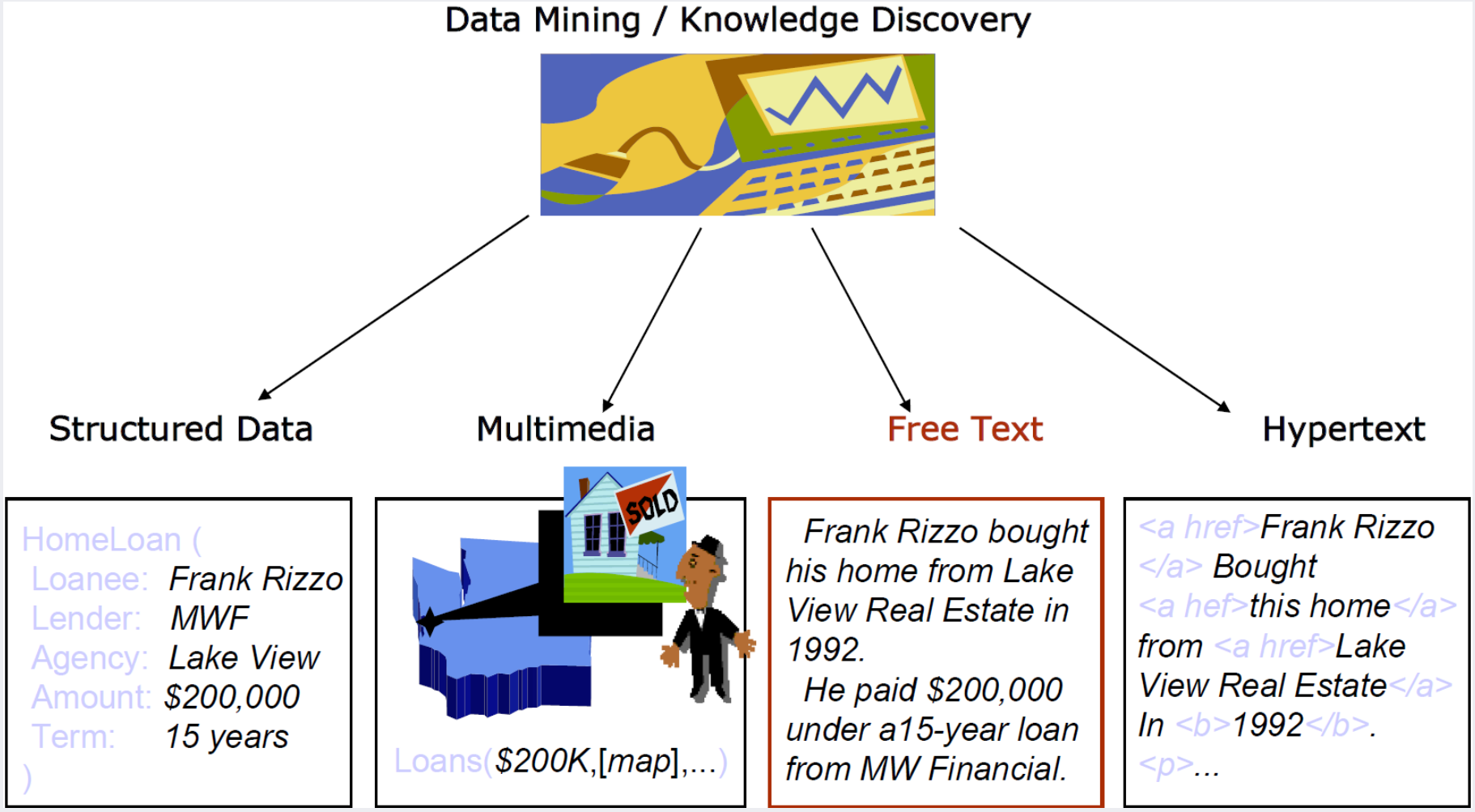
Mining Text Data

- Challenges
  - ✓ High number of possible "dimensions" (word, phrases, etc.)
  - ✓ Records (documents) are not structurally identical nor statistically independent
  - ✓ Complex and subtle relationship between concepts in texts
    - "AOL merges with Time-Warner"
    - "Time-Warner is bought by AOL"
  - ✓ Ambiguity and context sensitivity
    - automobile = car = vehicle = Hyundai

vs.

Mining Text Data

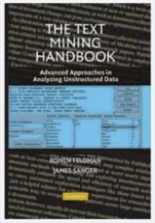
- Structure of text data



http://www.slideshare.net/pierluca.lanzi/machine-learning-and-data-mining-19-mining-text-and-web-data

Mining Text Data

Abbott (2013)

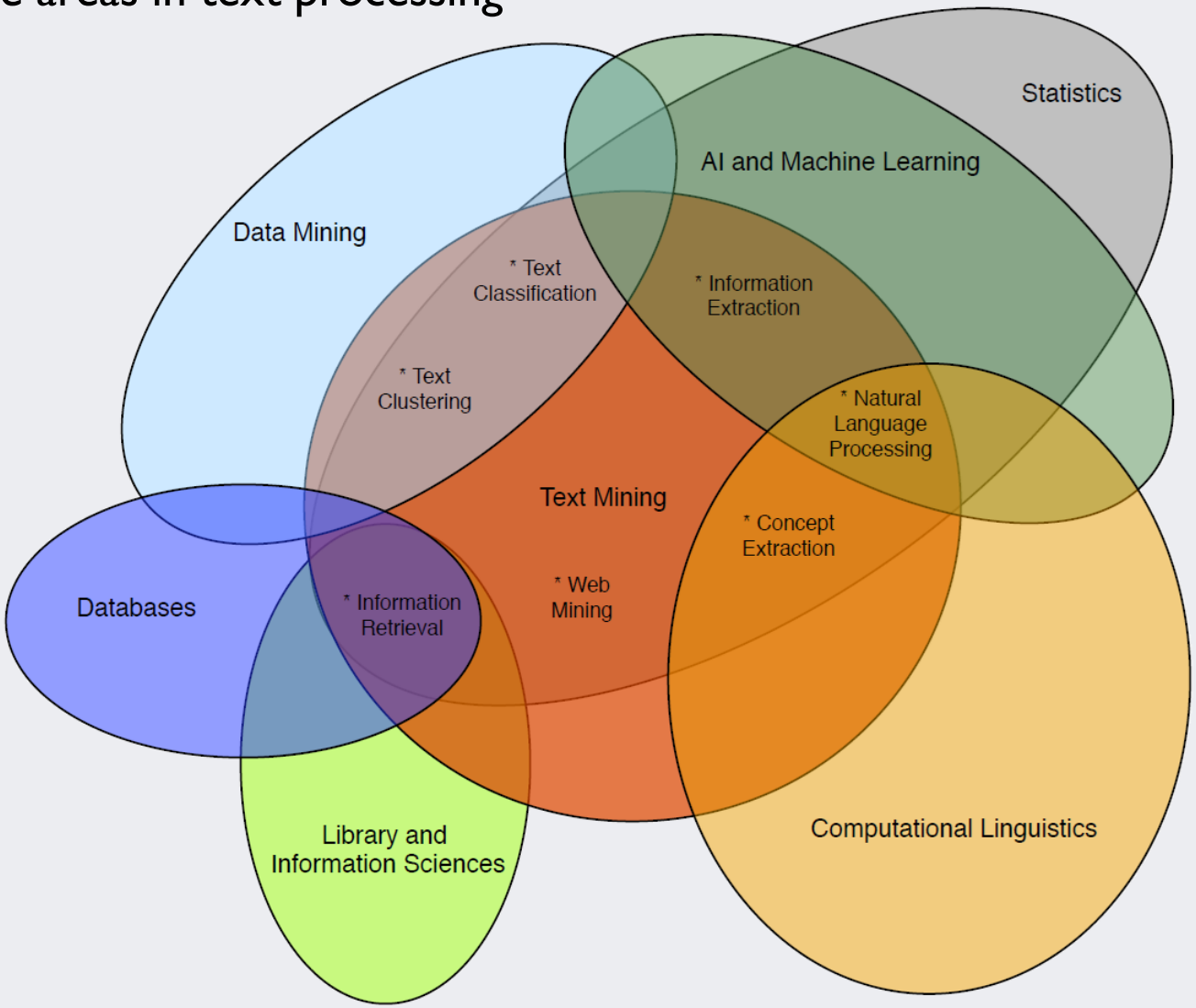


- How Unstructured is "Unstructured"? (by Feldman and Sanger)
  - ✓ Weakly structured
    - Few structural cues to text based layout or markups: research papers, legal memoranda, news stories, etc.
  - ✓ Semi-structured
    - Extensive format elements, metadata, field labels: E-mail, HTML/XML web pages, pdf files, etc.
- Why is Text Mining Hard?
  - ✓ Language itself is ambiguous
    - Contexts is needed to clarify
    - Same word with different meanings, different words with same meaning
    - Misspellings, abbreviations, etc.

Mining Text Data

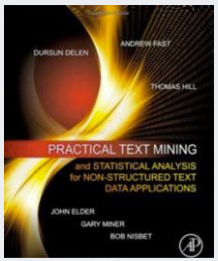
Abbott (2013)

- Active areas in text processing



Mining Text Data

Abbott (2013)



- Seven Types of Text Mining (by Elder et al.)
  - ✓ Document Classification
    - Grouping and categorizing snippets, paragraphs, or document using data mining classification methods, based on models trained on labeled examples
  - ✓ Document Clustering
    - Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods
  - ✓ Concept Extraction
    - Grouping or words and phrases into semantically similar groups

Mining Text Data

Abbott (2013)

- Seven Types of Text Mining (by Elder et al.)
  - ✓ Search and Information Retrieval (IR)
    - Storage and retrieval of text documents, including search engines and keyword search
  - ✓ Information Extraction (IE)
    - Identification and extraction of relevant facts and relationships from unstructured texts, the process of making structured data from unstructured and semi-structured texts
  - ✓ Web Mining
    - Data and text mining on the internet with a specific focus on the scale and interconnectedness of the web
  - ✓ Natural Language Processing (NLP)
    - Low-level language processing and understanding tasks (e.g., tagging part of speech)
    - Often used synonymously with computational linguistics

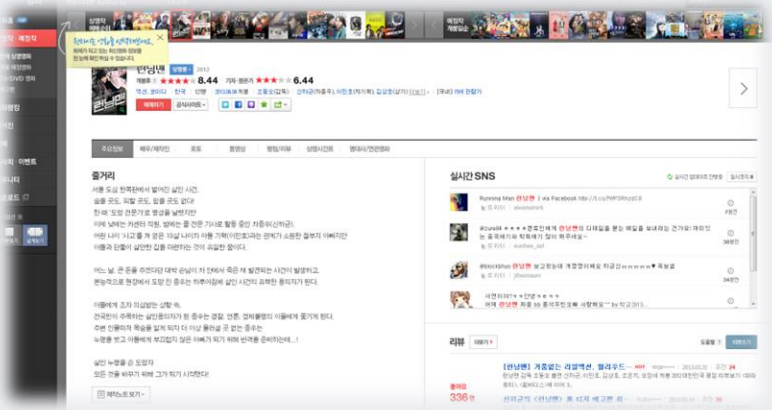
A Simplified Process of Text Mining

Source of text data

Digital library

Corporate document archive

Word Wide Web (WWW)



SNS

Step 1:
Decide what to mine
& Collect text data

A Simplified Process of Text Mining

From unstructured to structured!

S1: Jon likes to watch movies. Mary likes too.

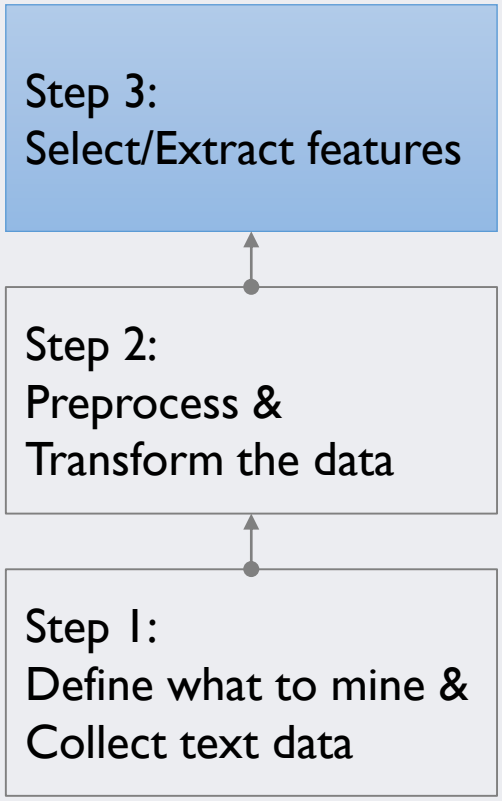
S2: John also likes to watch football game.

Word	S1	S2
John	1	1
Likes	2	1
To	1	1
Watch	1	1
Movies	1	0
Also	0	1
Football	0	1
Games	0	1
Mary	1	0
too	1	0

Step 2:
Preprocess &
Transform the data

Step 1:
Define what to mine &
Collect text data

A Simplified Process of Text Mining



Reduce the number of features

Word	S1	S2
John	1	1
Likes	2	1
To	1	1
Watch	1	1
Movies	1	0
Also	0	1
Football	0	1
Games	0	1
Mary	1	0
too	1	0

Word	S1	S2
Likes	2	1
Watch	1	1
Movies	1	0
Football	0	1
Games	0	1

A Simplified Process of Text Mining

Step 4:
Algorithm Learning & Evaluation

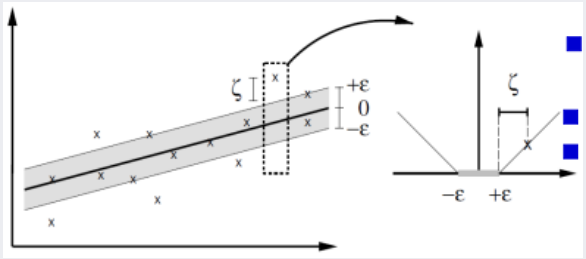
Step 3:
Select/Extract features

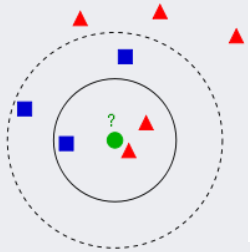
Step 2:
Preprocess & Transform the data

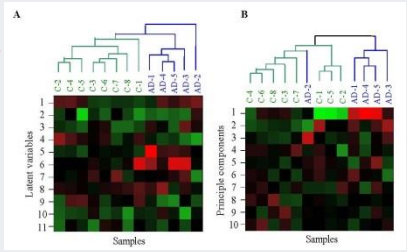
Step 1:
Define what to mine & Collect text data

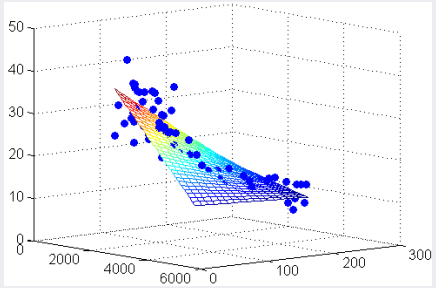
Select appropriate algorithm

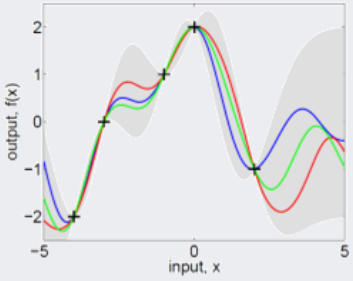
- Vector space model vs. Probabilistic model
- Classification vs. Clustering vs. Association

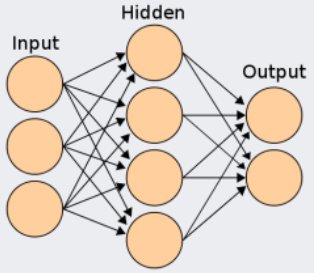












AGENDA

01 Text Mining Overview

02 TM Process 1: Collection & Preprocessing

03 TM Process 2: Transformation

04 TM Process 3: Dimensionality Reduction

05 TM Process 4: Learning & Evaluation

Decide What to Mine

Witte (2006)

- From a wide range of text sources…

Emails, Instant Messages, Blogs, . . .

Look for:

- Entities (Persons, Companies, Organizations, . . .)
- Events (Inventions, Offers, Attacks, . . .)

Biggest existing system: ECHELON (UKUSA)

News: Newspaper articles, Newswires, . . .

Similar to last, but additionally:

- collections of articles (e.g., from different agencies, describing the same event)
- contrastive summaries (e.g., event described by U.S. newspaper vs. Arabic newspaper)
- also needs temporal analysis
- main problems: cross-language and cross-document analysis

Many publicly accessible systems, e.g. *Google News* or *Newsblaster*.

Decide What to Mine

Witte (2006)

- From a wide range of text sources…

(Scientific) Books, Papers, . . .

- detect new trends in research
- automatic curation of research results in Bioinformatics

need to deal with highly specific language

Software Requirement Specifications, Documentation, . . .

- extract requirements from software specification
- detect conflicts between source code and its documentation

Web Mining

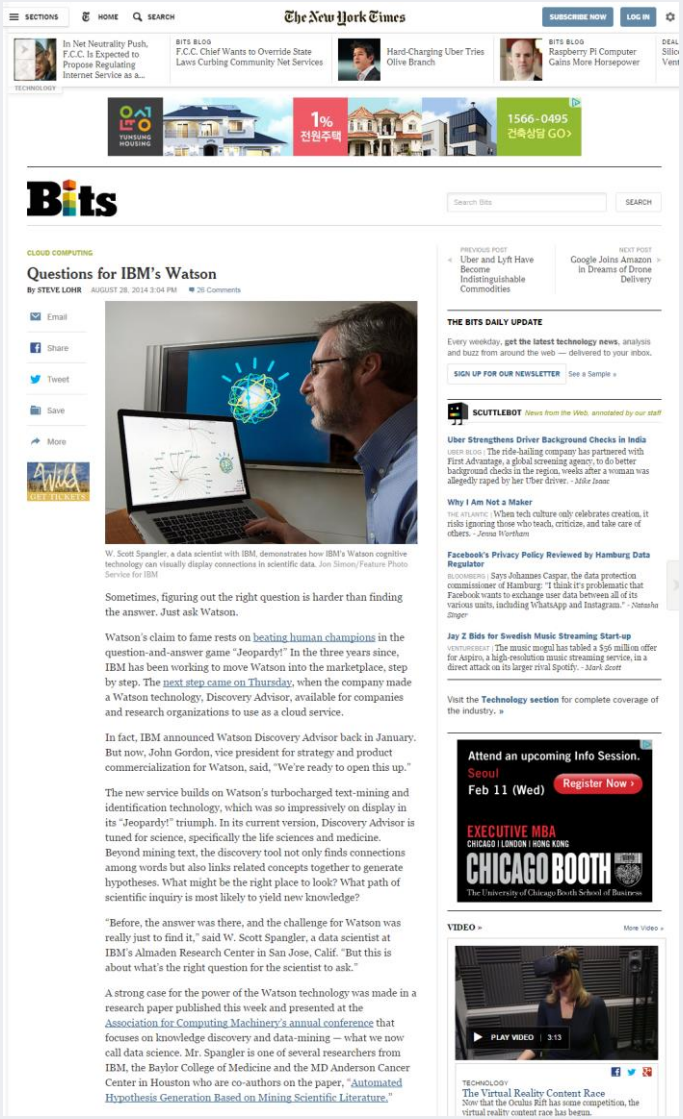
extract and analyse information from web sites

- mine companies' web pages (detect new products & trends)
- mine Intranets (gather knowledge, find "illegal" content, . . .)

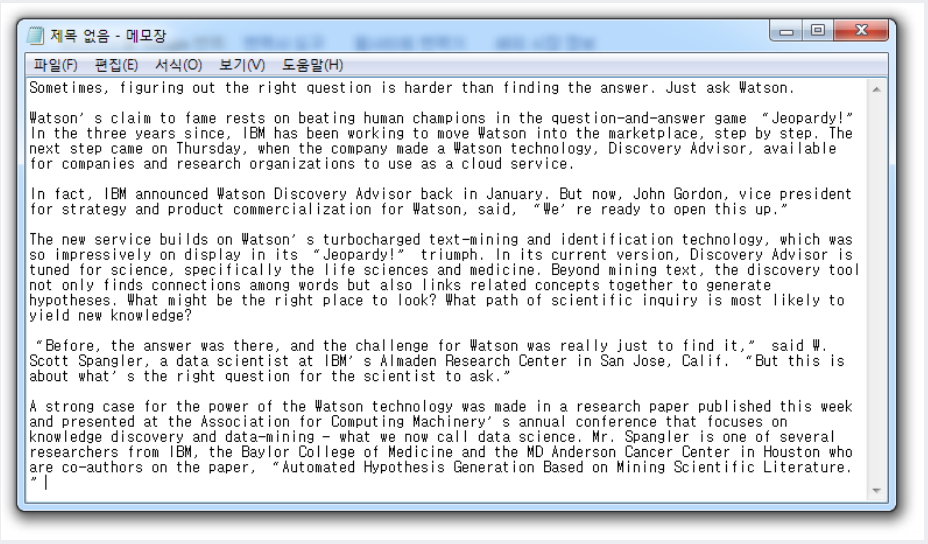
problems: not simply plain text, also hyperlinks and hidden information ("deep web")

Text Preprocessing Level 0: Text

- Remove unnecessary information from the collected data



- Remove figures, advertisements, html syntax, hyperlinks, etc.



Text Preprocessing Level 0: Text

- Do not remove meta-data, which contains significant information on the text
  - ✓ Ex) Newspaper article: author, date, category, language, newspaper, etc.
- Meta-data can be used for further analysis
  - ✓ Target class of a document
  - ✓ Time series analysis

Text Preprocessing Level 1: Token

- Extracting meaningful (worth being analyzed) tokens (word, number, space, etc.) from a text is not an easy task.
  - ✓ Is John's sick one token or two?
    - If one → problems in parsing (where is the verb?)
    - If two → what do we do with John's house?
  - ✓ What to do with hyphens?
    - database vs. data-base vs. data base
  - ✓ What to do with "C++", "A/C", ":-)", "…", "ㅋㅋㅋㅋㅋㅋㅋㅋㅋ"?
  - ✓ Some languages do not use whitespace (e.g., Chinese)

2013年5月，习主席在视察成都战区时，郑重提出在适当时候召开全军政治工作会议，并明确提出到古田召开这次会议，以更好弘扬我党我军的光荣传统和优良作风。6月，总政治部向中央军委提交《关于筹备召开全军政治工作会议的请示》，提出要通过召开会议形成一个指导性文件。习主席随即批示同意，明确要求这个文件要充分体现深厚的历史积淀和政治意蕴，能够管一个时期，起到历史性作用。

- Consistent tokenization is important for all later processing steps.

Text Preprocessing Level 1: Token

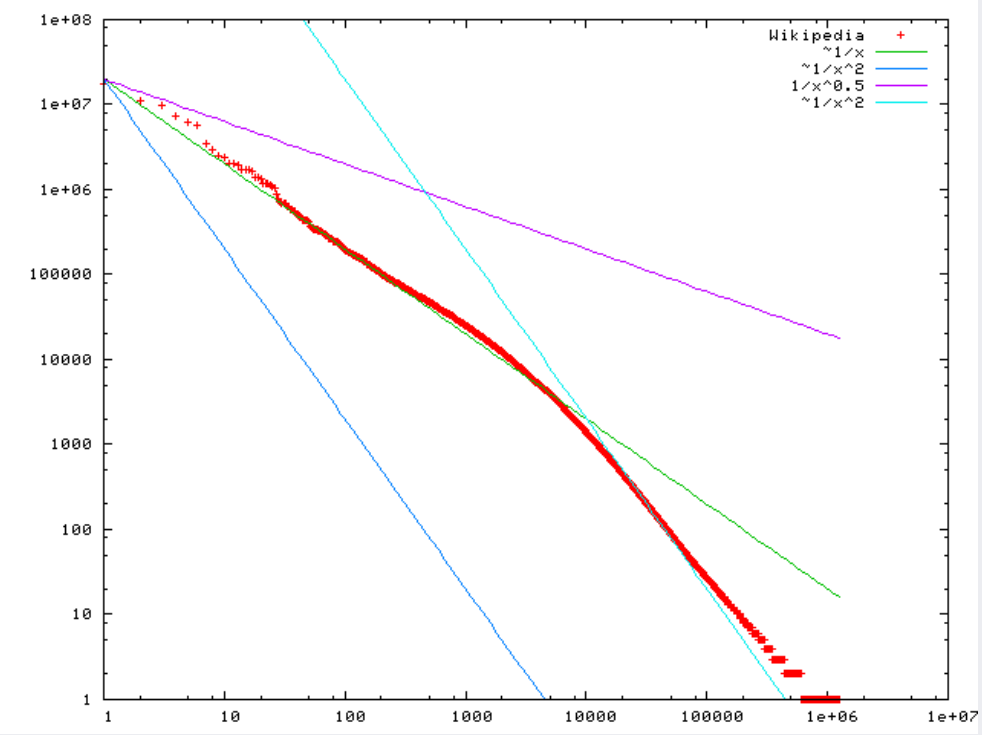
- Power distribution in word frequencies
  - ✓ It is not true that more frequently appeared words (tokens) are more important for text mining tasks.

100 common words in the Oxford English Corpus

Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use
4	of	24	by	44	if	64	your	84	two
5	and	25	from	45	about	65	good	85	how
6	a	26	they	46	who	66	some	86	our
7	in	27	we	47	get	67	could	87	work
8	that	28	say	48	which	68	them	88	first
9	have	29	her	49	go	69	see	89	well
10	I	30	she	50	me	70	other	90	way
11	it	31	or	51	when	71	than	91	even
12	for	32	an	52	make	72	then	92	new
13	not	33	will	53	can	73	now	93	want
14	on	34	my	54	like	74	look	94	because
15	with	35	one	55	time	75	only	95	any
16	he	36	all	56	no	76	come	96	these
17	as	37	would	57	just	77	its	97	give
18	you	38	there	58	him	78	over	98	day
19	do	39	their	59	know	79	think	99	most
20	at	40	what	60	take	80	also	100	us

http://en.wikipedia.org/wiki/Most_common_words_in_English

Word frequency distribution in Wikipedia



http://upload.wikimedia.org/wikipedia/commons/b/b9/Wikipedia-n-zipf.png

Text Preprocessing Level 1: Word

- Stop-words
  - ✓ Words that do not carry any information
    - Mainly functional role
    - Usually remove them to help the machine learning algorithms to perform better
  - ✓ Natural language dependent
    - English: a, about, above, across, after, again, against, all, also, etc.
    - 한국어: …습니다, …로서(써), …를 등

[Original text]

Information Systems Asia Web - provides research, IS-related commercial materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region.

[After removing stop words]

Information Systems Asia Web provides research IS-related commercial materials interaction research sponsorship interested corporations focus Asia Pacific region

http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf

Text Preprocessing Level 1: Word

- Stemming
  - ✓ Different forms of the same word are usually problematic for text data analysis, because they have different spelling and similar meaning
    - Learns, learned, learning, …
  - ✓ Stemming is a process of transforming a word into its stem (normalized form)
    - In English: Porter2 stemmer (http://snowball.tartarus.org/algorithms/english/stemmer.html)
    - In Korean, 꼬꼬마 형태소 분석기 (http://kkma.snu.ac.kr/documents/)

word	stem	word	stem
consign	consign	knack	knack
consigned	consign	knackeries	knackeri
consigning	consign	knacks	knack
consignment	consign	knag	knag
consist	consist	knave	knave
consisted	consist	knaves	knave
consistency	consist	knavish	knavish
consistent	consist	kneaded	knead
consistently	consist	kneading	knead
consisting	consist	knee	knee
consists	consist	kneel	kneel
consolation	consol	kneeled	kneel
consolations	consol	kneeling	kneel
consolatory	consolatori	kneels	kneel
console	consol	knees	knee
consoled	consol	knell	knell
consoles	consol	knelt	knelt
consolidate	consolid	knew	knew
consolidated	consolid	knick	knick
consolidating	consolid	knif	knif

Text Preprocessing Level 1: Word

- Lemmatization
  - ✓ Although stemming just finds any base form, which does not even need to be a word in the language, but lemmatization finds the actual root of a word

Word	Stemming	Lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

Text Preprocessing Level 2: Sentence

- Correct sentence boundary is also important
  - ✓ For many downstream analysis tasks
    - POS-Tagger maximize probabilities of tags within a sentence
    - Summarization systems rely on correct detection of sentence

Mark Sentence Boundaries

Detects sentence units. Easy case:

- often, sentences end with ".", "!", or "?"

Hard (or annoying) cases:

- difficult when a "." do not indicate an EOS:
  "MR. X", "3.14", "Y Corp.", ...
- we can detect common abbreviations ("U.S."), but what if a sentence ends with one?
  "...announced today by the U.S. The..
- Sentences can be *nested* (e.g., within quotes)

AGENDA

01 Text Mining Overview

02 TM Process 1: Collection & Preprocessing

03 TM Process 2: Transformation

04 TM Process 3: Dimensionality Reduction

05 TM Process 4: Learning & Evaluation

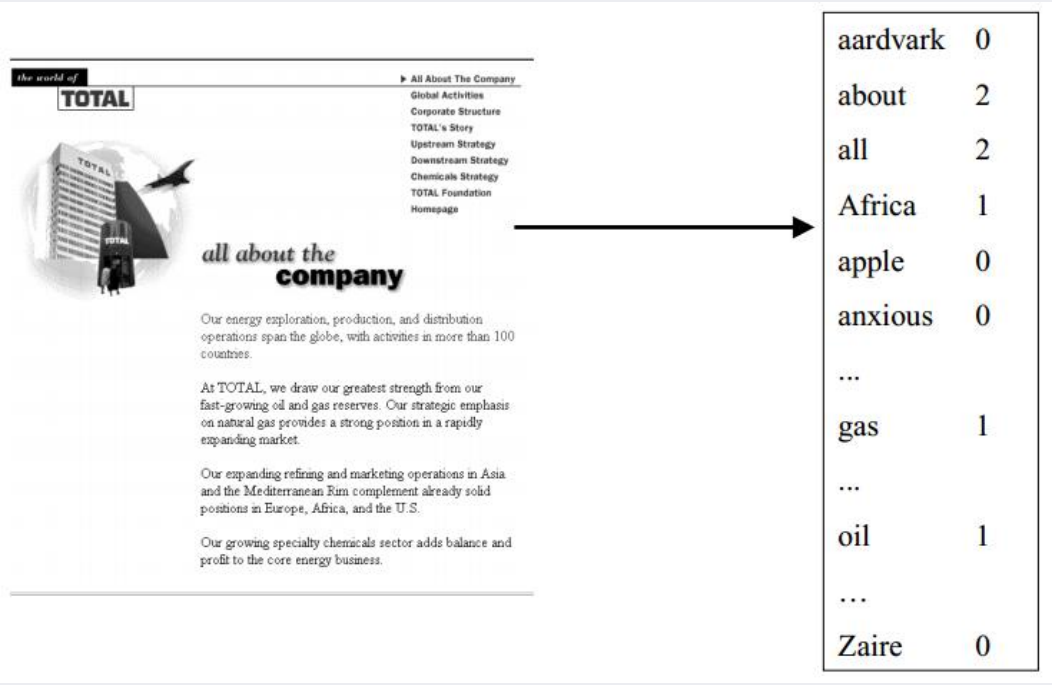
Text Transformation

- Document representation
  - ✓ Bag-of-words: simplifying representation method for documents where a text is represented in a vector of an unordered collection of words

S1: Jon likes to watch movies. Mary likes too.

S2: John also likes to watch football game.

Word	S1	S2
John	1	1
Likes	2	1
To	1	1
Watch	1	1
Movies	1	0
Also	0	1
Football	0	1
Games	0	1
Mary	1	0
too	1	0



Text Transformation

- Word Weighting
  - ✓ Each word is represented as a separate variable with a numeric weight
  - ✓ Term frequency and inverse document frequency (**TF-IDF**)
    - $tf(w)$: term frequency (number of word occurrences in a document)
    - $df(w)$: number of documents containing the word (number of documents containing the word)

$$TF - IDF(w) = tf(w) \times \log\left(\frac{N}{df(w)}\right)$$

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Text Transformation

- Word weighting (cont'): example

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0.35
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

Text Transformation

- One-hot-vector representation
  - ✓ The most simple & intuitive representation

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{a} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

  - ✓ Can make a vector representation, but similarities between words cannot be preserved.

$$(w^{hotel})^{\top} w^{motel} = (w^{hotel})^{\top} w^{cat} = 0$$

Text Transformation

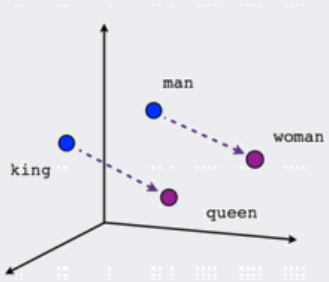
- Word vectors: distributed representation
  - ✓ A parameterized function mapping words in some language to a certain dimensional vectors

$$W : \text{words} \rightarrow \mathbb{R}^n$$

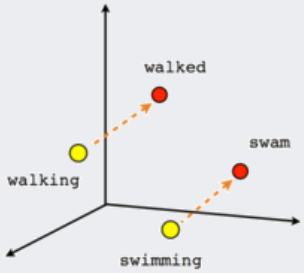
$$W(\text{“cat”}) = (0.2,\ \text{-}0.4,\ 0.7,\ \ldots)$$

$$W(\text{“mat”}) = (0.0,\ 0.6,\ \text{-}0.1,\ \ldots)$$

- Interesting feature of word embedding
  - ✓ Semantic relationship between words can be preserved



Male-Female

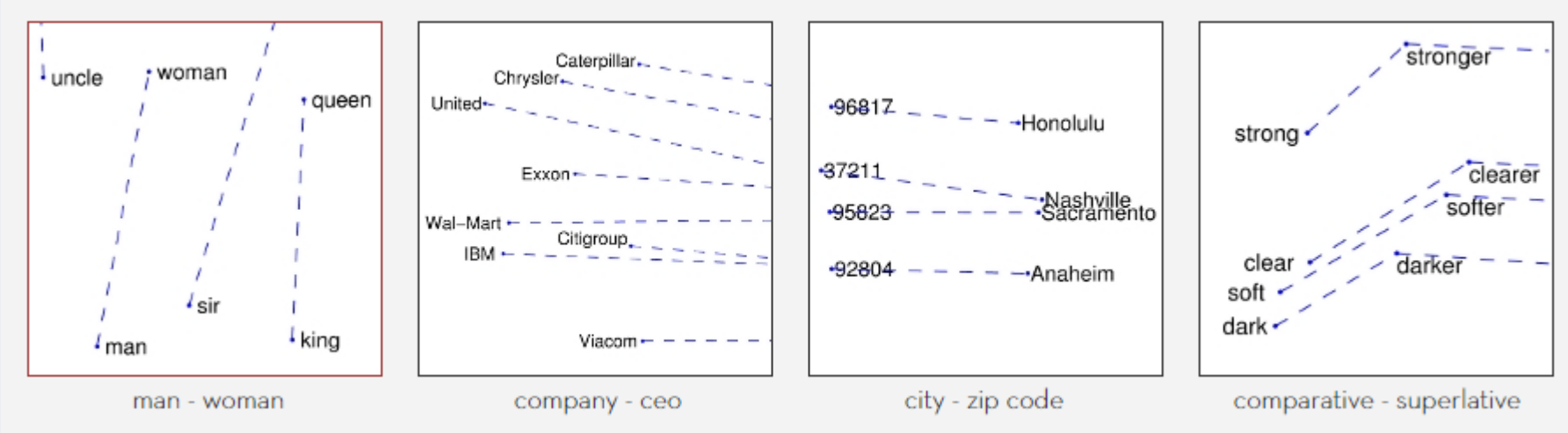


Verb tense

Country-Capital

Text Transformation

- Word vectors: distributed representation



http://nlp.stanford.edu/projects/glove/

AGENDA

01 Text Mining Overview

02 TM Process 1: Collection & Preprocessing

03 TM Process 2: Transformation

04 TM Process 3: Dimensionality Reduction

05 TM Process 4: Learning & Evaluation

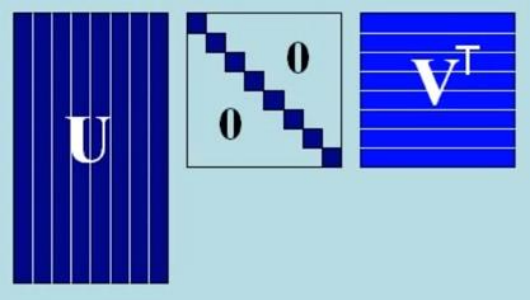
Feature Selection/Extraction

- Feature subset selection
  - ✓ Select only the best features for further analysis
    - The most frequent
    - The most informative relative to the all class values, …
  - ✓ Scoring methods for individual feature (for supervised learning tasks)
    - Information gain: $\sum_{F=W,\overline{W}} P(F) \sum_{C=pos,neg} P(C|F) \log \frac{P(C|F)}{P(C)}$
    - Cross-entropy: $P(W) \sum_{C=pos,neg} P(C|W) \log \frac{P(C|W)}{P(C)}$
    - Mutual information: $\sum_{C=pos,neg} P(C) \log \frac{P(W|C)}{P(W)}$
    - Weight of evidence: $\sum_{C=pos,neg} P(C)P(W) \left| \log \frac{P(C|W)(1-P(C))}{P(C)(1-P(C|W))} \right|$
    - Odds ratio: $\log \frac{P(W|pos) \times (1-P(W|neg))}{(1-P(W|pos)) \times P(W|neg)}$
    - Frequency: $Freq(W)$

Feature Selection/Extraction

- Feature subset extraction
  - ✓ Feature extraction: construct a set of variables that preserve the information of the original data by combining them in a linear/non-linear form
  - ✓ Latent Semantic Analysis (LSA) that is based on singular value decomposition is commonly used

m x r r x r r x n

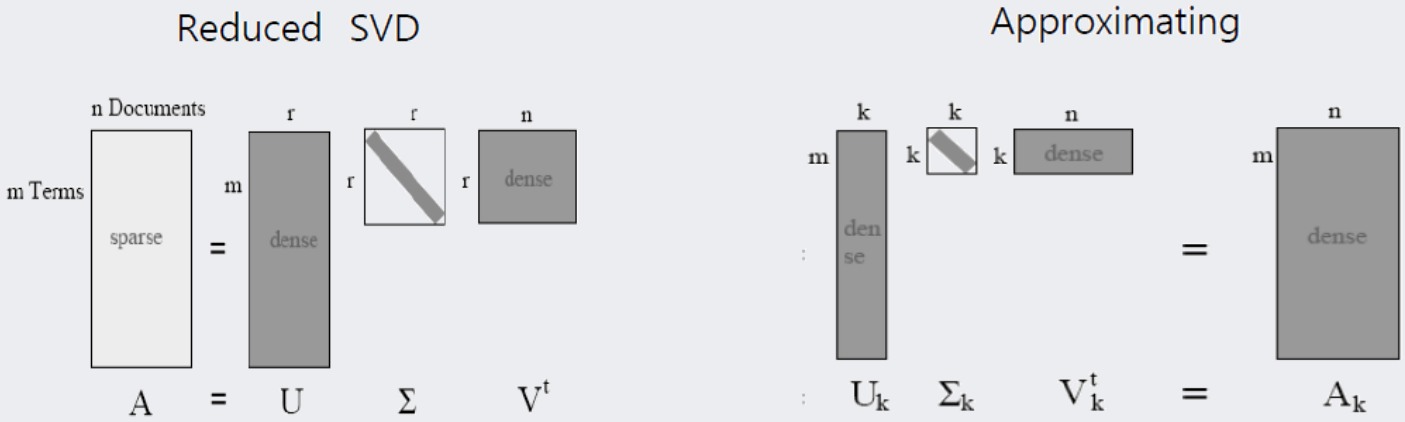


- Rectangular matrix A(m x n) can be decomposed as $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$
- All singular vectors of the matrices U and V are orthogonal $\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$
- Eigenvalues of the matrix ∑ is positive and sorted in a descending order

Feature Selection/Extraction

Lee (2010)

- SVD in Text Mining (Latent Semantic Analysis/Indexing)



✓ Step 1) Using SVD, a term-document matrix D is reduced to D_k

$$\mathbf{D} \approx \mathbf{D}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^{\mathrm{T}}$$

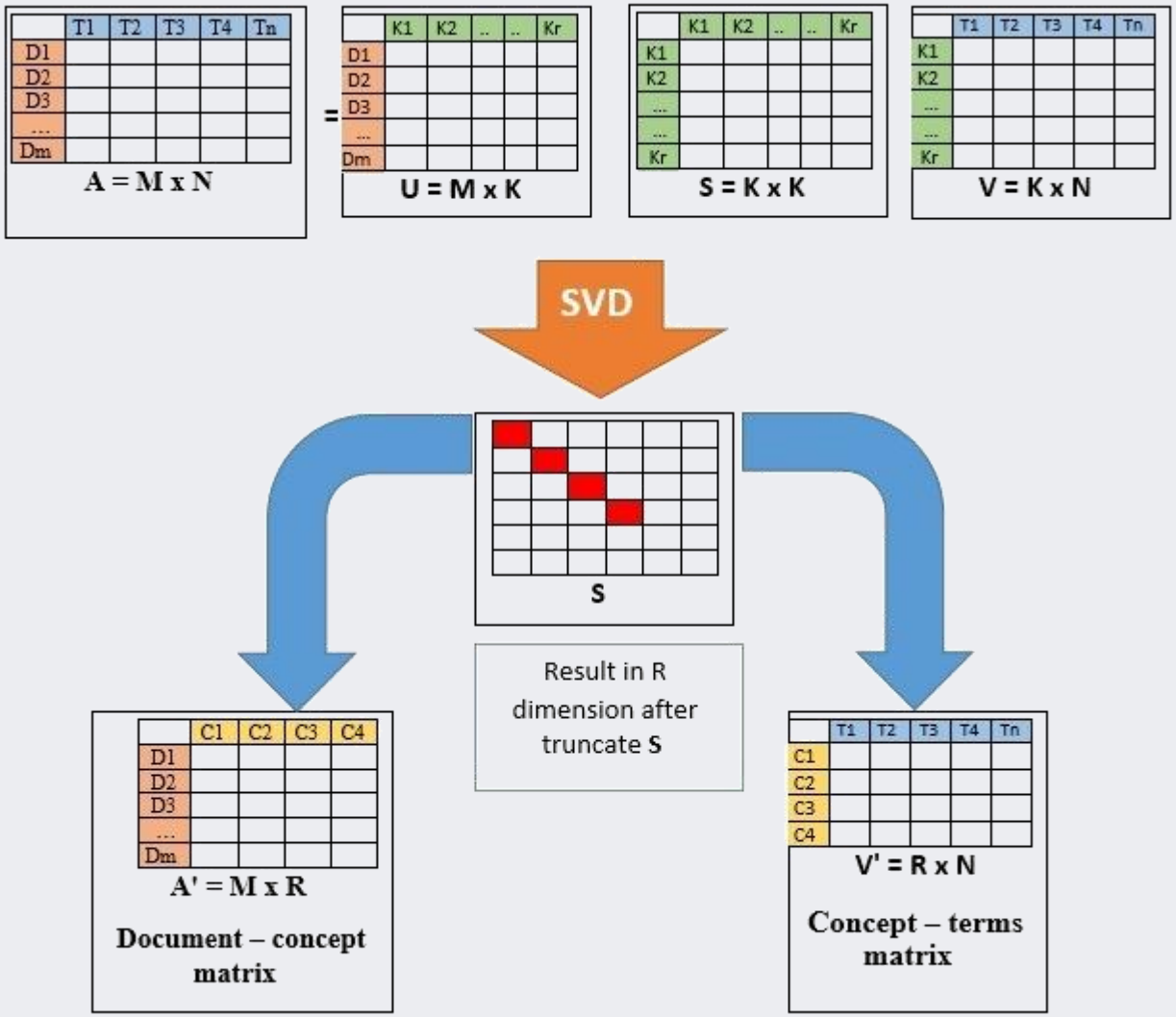
✓ Step 2) Multiply the transpose of the matrix U_k

$$\mathbf{U}_k^{\mathrm{T}} \mathbf{D}_k = \mathbf{\Sigma}_k \mathbf{V}_k^{\mathrm{T}}$$

✓ Apply data mining algorithms to the matrix obtained in the Step 2)

Feature Selection/Extraction

- LSA



https://www.quora.com/How-is-LSA-used-for-a-text-document-clustering

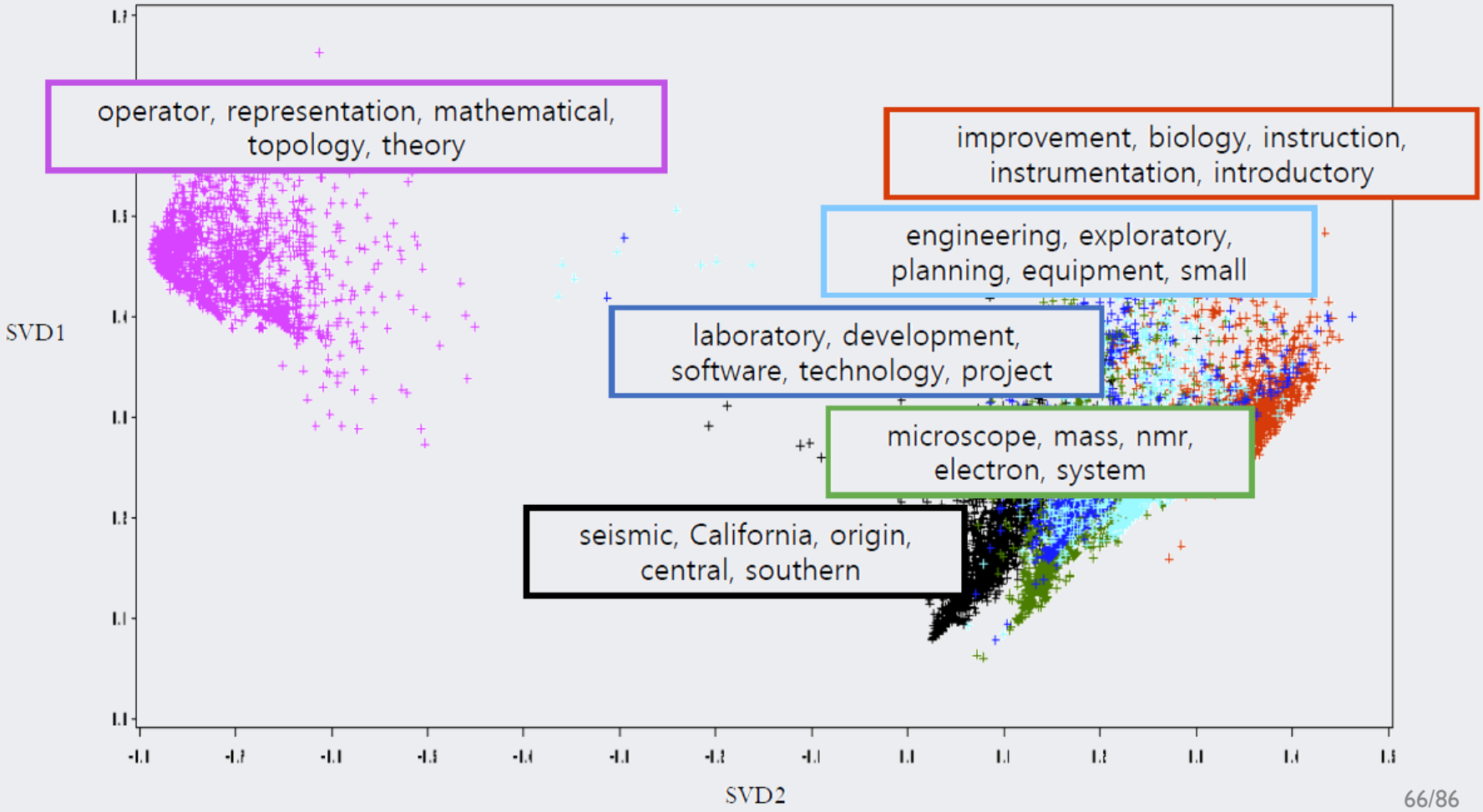
Feature Selection/Extraction

- LSA example
  - ✓ Data: 41,717 abstracts of the research projects that were funded by National Science Foundation (NSF) between 1990 and 2003
  - ✓ Top 10 positive and negative keywords for each SVD

(−) SVD1	(+)	(−) SVD2	(+)	(−) SVD3	(+)
sister−chromatid	real	sum	electronics	revised	evaporation
plastids	other	diophantine	enhanced	major	agglomerate
segregation	reduction	mathematical	video	introductory	mobility
pneumoniae	dimensional	yang−mills	computer−based	miniority	transient
males	complex	noetherian	improved	computer−based	multicomponent
candida	multiple	differential equations	major	micro computer	lateral
trait	expansion	equations	support	laboratory	distribution
decline	domain	invariant	instructional	unix−based	copolymer
factorial	vector	asymtotic	theme	ample	compacted
gmp	stability	non−commutative	capability	inquiry	long

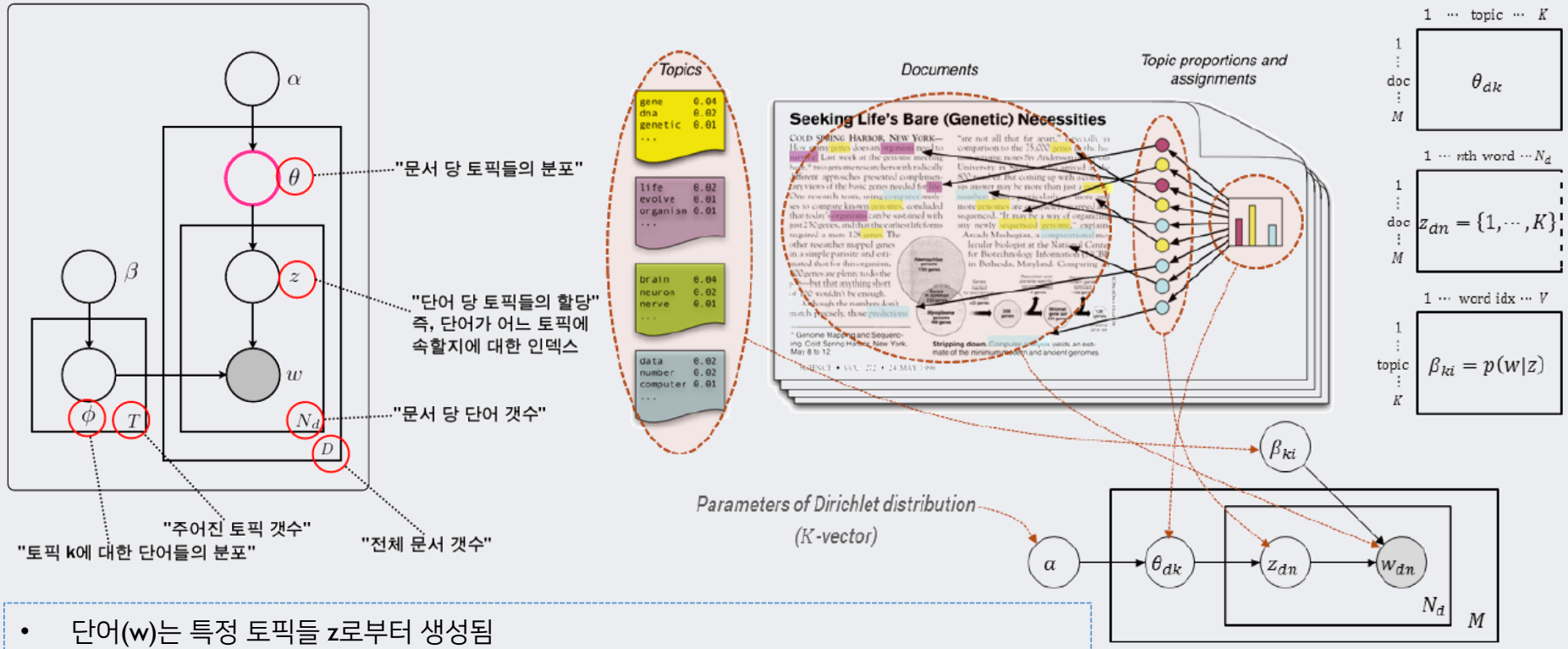
Feature Selection/Extraction

- LSA example
  - ✓Visualize the project in the reduced 2-D space



Feature Selection/Extraction

- Topic Modeling as a Feature Extractor
  - ✓ Latent Dirichlet Allocation (LDA)



- 단어(w)는 특정 토픽들 z로부터 생성됨
- 해당 문서가 어떤 토픽 비율(topic proportion, θ)을 가질 것인지는 파라미터가 α인 Dirichlet distribution에 의해 결정됨
- w만 실제로 관찰할 수 있는 값이고 θ, z, Φ는 숨겨진 값임
- 문서 생성 프로세스는 먼저 다항분포 θ로부터 토픽이 나타날 확률 θ를 추출하고 이를 바탕으로 단어가 나타날 확률인 w를 산출함

Feature Selection/Extraction

- Topic Modeling as a Feature Extractor
  - ✓ Two outputs of topic modeling
    - Per-document topic proportion
    - Per-topic word distribution

(a) Per-document topic proportions (θ_d)

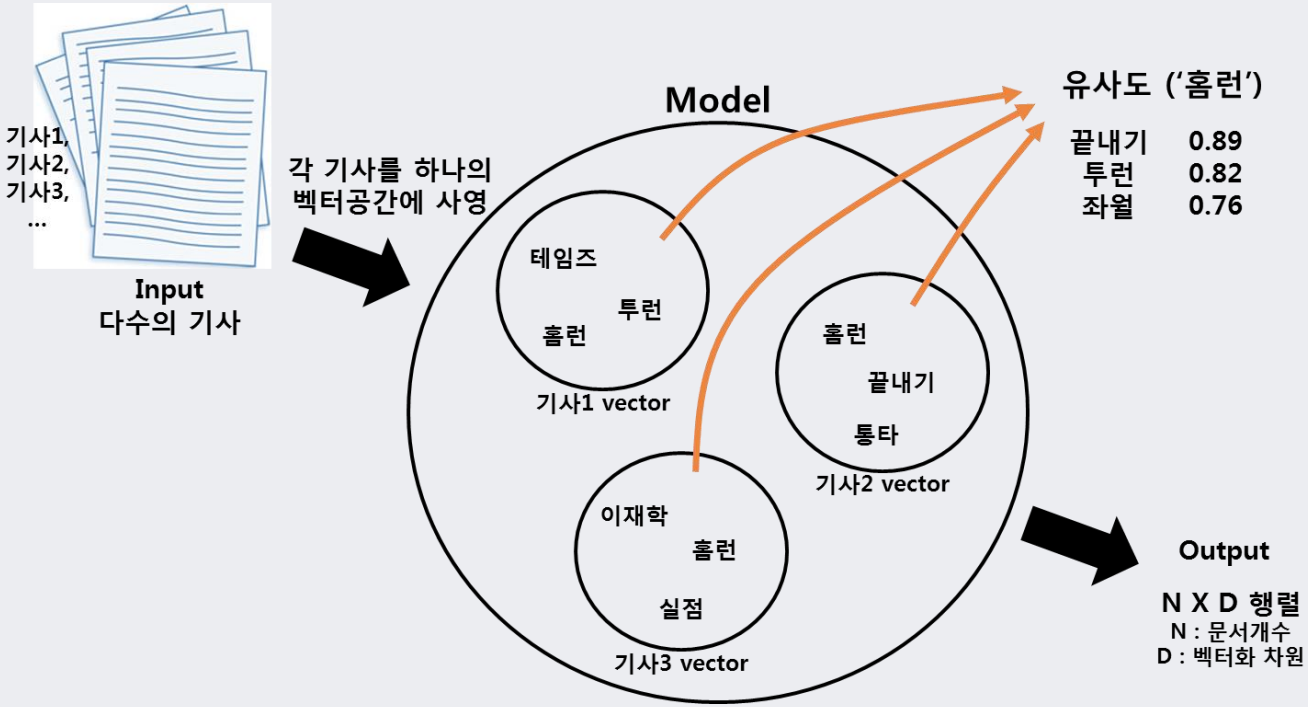
	Topic 1	Topic 2	Topic 3	…	Topic K	Sum
Doc 1	0.20	0.50	0.10	…	0.10	1
Doc 2	0.50	0.02	0.01	…	0.40	1
Doc 3	0.05	0.12	0.48	…	0.15	1
…	…	…	…	…	…	1
Doc N	0.14	0.25	0.33	…	0.14	1

(b) Per-topic word distributions (ϕ_k)

	Topic 1	Topic 2	Topic 3	…	Topic K
word 1	0.01	0.05	0.05	…	0.10
word 2	0.02	0.02	0.01	…	0.03
word 3	0.05	0.12	0.08	…	0.02
…	…	…	…	…	…
word V	0.04	0.01	0.03	…	0.07
Sum	1	1	1	1	1

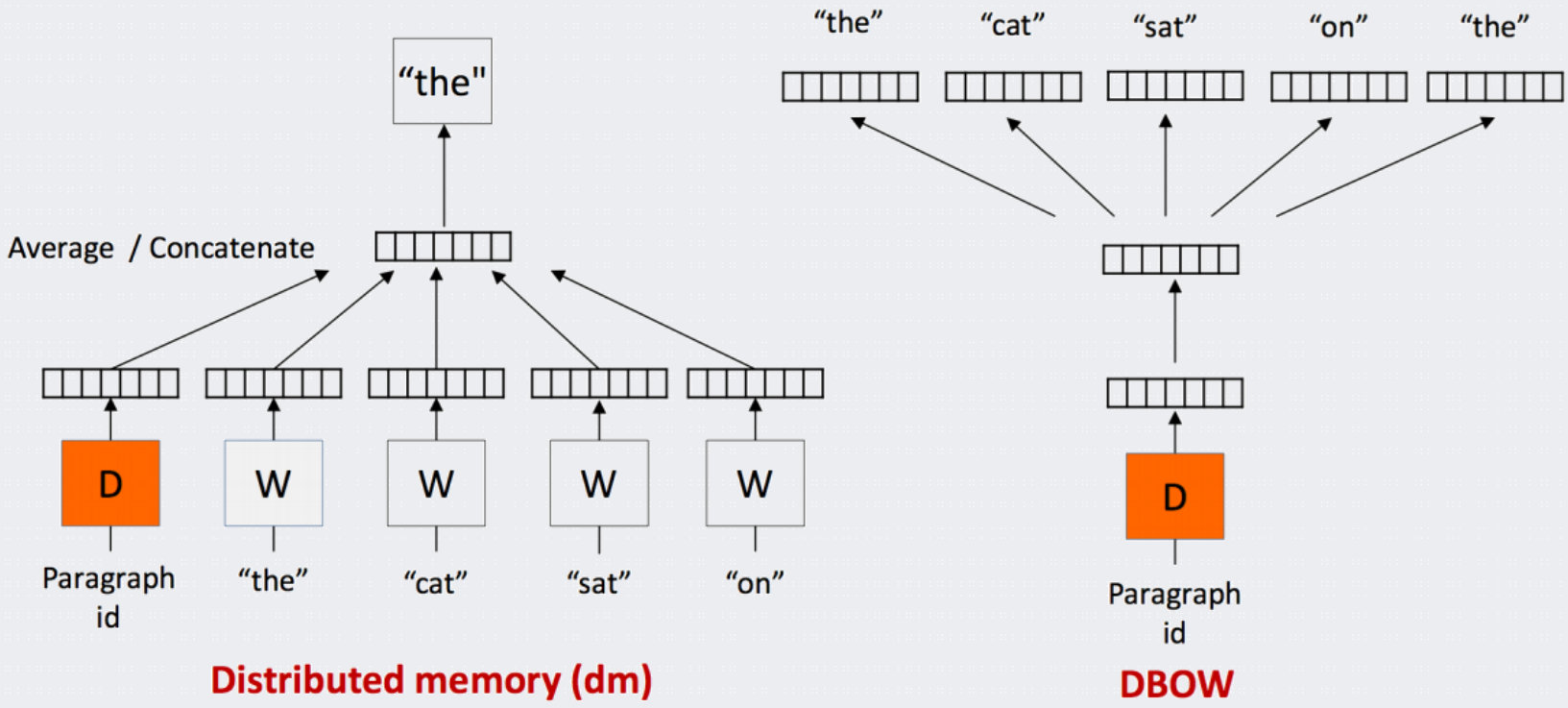
Feature Selection/Extraction

- Document to vector (Doc2Vec)
  - ✓ A natural extension of word2vec
  - ✓ Use a distributed representation for each document



Feature Selection/Extraction

- Document to vector (Doc2Vec)
  - ✓ A natural extension of word2vec
  - ✓ Use a distributed representation for each document



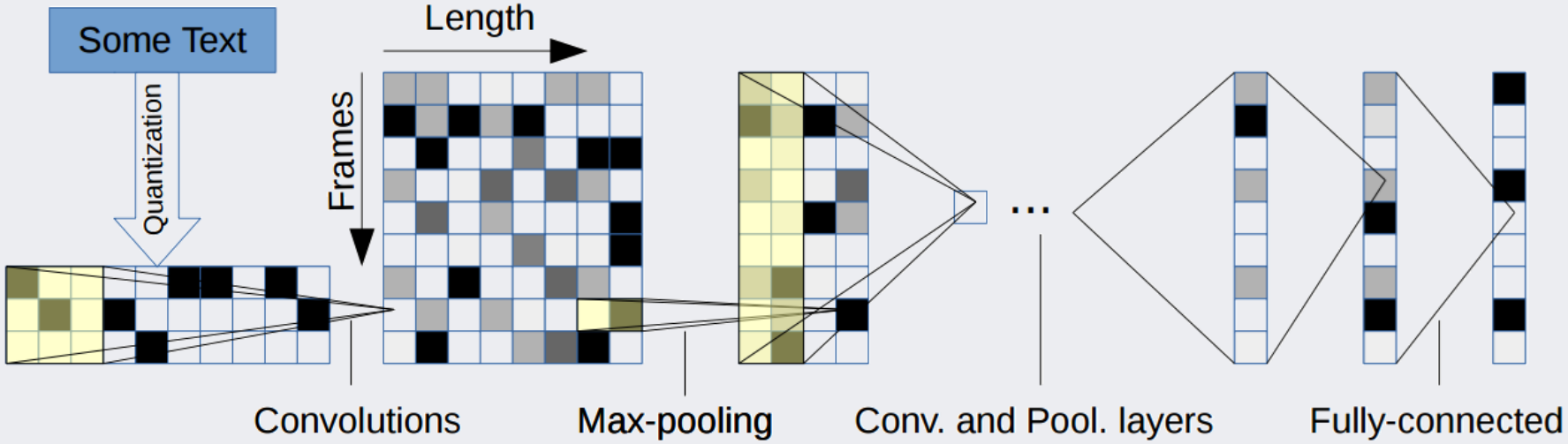
Feature Selection/Extraction

Zhang and LeCun (2015)

- No needs for feature engineering
  - ✓ 1-of-M encoding
    - 70 characters are encoded

```
abcdefghijklmnopqrstuvwxyz0123456789
-,;.!?:'''/\|_@#$%^&*~`+-=<>()[]{}
```

  - ✓ ConvNet Structure



Feature Selection/Extraction

Zhang and LeCun (2015)

- No needs for feature engineering
  - ✓ Performances for various TM tasks
    - Amazon review sentiment analysis

	Total	Full	Polarity
1	2,746,559	730,000	1,275,000
2	1,791,219	730,000	725,000
3	2,892,566	730,000	0
4	6,551,166	730,000	725,000
5	20,705,260	730,000	1,275,000

Full score

Model	Thesaurus	Train	Test
Large ConvNet	No	62.96%	58.69%
Large ConvNet	Yes	68.90%	59.55%
Small ConvNet	No	**69.24%**	59.47%
Small ConvNet	Yes	62.11%	**59.57%**
Bag of Words	No	54.45%	54.17%
word2vec	No	36.56%	36.50%

Polarity

Model	Thesaurus	Train	Test
Large ConvNet	No	**97.57%**	94.49%
Large ConvNet	Yes	96.82%	**95.07%**
Small ConvNet	No	96.03%	94.50%
Small ConvNet	Yes	95.44%	94.33%
Bag of Words	No	89.96%	89.86%
word2vec	No	72.95%	72.86%

Feature Selection/Extraction

Zhang and LeCun (2015)

- No needs for feature engineering
  - ✓ Performances for various TM tasks
    - Yahoo! Answers topic classification dataset

Category	Total	Train	Test
Society & Culture	295,340	140,000	6,000
Science & Mathematics	169,586	140,000	6,000
Health	278,942	140,000	6,000
Education & Reference	206,440	140,000	6,000
Computers & Internet	281,696	140,000	6,000
Sports	146,396	140,000	6,000
Business & Finance	265,182	140,000	6,000
Entertainment & Music	440,548	140,000	6,000
Family & Relationships	517,849	140,000	6,000
Politics & Government	152,564	140,000	6,000

Model	Thesaurus	Train	Test
Large ConvNet	No	73.42%	70.45%
Large ConvNet	Yes	**75.55%**	**71.10%**
Small ConvNet	No	72.84%	70.16%
Small ConvNet	Yes	72.51%	70.16%
Bag of Words	No	66.83%	66.62%
word2vec	No	56.37%	56.47%

Feature Selection/Extraction

Zhang and LeCun (2015)

- No needs for feature engineering
  - ✓ Performances for various TM tasks
    - News categorization

Category	Total	Train	Test
World	81,456	30,000	1,900
Sports	62,163	30,000	1,900
Business	56,656	30,000	1,900
Sci/Tech	41,194	30,000	1,900

Model	Thesaurus	Train	Test
Large ConvNet	No	99.44%	**87.18%**
Large ConvNet	Yes	**99.49%**	86.61%
Small ConvNet	No	99.20%	84.35%
Small ConvNet	Yes	96.81%	85.20%
Bag of Words	No	88.02%	86.69%
word2vec	No	78.20%	76.73%

Category	Total	Train	Test
Sports	645,931	90,000	12,000
Finance	315,551	90,000	12,000
Entertainment	160,409	90,000	12,000
Automobile	167,647	90,000	12,000
Technology	188,111	90,000	12,000

Model	Thesaurus	Train	Test
Large ConvNet	No	**99.14%**	**95.12%**
Small ConvNet	No	93.05%	91.35%
Bag of Words	No	92.97%	92.78%

AGENDA

01 Text Mining Overview

02 TM Process 1: Collection & Preprocessing

03 TM Process 2: Transformation

04 TM Process 3: Dimensionality Reduction

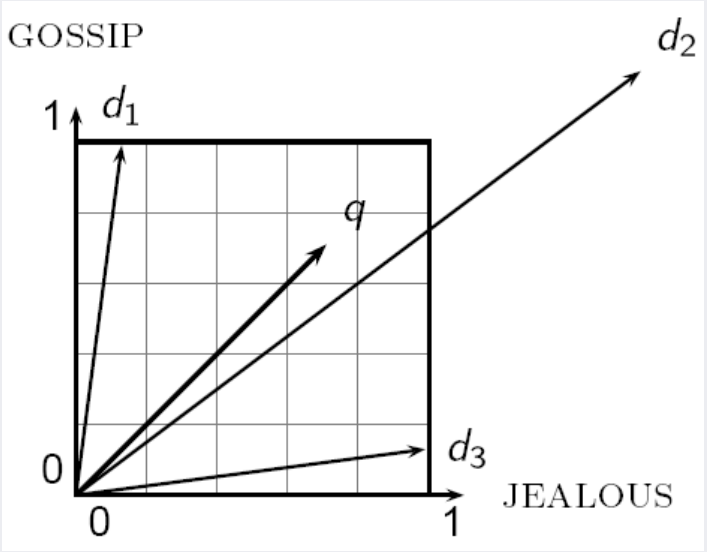
05 TM Process 4: Learning & Evaluation

Similarity Between Documents

- Document similarity
  - ✓ Use the cosine similarity rather than Euclidean distance
    - Which two documents are more similar?

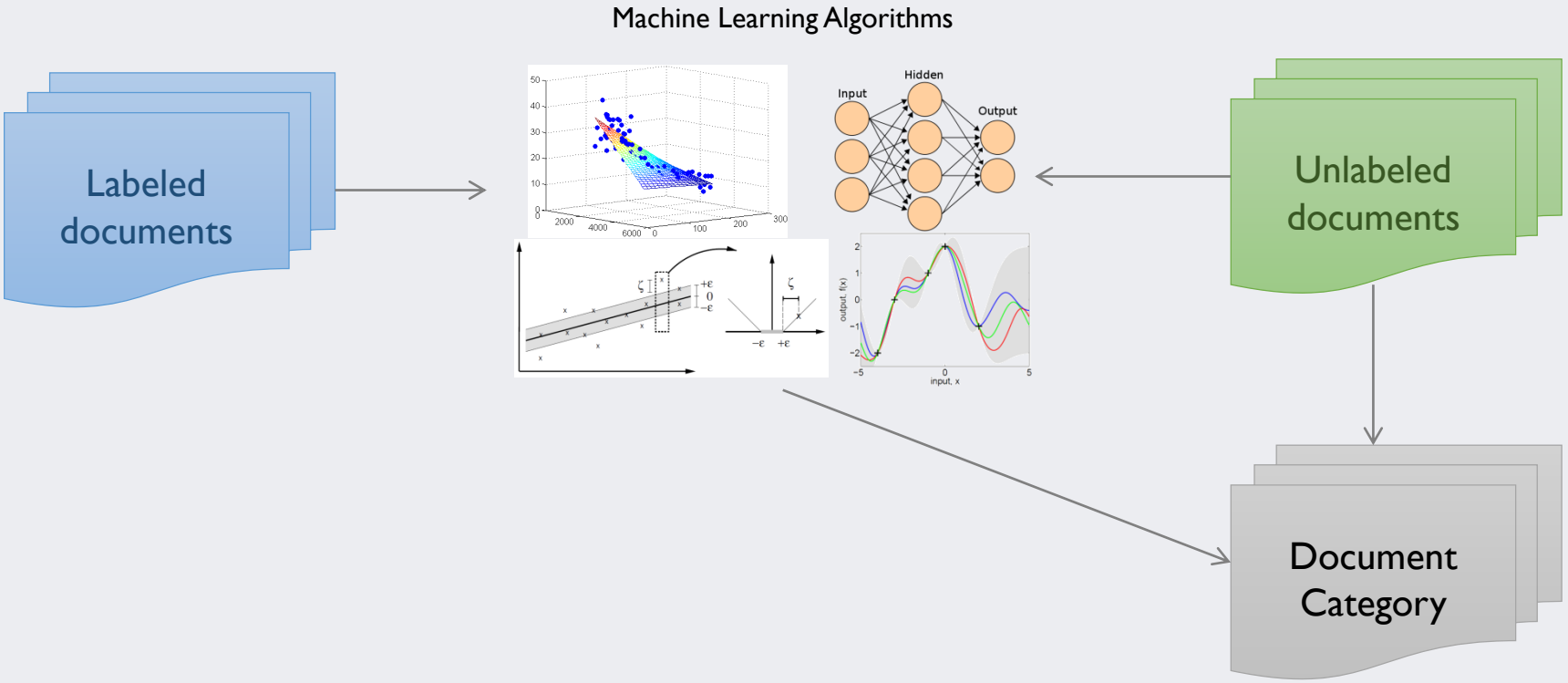
Doc.	Word 1	Word 2	Word 3
Document 1	1	1	1
Document 2	3	3	3
Document 3	0	2	0

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\sqrt{\sum_k x_k^2}}$$



Learning Task 1: Classification

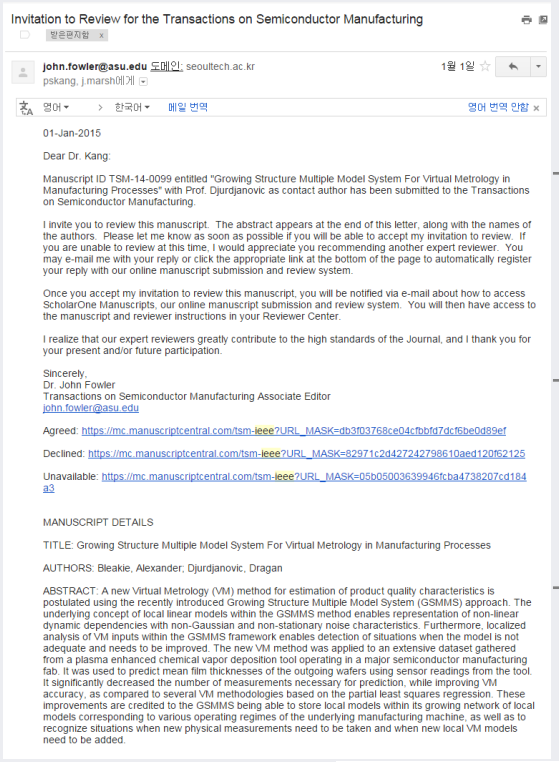
- Document categorization (classification)
  - ✓ Automatically classify a document into one of the pre-defined categories



Learning Task 1: Classification

- Spam filtering

Raw Data


Invitation to Review for the Transactions on Semiconductor Manufacturing

john.fowler@asu.edu 도메인: seoultech.ac.kr 1월 1일

pskang, j.marsh에게

영어 › 한국어 메일 번역 영어 번역 안함

01-Jan-2015

Dear Dr. Kang:

Manuscript ID TSM-14-0099 entitled "Growing Structure Multiple Model System For Virtual Metrology in Manufacturing Processes" with Prof. Djurdjanovic as contact author has been submitted to the Transactions on Semiconductor Manufacturing.

I invite you to review this manuscript. The abstract appears at the end of this letter, along with the names of the authors. Please let me know as soon as possible if you will be able to accept my invitation to review. If you are unable to review at this time, I would appreciate you recommending another expert reviewer. You may e-mail me with your reply or click the appropriate link at the bottom of the page to automatically register your reply with our online manuscript submission and review system.

Once you accept my invitation to review this manuscript, you will be notified via e-mail about how to access ScholarOne Manuscripts, our online manuscript submission and review system. You will then have access to the manuscript and reviewer instructions in your Reviewer Center.

I realize that our expert reviewers greatly contribute to the high standards of the Journal, and I thank you for your present and/or future participation.

Sincerely,
Dr. John Fowler
Transactions on Semiconductor Manufacturing Associate Editor
john.fowler@asu.edu

Agreed: https://mc.manuscriptcentral.com/tsm-ieee?URL_MASK=db3f03768ce04cfbbfd7dcf6be0d89ef

Declined: https://mc.manuscriptcentral.com/tsm-ieee?URL_MASK=82971c2d42724279861 0aed120f62125

Unavailable: https://mc.manuscriptcentral.com/tsm-ieee?URL_MASK=05b05003639946fcba4738207cd184a3

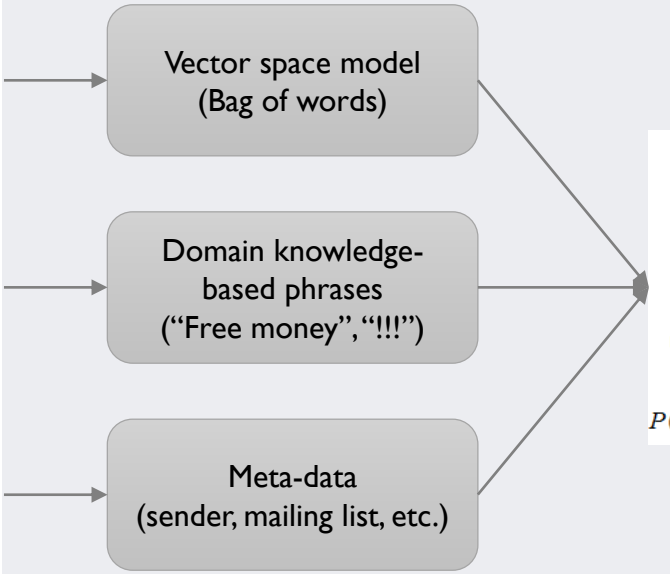
MANUSCRIPT DETAILS

TITLE: Growing Structure Multiple Model System For Virtual Metrology in Manufacturing Processes

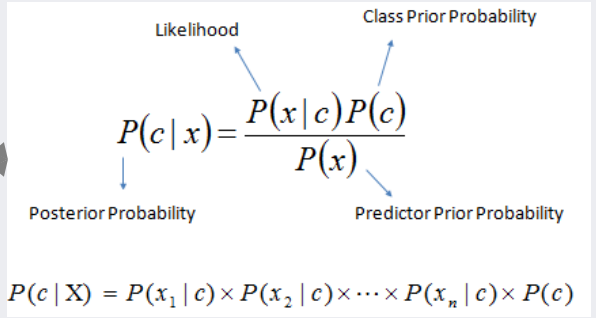
AUTHORS: Bleakie, Alexander; Djurdjanovic, Dragan

ABSTRACT: A new Virtual Metrology (VM) method for estimation of product quality characteristics is postulated using the recently introduced Growing Structure Multiple Model System (GSMMS) approach. The underlying concept of local linear models within the GSMMS method enables representation of non-linear dynamic dependencies with non-Gaussian and non-stationary noise characteristics. Furthermore, localized analysis of VM inputs within the GSMMS framework enables detection of situations when the model is not adequate and needs to be improved. The new VM method was applied to an extensive dataset gathered from a plasma enhanced chemical vapor deposition tool operating in a major semiconductor manufacturing fab. It was used to predict mean film thicknesses of the outgoing wafers using sensor readings from the tool. It significantly decreased the number of measurements necessary for prediction, while improving VM accuracy, as compared to several VM methodologies based on the partial least squares regression. These improvements are credited to the GSMMS being able to store local models within its growing network of local models corresponding to various operating regimes of the underlying manufacturing machine, as well as to recognize situations when new physical measurements need to be taken and when new local VM models need to be added.

Features



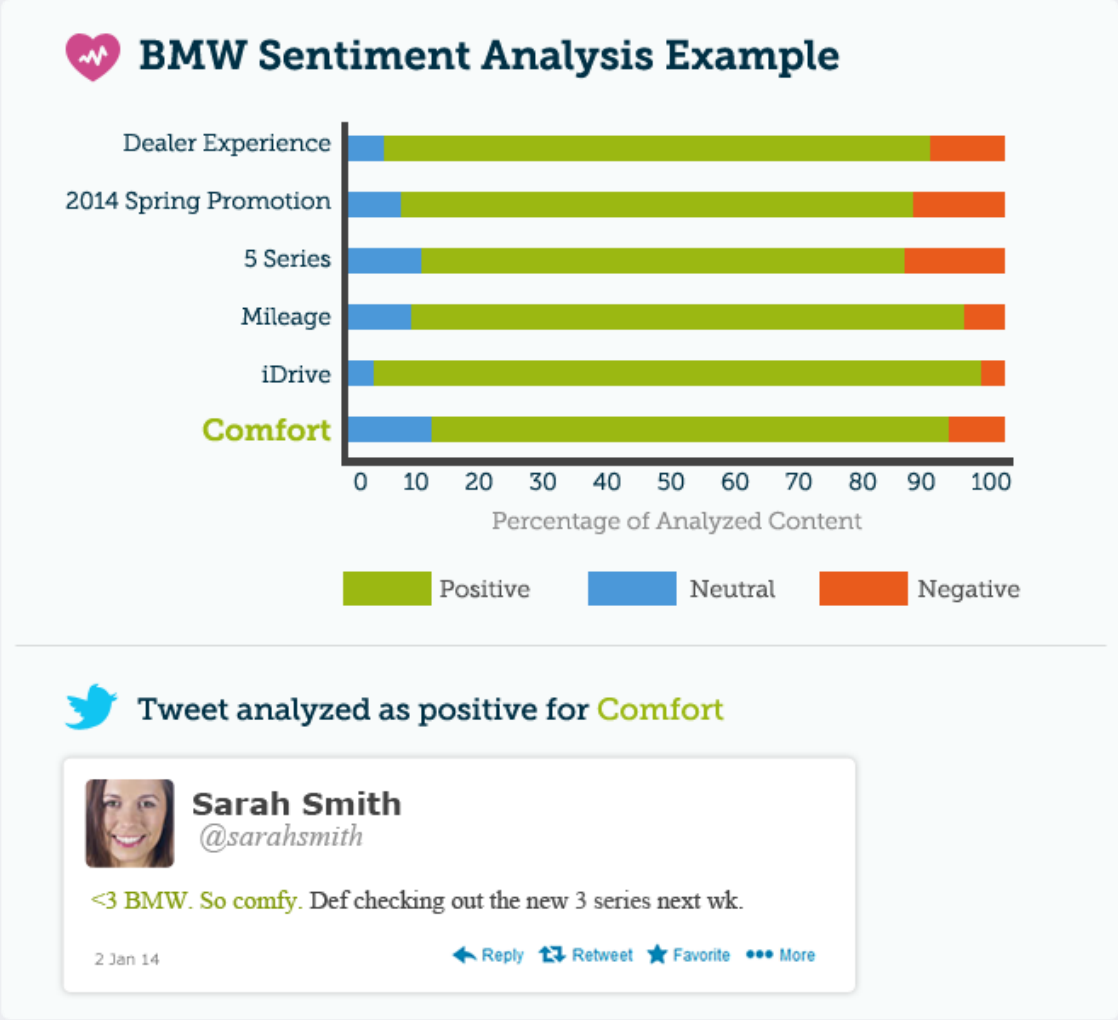
Model



Feature Regime	Junk		Legitimate	
	Precision	Recall	Precision	Recall
Words only	97.1%	94.3%	87.7%	93.4%
Words + Phrases	97.6%	94.3%	87.8%	94.7%
Words + Phrases + Domain-Specific	100.0%	98.3%	96.2%	100.0%

Learning Task 1: Classification

- Sentiment Analysis



http://www.crowdsource.com/solutions/content-moderation/sentiment-analysis/

Learning Task 1: Classification

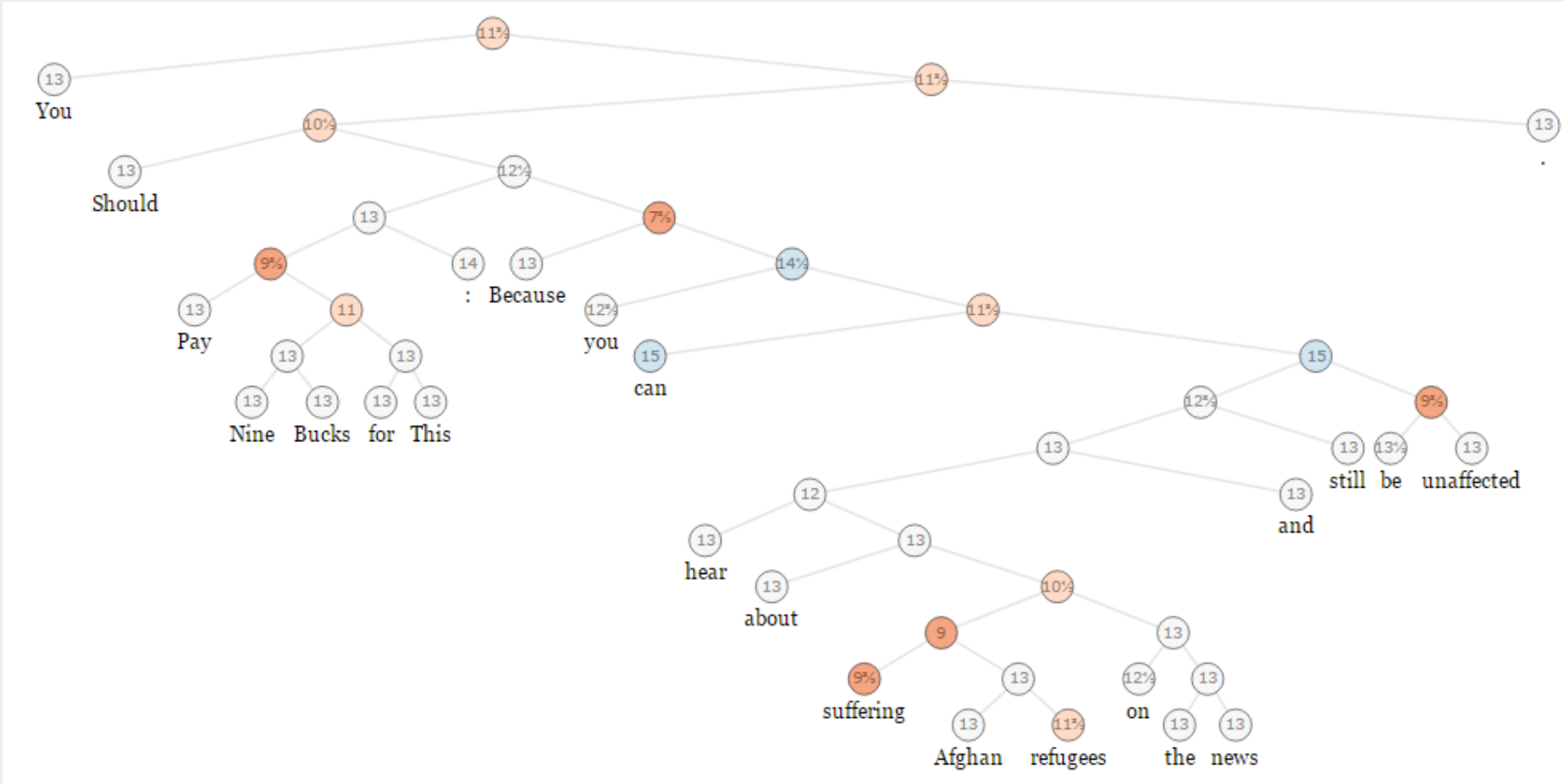
Socher et al. (2013)

- Sentiment Analysis
  - ✓ Sentiment tree bank @Stanford NLP Lab

Learning Task 1: Classification

Socher et al. (2013)

- Sentiment Analysis
  - ✓ Sentiment tree bank @Stanford NLP Lab



Learning Task 2: Clustering

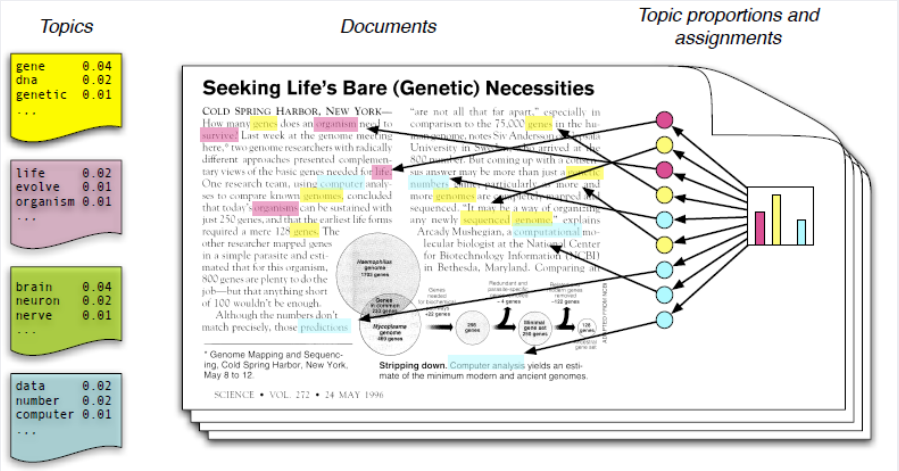
- Document Clustering & Visualization
  - ✓ Have a top level view of the topics in the corpora
  - ✓ See relationships between topics
  - ✓ Understand better what's going on

Raw Data

Source.title	Year	Title	Abstract
IEEE Transactions on Enginee	2014	Impact of interorganizational relationshi	Interorganizational relationships are an integral part of an organization's decision-
IEEE Transactions on Enginee	2014	Exploring the boundary conditions of dis	We explore the boundary conditions of Christensen's theory of disruption. Using te
IEEE Transactions on Enginee	2014	Two stage procurement processes with c	This paper considers a sourcing problem faced by a manufacturer who outsources
IEEE Transactions on Enginee	2014	Identifying emerging customer requirem	Rapid changes of new technologies, market dynamics, and swift fluctuation of cus
IEEE Transactions on Enginee	2014	Supply chain quality integration: Antece	This study extends quality management from an individual company perspective to
IEEE Transactions on Enginee	2014	Achieving IT-enabled enterprise agility ir	An ongoing challenge that enterprises currently face is that of fostering agile minc
IEEE Transactions on Enginee	2014	Accessing external technological knowle	This study elucidates how a firm accesses external technological knowledge (TK) a
IEEE Transactions on Enginee	2014	An investigation on centralized and dece	Design for supply chain has become an essential consideration while designing a
IEEE Transactions on Enginee	2014	ISO 9000 internalization and organizatio	This study presents an investigation of the relationship between internalization of
IEEE Transactions on Enginee	2014	Knowledge specialization in ph.D. Studer	Researchers have argued that specialization within groups yields productivity gains
IEEE Transactions on Enginee	2014	An overlapping-based design structure r	To gain competitive leverage, firms that design and develop complex products see
IEEE Transactions on Enginee	2014	The TOC-based algorithm for solving mu	In a previous paper in this journal named The TOC-Based Algorithm for Solving M
IEEE Transactions on Enginee	2014	Toward environmental and quality sustai	In the last decade, industrial research for sustainable development and environmer
IEEE Transactions on Enginee	2014	Balancing exploration and exploitation ir	This study investigates the influence of supply chain portfolios on the paradox of
IEEE Transactions on Enginee	2014	Managing nuclear spare parts inventorie	This paper presents a methodology for developing a spare parts inventory manage
IEEE Transactions on Enginee	2014	Scenario-based modeling of interdepenc	This paper develops a scenario generation framework for modeling demand and s
IEEE Transactions on Enginee	2014	Governing the portfolio management pr	Past strategy and innovation research has basically ignored potential effects of po
IEEE Transactions on Enginee	2014	Is an efficacious operation a safe operat	One of the most fundamental indicators of a facility's social sustainability is the he
IEEE Transactions on Enginee	2014	Managing highly innovative projects: Thr	The climate change debate and economic recovery strategies in various industries
IEEE Transactions on Enginee	2014	Improved mean and variance estimating	The program evaluation and review technique (PERT) is a popular method for mea
IEEE Transactions on Enginee	2014	Openness and appropriation: Empirical e	The adoption of open innovation creates a dilemma for firms. On one hand, a com
IEEE Transactions on Enginee	2014	The redundancy queuing-location-alloca	Redundancy queuing-location-allocation problems (RQLAPs) involve the economic
IEEE Transactions on Enginee	2014	Response to buyout options in internet a	A buyout option allows a bidder to acquire an auctioned product immediately at a
IEEE Transactions on Enginee	2014	Contracting a development supplier in th	The development of important components for a new product is frequently outsou
IEEE Transactions on Enginee	2014	Exploration, exploitation, and growth thr	This study investigates how firm age and environmental adversity influence explora
IEEE Transactions on Enginee	2014	Innovation and performance: The role of	Innovation has become the cornerstone for achieving competitive advantage and i
IEEE Transactions on Enginee	2014	Spotting lemons in platform markets: A (	This study addresses the understudied question of how content integrity for digita
IEEE Transactions on Enginee	2014	Business models and tactics in new proc	Effectuation and causation are two contrasting approaches to new business develc
IEEE Transactions on Enginee	2014	Energy technological change and capaci	This paper explores the role of learning in managing the capacities of existing and

- 8,850 articles from 11 Journals for the recent 10 years
- 21,434 terms after preprocessing

Features

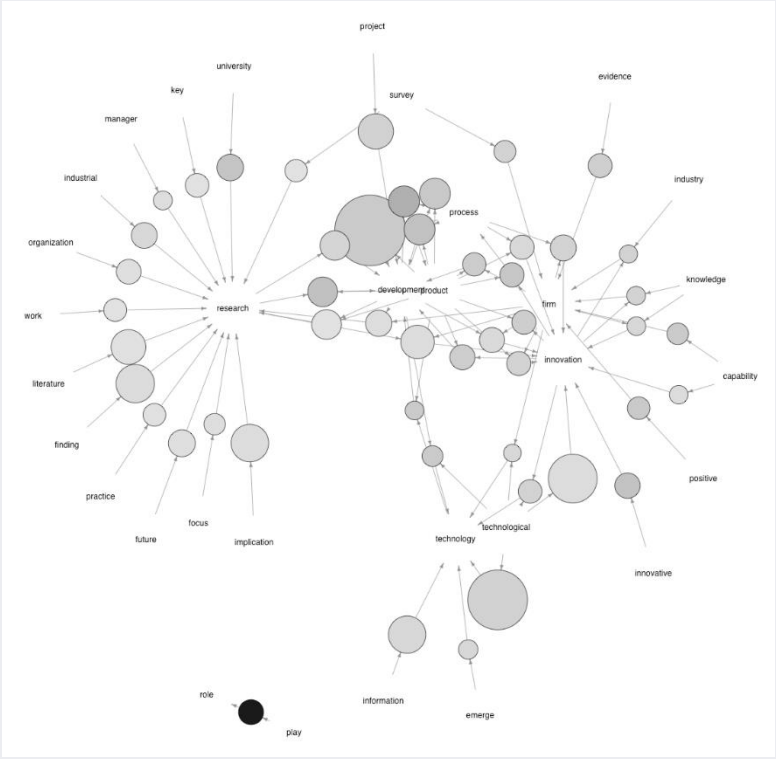


- 50 topics from Latent Dirichlet Allocation (LDA)

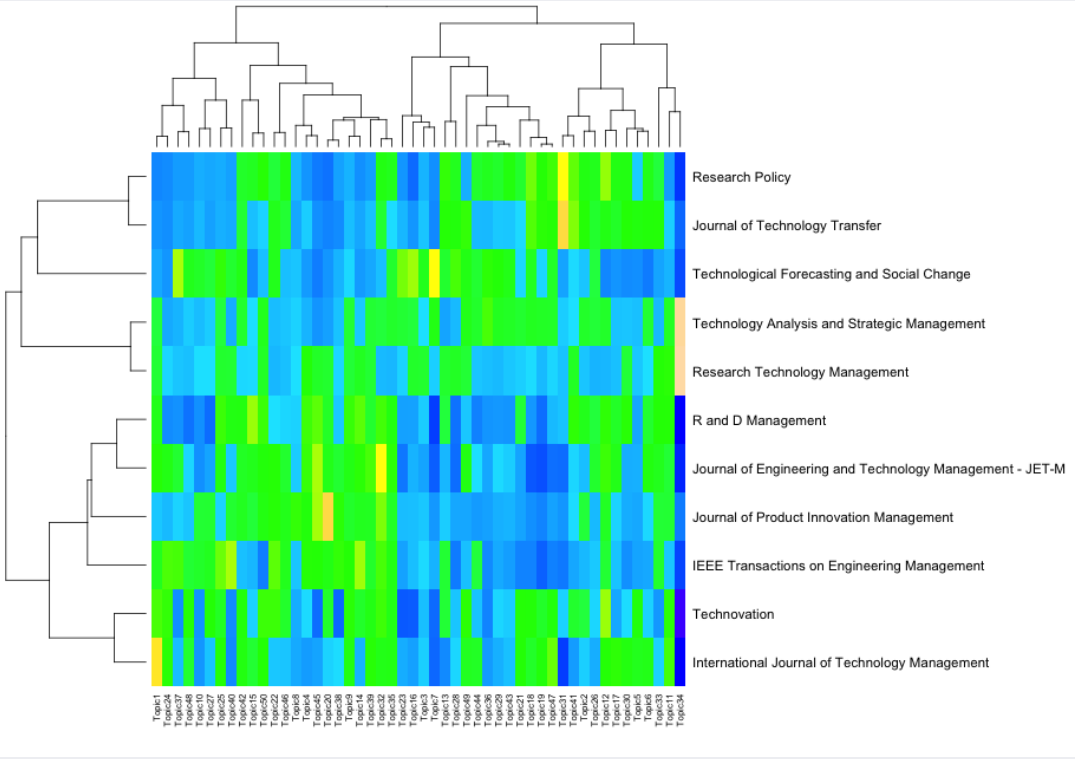
Learning Task 2: Clustering

- Document Clustering & Visualization

Keywords association

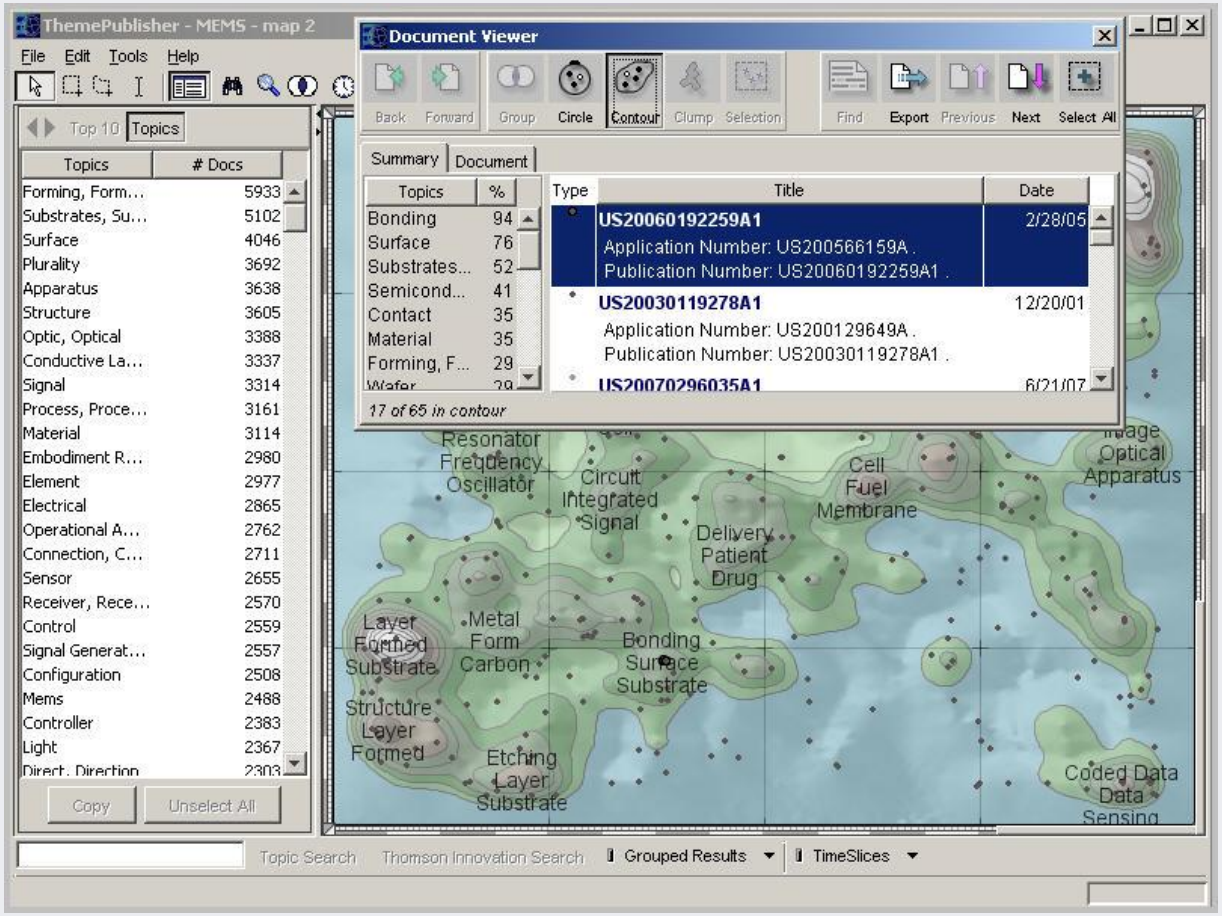


Journal/Topic Clustering



Learning Task 2: Clustering

- Document Clustering & Visualization
  - ✓ Themescape: contents maps from Thomson innovation full text patent data



Learning Task 3: Information Extraction/Retrieval

Yang et al. (2013)

- Information extraction/retrieval
  - ✓ Find useful information from text databases
  - ✓ Examples: identify the relationships between the symptoms and medicines (Yang et al., 2013)

연구대상 : 방약합편 (方藥合編)

- 한의학의 경험과 지식 축적은 주로 서적의 형태로 정리
- 방약합편 : 동의보감이 발간된 이후, 병증을 중점으로 두고 각 병증의 말미에 처방을 제시한 서적
- 병증에 따라 유용한 처방을 한눈에 제시
- 처방에 필요한 약재(본초)들을 명시

Learning Task 3: Information Extraction/Retrieval

- Information retrieval (cont')
  - ✓ Construct keyword database

Materia medica/prescription	Ginseng Radix	Citri Pericarpium	Pinelliae Rhizoma	Poria(red)	...	...
Prescription 1	1	0	0	0		
Prescription 2	1	1	0	1		
Prescription 3	0	0	0	1		
Prescription 4	0	0	1	1		
...						
Prescription 521	1	1	0	0		

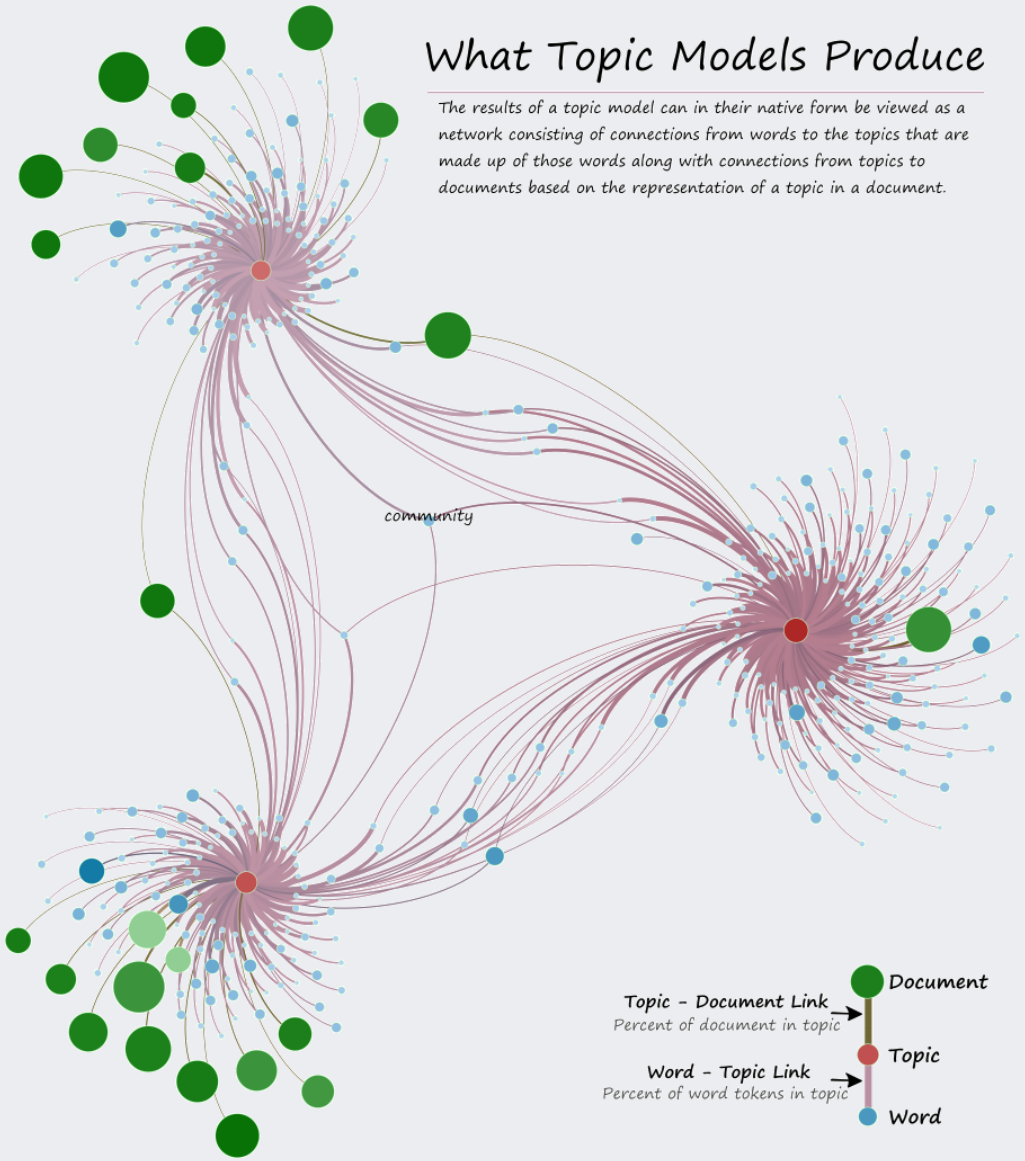
Learning Task 3: Information Extraction/Retrieval

- Information retrieval (cont')
  - ✓ Perform association rule to extract symptom-medicine relationships

Antecedent (symptom)	Consequent (herbal materials)	Support (%)	Confidence (%)	Lift
Coughs	Pinelliae Rhizoma	4.2	46.0	2.2
	Citri Pericarpium	3.9	42.9	1.3
	Rehmanniae Radix Preparat	2.2	23.8	1.8
	Poria(red)	2.0	22.2	1.3
	Poria(white)	2.0	22.2	1.1
	Armeniacae Semen	1.9	20.6	3.0
Overexertion/Fatigue	Rehmanniae Radix Preparat	3.5	70.5	4.5
	Angelicae Gigantis Radix	3.3	67.6	1.9
	Ginseng Radix	2.0	41.1	1.4
	Poria(white)	1.9	38.2	1.9
	Dioscoreae Rhizoma	1.7	35.3	5.0
	Paeoniae Radix Alba	1.7	35.3	1.8
	Atractylodis Rhizoma Alba	1.6	32.3	1.2
	Corni Fructus	1.4	29.4	5.8
	Capreoli Cornu	1.3	26.5	8.2
	Astragali Radix	1.3	26.5	2.4
	Achyranthis Radix	1.2	23.5	5.0
	Pulvis Aconiti Tuberis Purificatum	1.2	23.5	2.8
	Cinnamomi Cortex Spissus	1.2	23.5	1.8
	Moutan Cortex	1.0	20.6	3.6
	Schizandrae Fructus	1.0	20.6	3.4

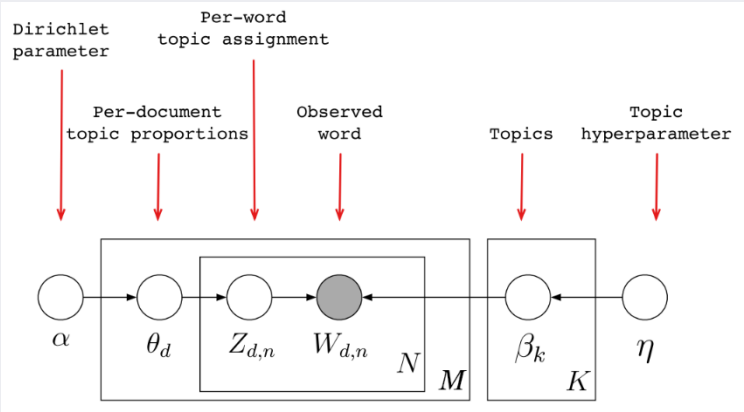
Learning Task 3: Information Extraction/Retrieval

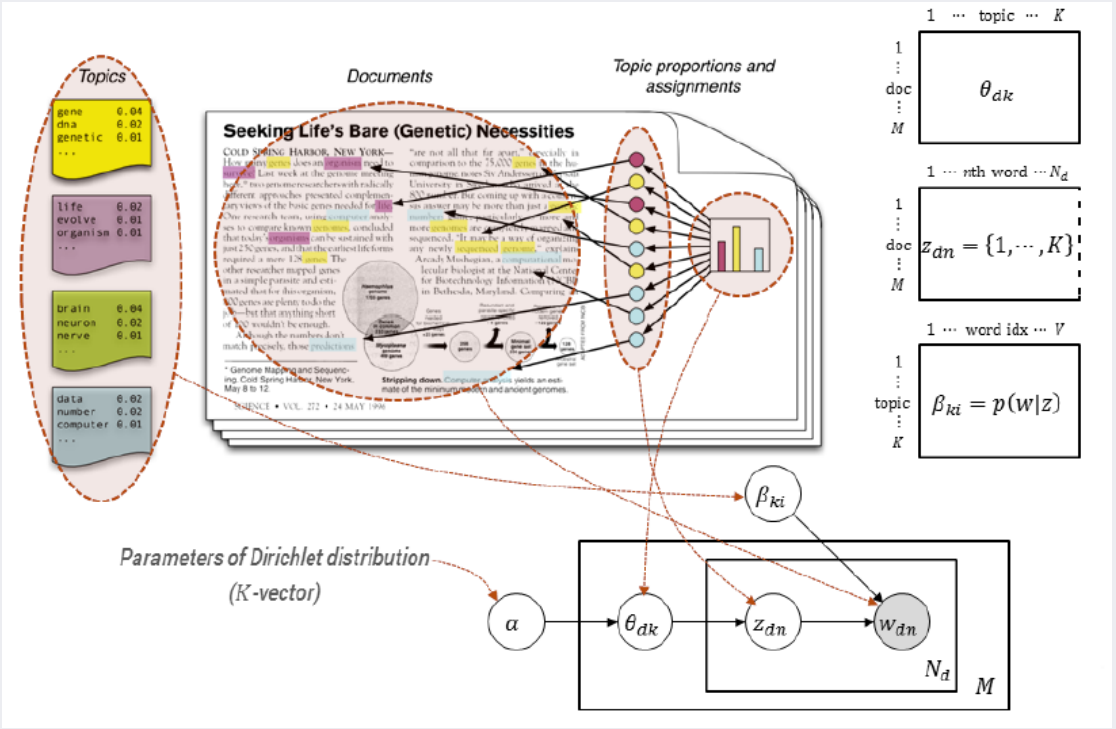
- Topic Modeling
  - ✓ A suite of algorithms that aim to discover and annotate large archives of documents with thematic information
  - ✓ Statistical methods that analyze the words of the original texts to discover
    - the themes that run through them
    - how themes are connected to each other
    - how they change over time



Learning Task 3: Information Extraction/Retrieval

- Latent Dirichlet Allocation (LDA)





- Words (w) are statistically generated from the topics Z
- Topic proportion for a document (θ_d) is determined by the Dirichlet distribution with the parameter α
- We can only observe W; θ, z, Φ are latent variables (hidden, cannot be observed)
- Document generation process: (1) Estimate the topic proportions θ, (2) estimate the word probability w from θ


ANY
questions?

References

Research Papers

- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
- Kim, T., Hong, J., & Kang, P. (2015). Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting 31(2), 364-390.
- Kim, H., Park, M., & Kang, P. (2016). 토픽모델링과 사회연경망을 통한 딥러닝 연구동향 분석. 대한산업공학회 춘계공동학술대회, 제주.
- Lee, G., Jeong, J., Seo, S., Kim, C., & Kang, P. (2017+). Sentiment classification with word attention based on weakly supervised learning with a convolutional neural network, under review.
- Lee, H. & Kang, P. (2017+). Identifying core topics in technology and innovation management studies: A topic model approach. Journal of Technology Transfer. Online available https://link.springer.com/content/pdf/10.1007%2Fs10961-017-9561-4.pdf
- Mo, K. Park, J., Jang, M., & Kang, P. (2017+). 단어와 자소 기반 합성공 신경망을 이용한 문서 분류, under review.
- Park, Y., Kim, H., Lee, H., Kim, D., Kim, S., & Kang, P. (2017). A deep learning-based sports player evaluation model based on game statistics and news articles, Knowledge-Based Systems 138, 15-36.
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, p. 1642).
- Yang, D.H, Kang, J.H., Park, Y.B., Park, Y.J., Oh, H.S., & Kim, S.B.* (2013). Association rule mining and network analysis in oriental medicine, PLOS ONE, Vol. 8, Issue 3 (March 15), e59241.
- Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. arXiv preprint arXiv:1502.01710.

References

Other materials

- Abbott, D. (2013). Introduction to Text Mining. Virtual Data Intensive Summer School. http://www.vscse.org/summerschool/2013/Abbott.pdf
- Lee, Y. (2010). Data Mining & Text Mining
- Sahami, M., Dumais, S., Hecherman, D. (1998). A Bayesian Approach to Filtering Junk E-mail, Learning for Text Categorization: Papers from the 1998 Workshop.
- Witte, R. (2006). Introduction to Text Mining. Tutorial at EDBT'06. http://rene-witte.net/text-mining-tutorial-edbt2006