

Lecture I: Introduction to Text Mining

Pilsung Kang

School of Industrial Management Engineering
Korea University

AGENDA

- 01 **Text Mining Overview**
- 02 TM Process I: 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



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>

Example: The BioTech Industry

Access to information is a serious problem

- 80% of biological knowledge is **only** in research 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

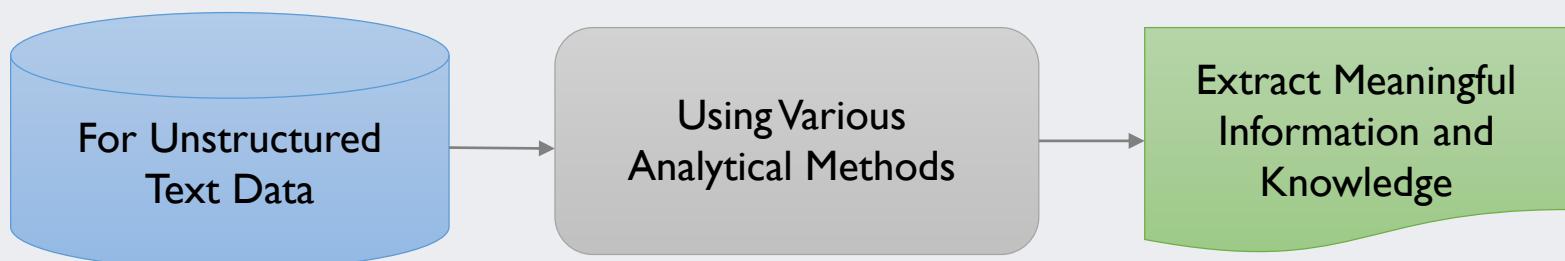
- MedLine adds more than 10 000 abstracts each *month*!
- Chemical Abstracts Registry (CAS) registers 4000 entities **each day**, 2.5 million in 2004 alone

[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.



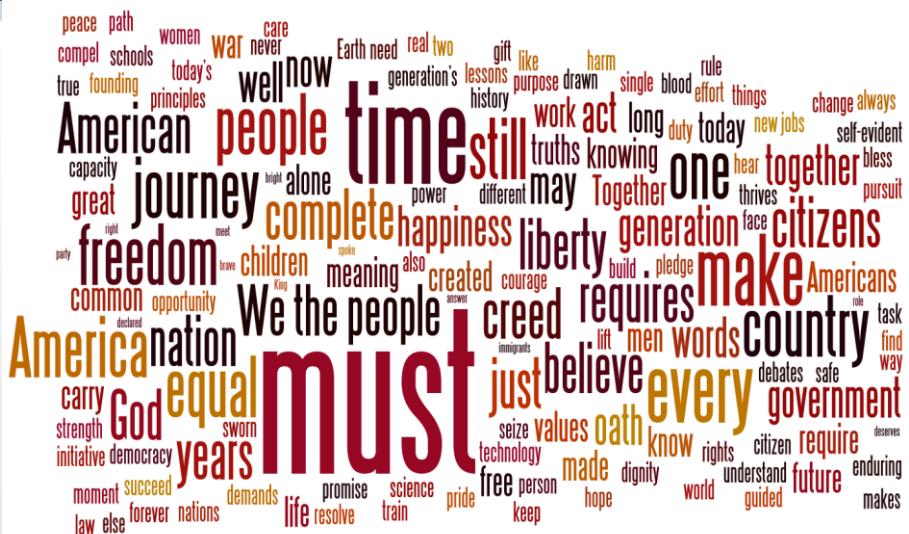
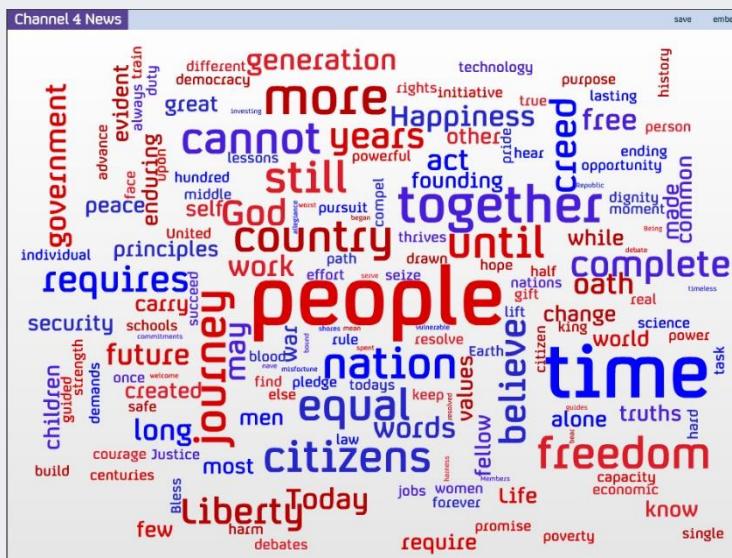
Text Mining: Applications

- Information Abstraction/Summarization/Visualization



Text Mining: Applications

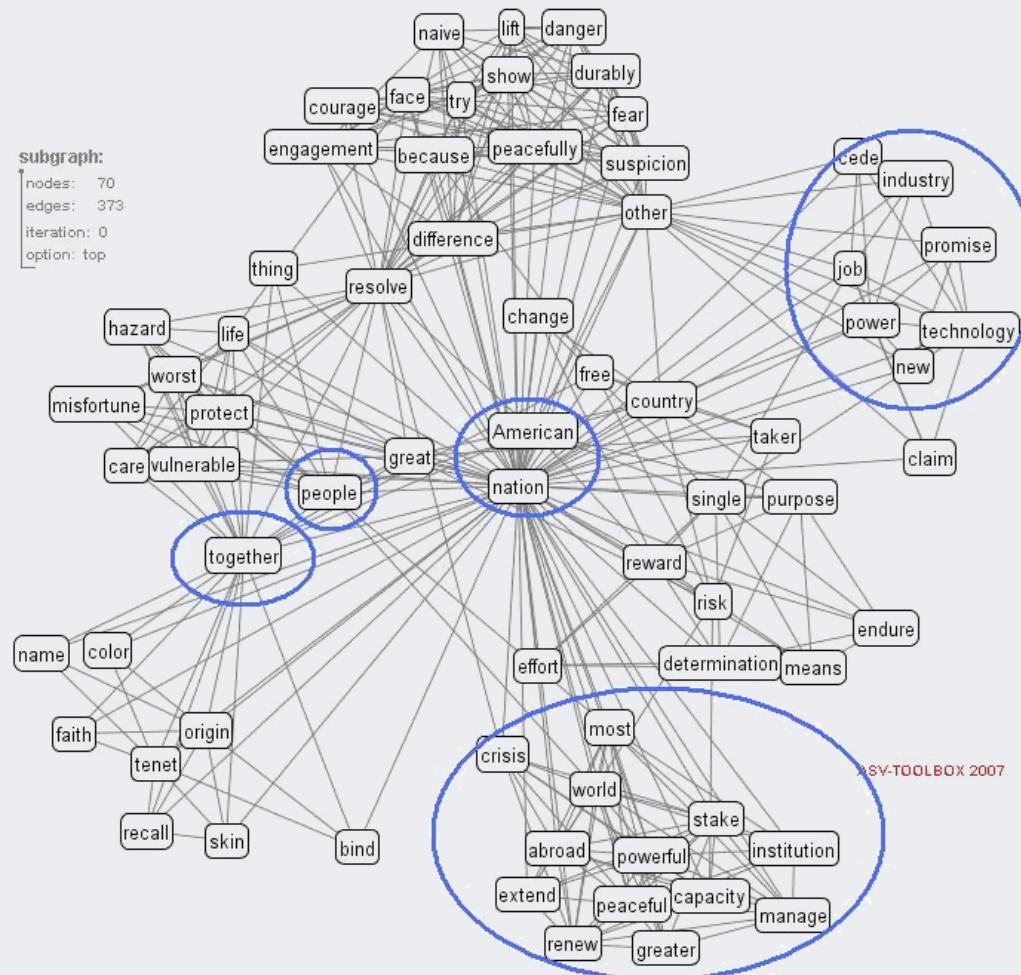
- Information Abstraction/Summarization/Visualization



Text Mining: Applications

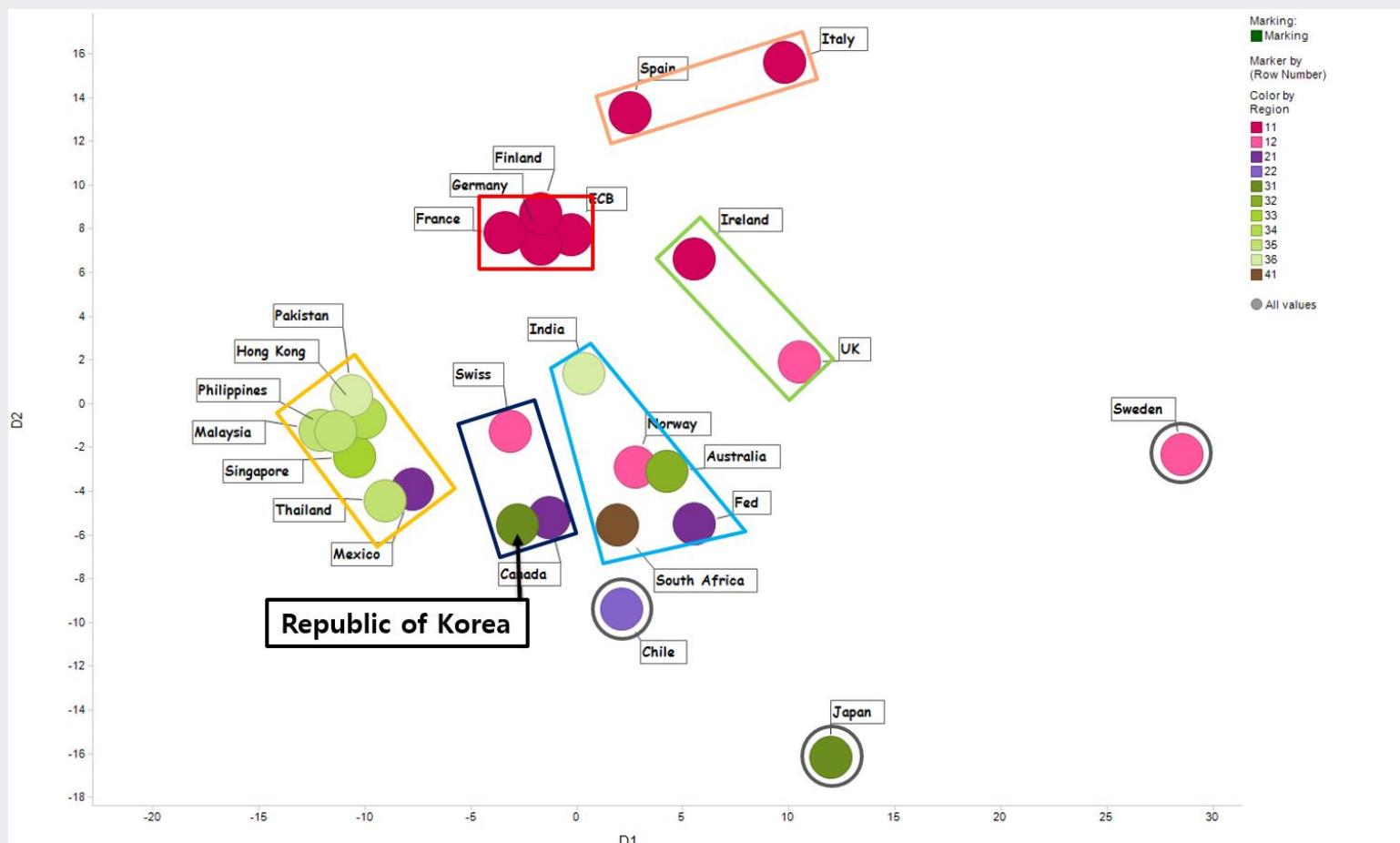
- Information Abstraction/Summarization/Visualization

2013년 오바마 취임 연설 주제어 연결망 - nation



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

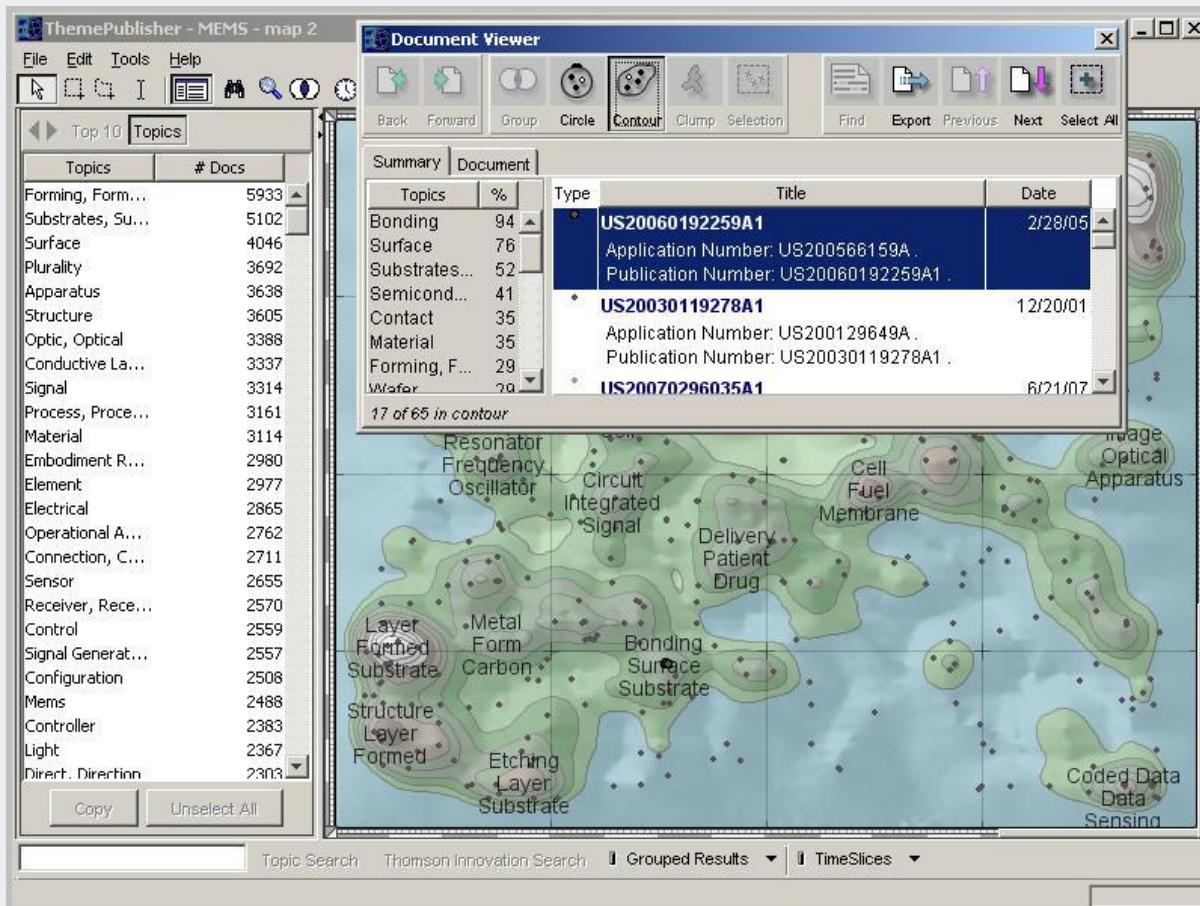
✓ Central Bank speech analysis: Similarities between the central banks in the world

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 Macropredential	Recovery/ QE VAT/ TAX	Deflation
2011	Sovereign Debt Basel III	Dodd-Frank 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

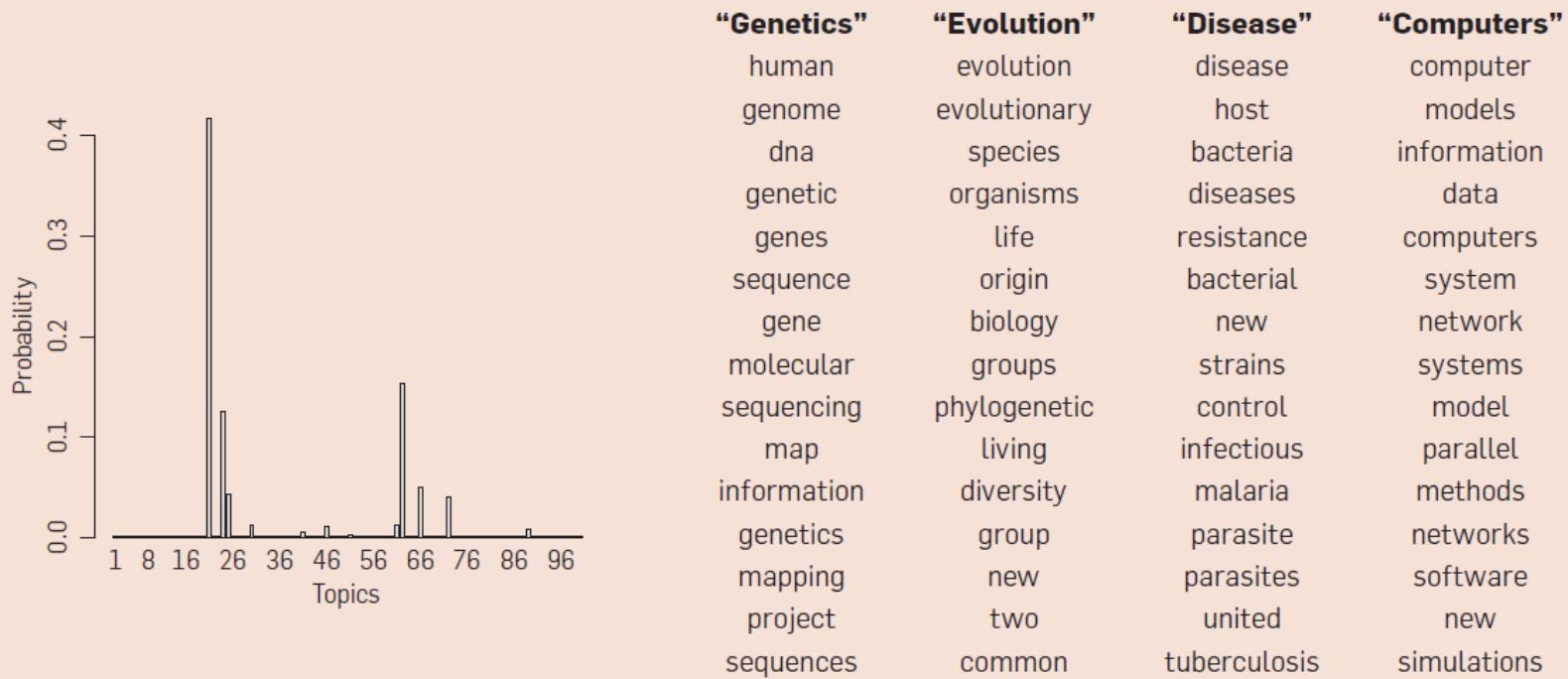


Text Mining: Applications

- 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.



Text Mining: Applications

- Topic Extraction

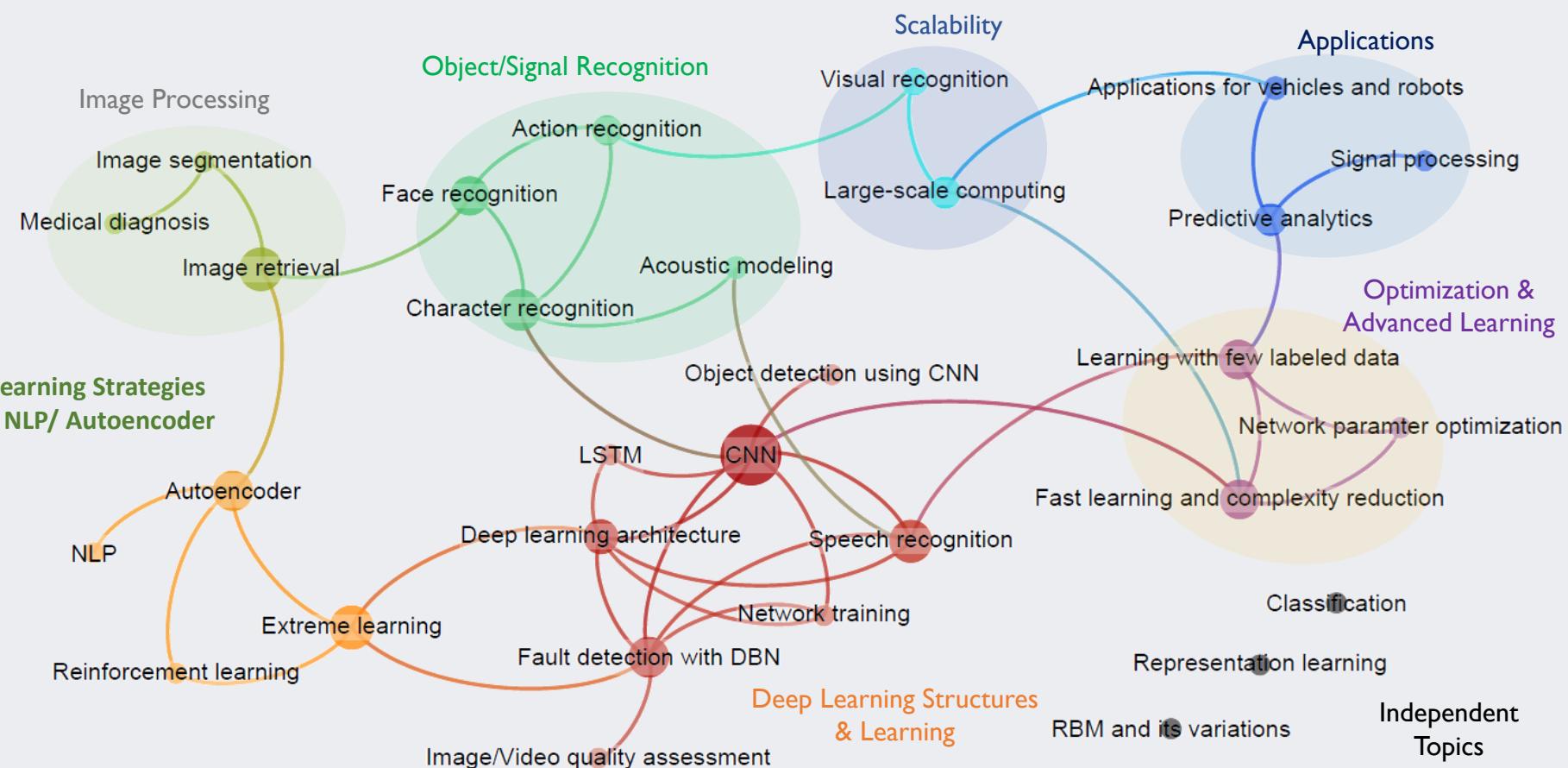
✓ 30 Topics discovered by LDA

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 belief network dbn fault	neural convolutional pool convolution convnet	layer input output unit hide function	feature level extract learn extraction	face recognition estimation facial shape	speaker speech noise adaptation source	speech recognition acoustic hmm neural	deep learn algorithm structure extreme	deep architecture neural standard explore	image scene scale segmentation pixel
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 recurrent long lstm network	data prediction technique information research	analysis filter signal component audio	classification classifier class vector support	application implementation efficient process power	domain state quality resolution relationship	pattern process compute visual field	word text language representation semantic	cnn detection convolutional neural detect	video human temporal action track
Image retrieval	Medical image diagnosis	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 visual retrieval descriptor attribute	image segmentation disease cell medical	learn question state answer reinforcement	train algorithm gradient sample optimization	representation learn sparse encode stack	machine boltzmann rbm restrict distribution	train data label few transfer	fast reduce parameter weight complexity	time real application drive Vehicle	recognition system character network neural

Text Mining: Applications

- Topic Extraction

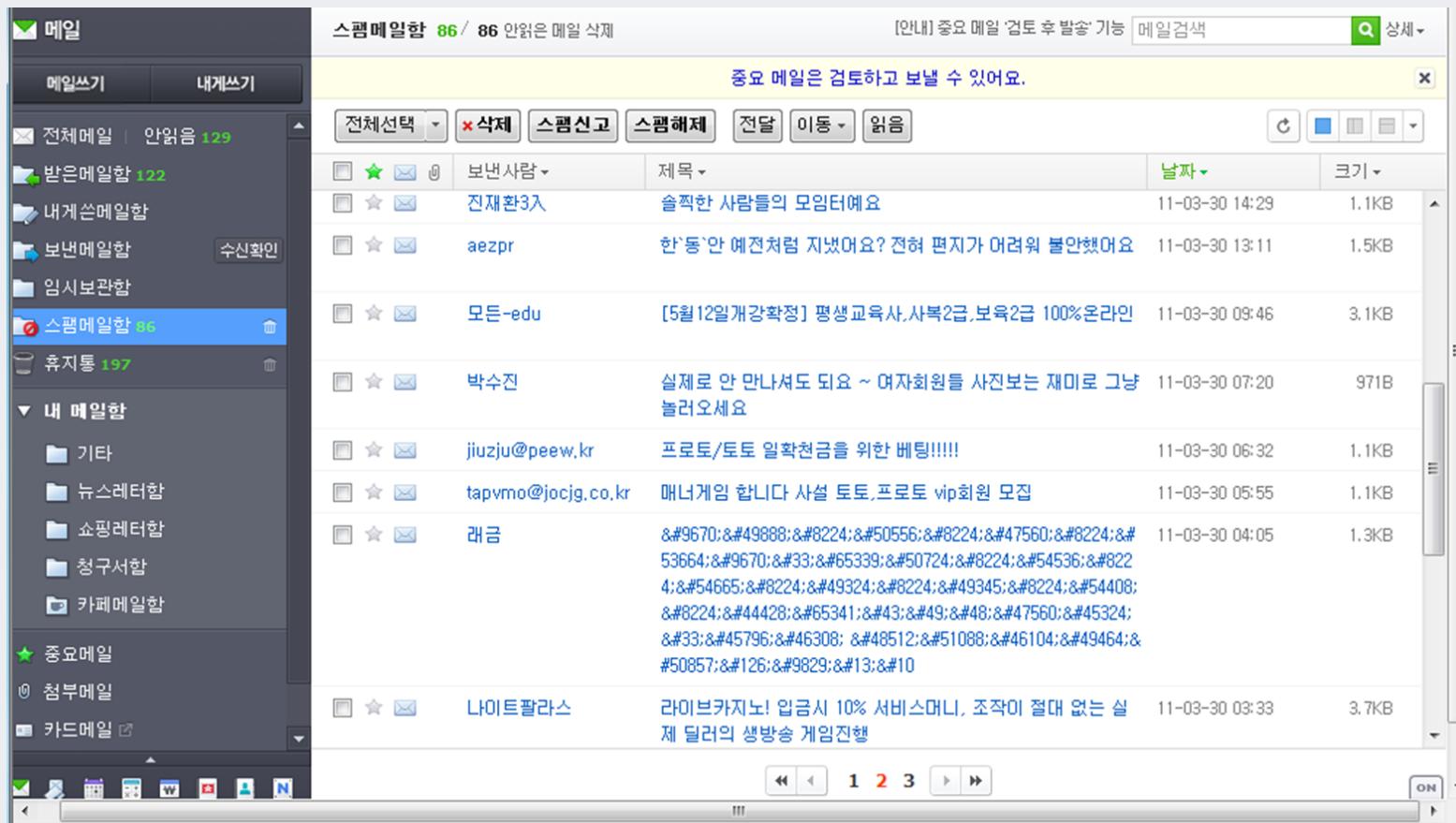
- ✓ Relations between topics



Text Mining: Applications

- Document Categorization/Classification

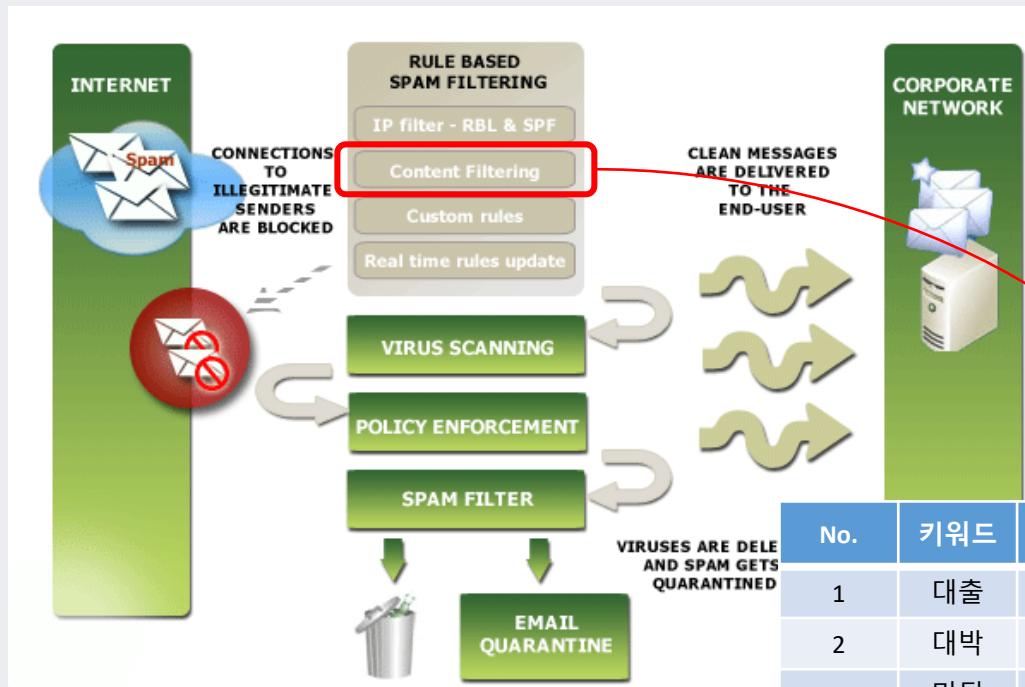
- ✓ Spam mail filtering



Text Mining: Applications

- Document Categorization/Classification

- ✓ Spam mail filtering

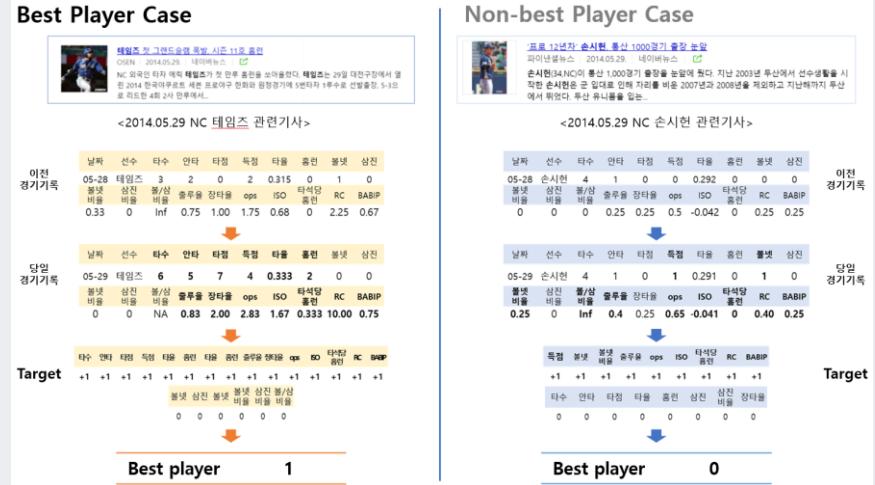
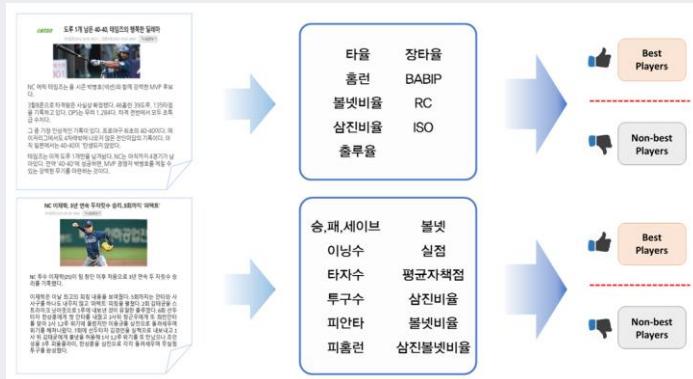


No.	키워드	Mail 1	Mail 2	Mail 3	...	Mail N
1	대출	0	2	0	...	0
2	대박	0	0	0	...	0
3	미팅	0	0	2	...	0
4	이상형	0	0	2	...	0
5	머니	0	2	0	...	0
6	외로	0	0	3	...	1
스팸 여부		N	Y	Y	...	N

Text Mining: Applications

- Document Categorization/Classification

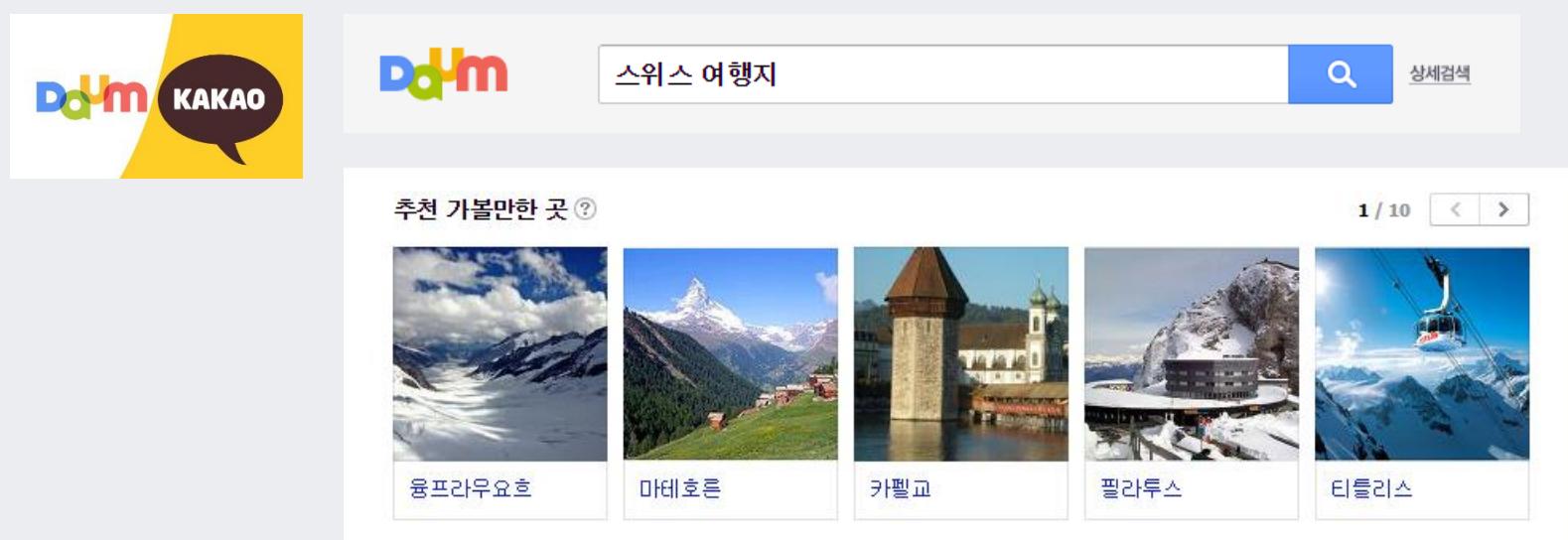
- ✓ Sport player evaluation



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

- 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 $\geq 100,000$

Screening data

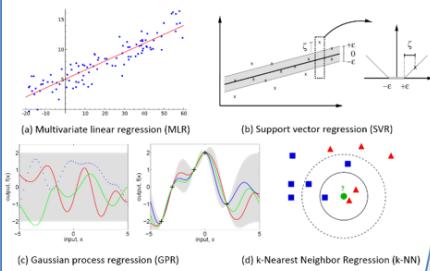
- KOFIC
- www.kobis.or.kr



4 Single Forecasting Model Development & Evaluation

Single forecasting models

- 1 linear algorithm
- 3 machine learning-based algorithms



2 Model Configuration & Variable Definition

Model configuration

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

Three weeks prior to release Two weeks prior to release One week prior to release One week after release Two weeks after release Three weeks after release



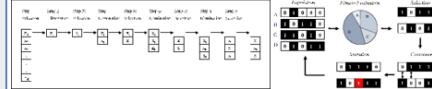
Variable definition

- Model R : 25 variables (1 screening, 24 SNS)
- Model W₁: 32 variables (4 screening, 28 SNS)
- Model W₂: 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

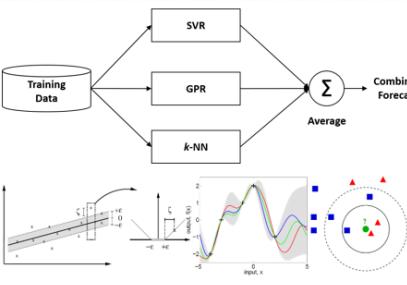


		w/o SNS	with SNS			
	Screening	Baseline	MLR	SVR	GPR	k-NN
Model R	Screening	1	1	1	1	1
	SNS		7	7	3	5
Model W ₁	Screening	2	2	2	4	2
	SNS		2	6	0	8
Model W ₂	Screening	4	3	6	5	6
	SNS		7	6	10	9

5 Combining Forecasts & Evaluation

Forecasting Combination

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



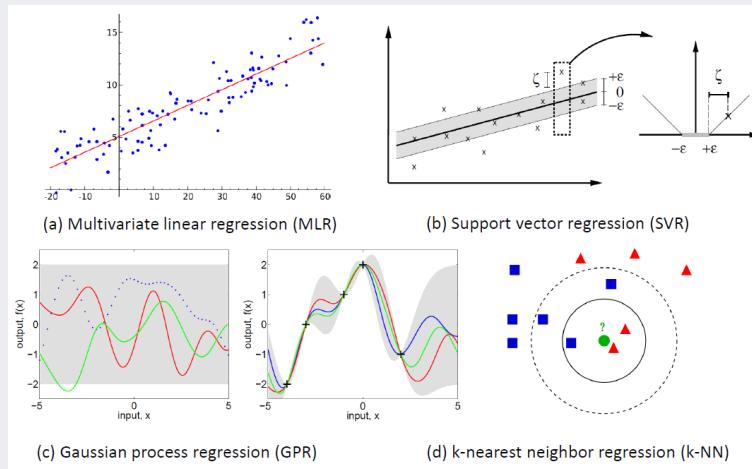
Text Mining: Applications

- 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	Box_office ¹ +Box_office ² N_seat ₃ N_seat ²
	Derived		Weekly_seat_increase	N_seat ¹ +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_mention	Cumulative_mention
		Cumulative_emotional	Cumulative_emotional	Cumulative_emotional
		Cumulative_positive	Cumulative_positive	Cumulative_positive
		Cumulative_negative	Cumulative_negative	Cumulative_negative
		Avg_mention_increase	Avg_mention_increase	Avg_mention_increase
		Weekly_mention_increase	Weekly_mention_increase	Weekly_mention_increase
		Avg_emotional_increase	Avg_emotional_increase	Avg_emotional_increase
		Weekly_emotional_increase	Weekly_emotional_increase	Weekly_emotional_increase
		Avg_positive_increase	Avg_positive_increase	Avg_positive_increase
		Weekly_positive_increase	Weekly_positive_increase	Weekly_positive_increase
		Avg_negative_increase	Avg_negative_increase	Avg_negative_increase
		Weekly_negative_increase	Weekly_negative_increase	Weekly_negative_increase

Text Mining: Applications

- 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

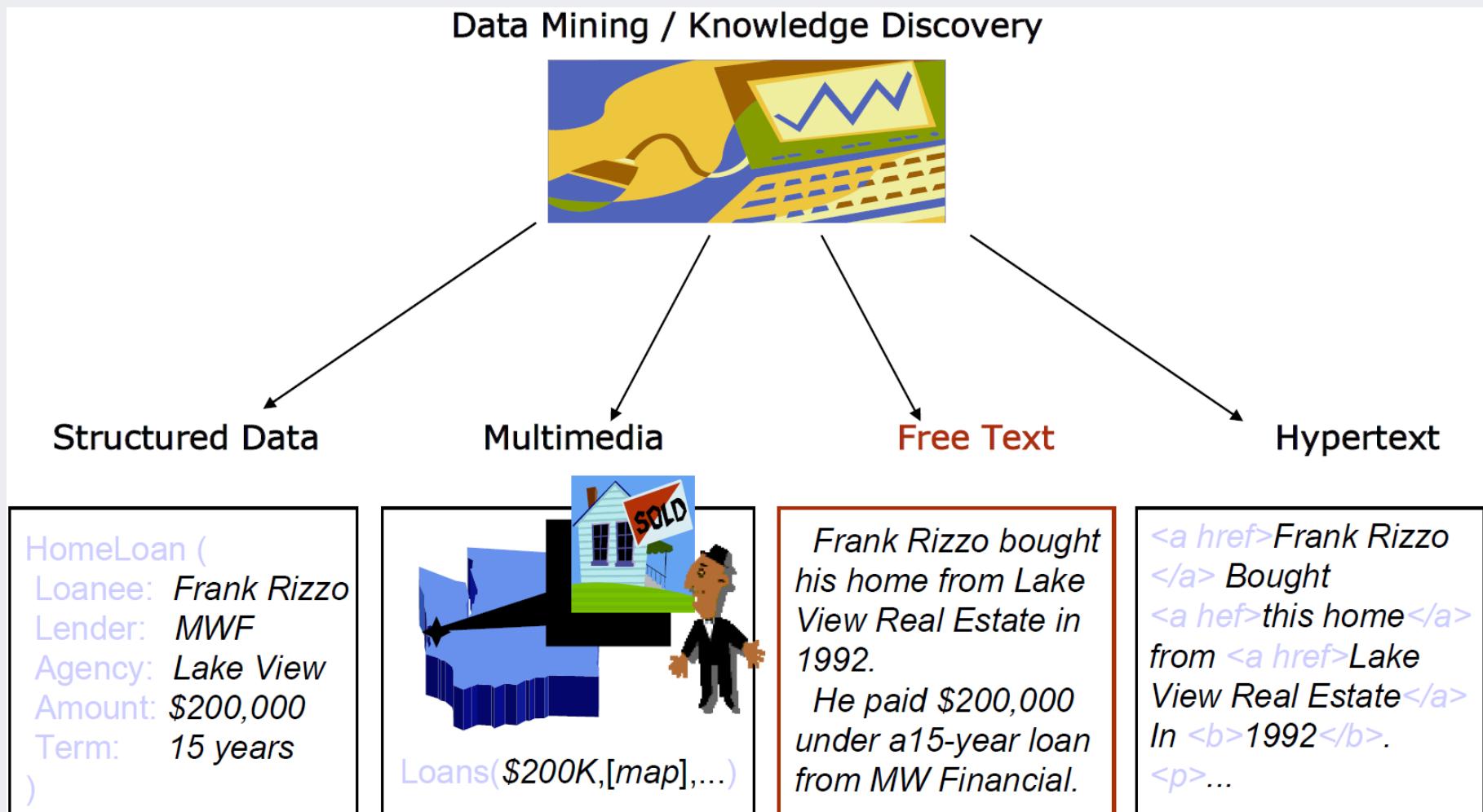


vs.



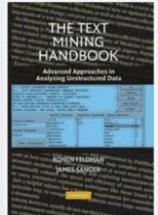
Mining Text Data

- Structure of text data



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?

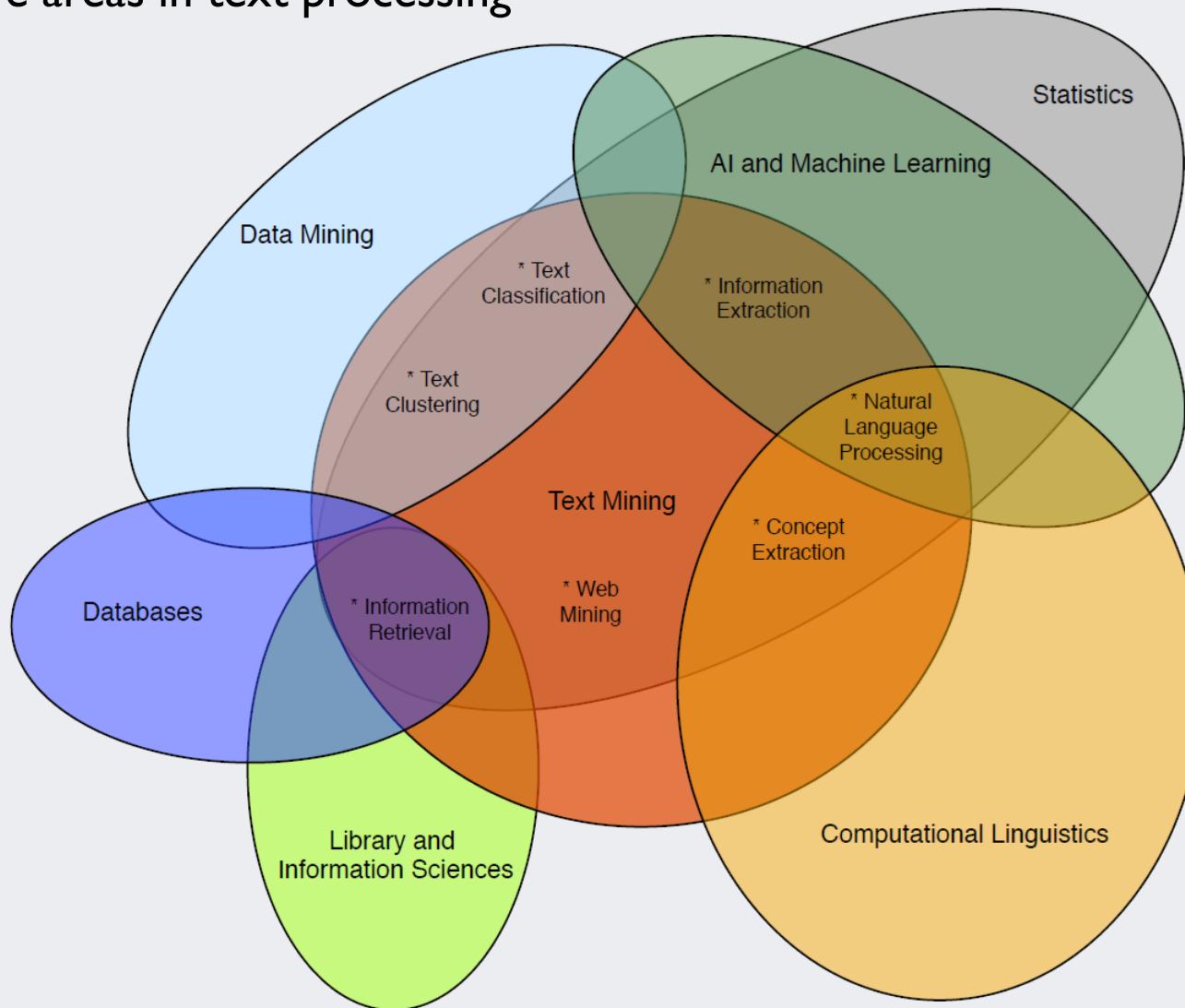
- ✓ Language itself is ambiguous

- Contexts is needed to clarify
 - Same word with different meanings, different words with same meaning
 - Misspellings, abbreviations, etc.

Mining Text Data

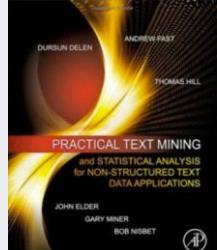
Abbott (2013)

- 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

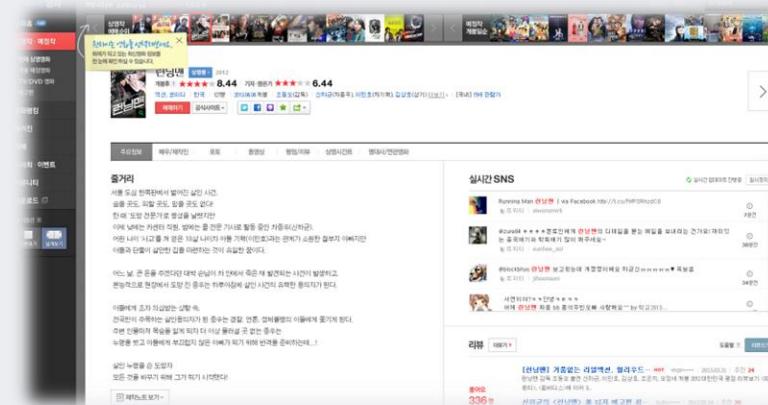
Digital library



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.

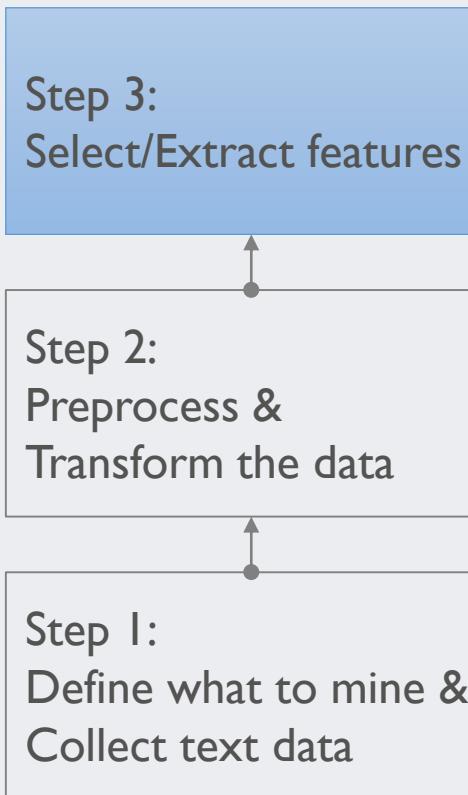


Step 2:
Preprocess &
Transform the data

Step 1:
Define what to mine &
Collect text data

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

A Simplified Process of Text Mining



Reduce the number of features

Word	S ₁	S ₂
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	S ₁	S ₂
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

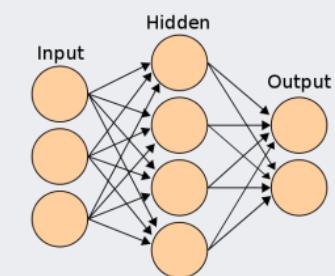
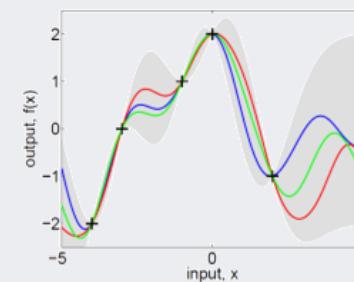
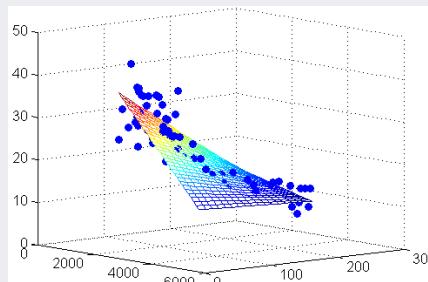
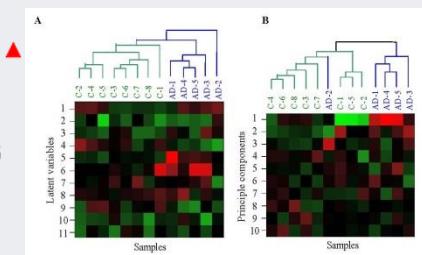
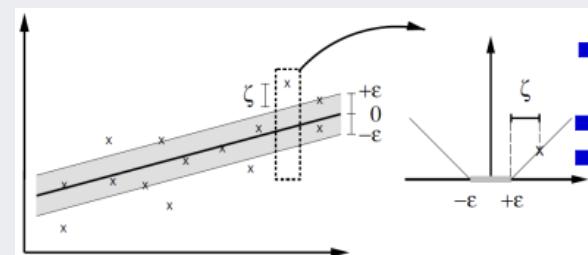
Step 3:
Select/Extract features

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 I: 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, ...)

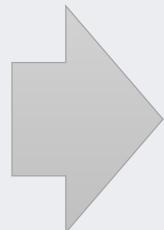
problems: not simply plain text, also hyperlinks and hidden information ("deep web")

Text Preprocessing Level 0: Text

- Remove unnecessary information from the collected data

The screenshot shows the Bits section of The New York Times website. At the top, there are several news headlines: "In Net Neutrality Push, F.C.C. Is Expected to Propose Regulating Internet Services as a...," "BITS BLOG F.C.C. Chief Wants to Override State Laws Curbing Community Net Services," "Hard-Charging Uber Tries Olive Branch," "BITS BLOG Raspberry Pi Computer Gains More Horsepower," and "DEAL Silico Vent". Below these are two large thumbnail images: one for "1% 산원주택" (1% House) and another for "1566-0495 건축설계 GO>". The main article visible is titled "Questions for IBM's Watson" by Steve Lohr, dated August 28, 2014, 3:04 PM, with 26 comments. It features a photo of W. Scott Spangler, a data scientist at IBM, demonstrating how Watson's cognitive technology can visually display connections in scientific data. The text discusses the challenges of finding the right question and the company's efforts to move Watson into the marketplace.

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



The screenshot shows a Microsoft Word document window titled "제목 없음 - 메모장". The text has been cleaned up, removing unnecessary elements. It discusses Watson's claim to fame, its work on "Jeopardy!", and its new service for scientific inquiry. The text is presented in a single column, free of figures, advertisements, or original hyperlinks.

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 I: 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”, “:-)”, “...”, “ㅋㅋㅋㅋㅋㅋㅋㅋ”?

✓ 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 I: Token

- Power distribution in word frequencies

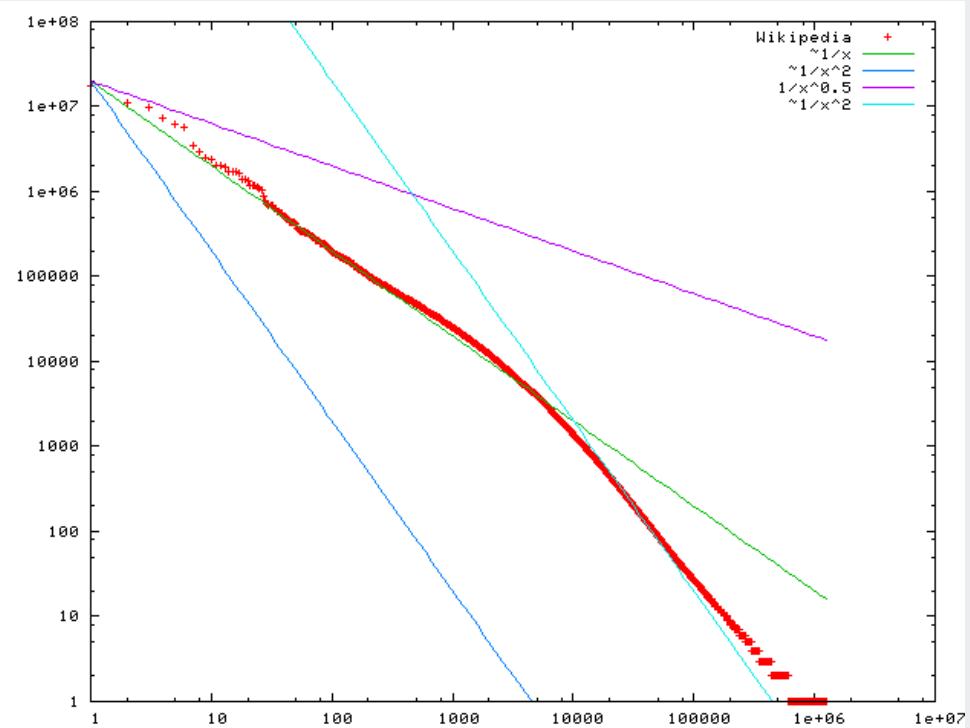
- ✓ It is **not true** that more frequently appeared words (tokens) are more important for text mining tasks.

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 I: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

Text Preprocessing Level I: 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 I: 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 I: 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	S ₁	S ₂
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



aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	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)

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

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

- Distributed Representation
 - ✓ The most simple & intuitive representation

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

- ✓ 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

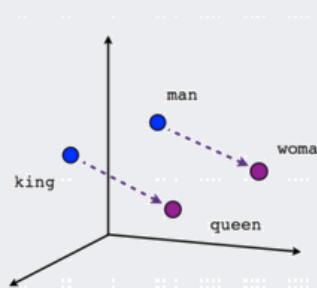
- Word vectors: distributed representation
 - ✓ A parameterized function mapping words in some language to a certain dimensional vectors

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

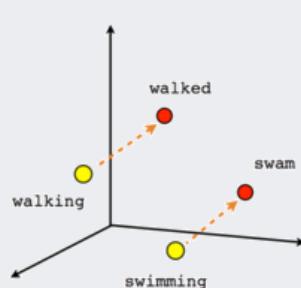
$$W(\text{"cat"}) = (0.2, -0.4, 0.7, \dots)$$

$$W(\text{"mat"}) = (0.0, 0.6, -0.1, \dots)$$

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



Male-Female



Verb tense



Country-Capital

Text Transformation

- Word vectors: distributed representation

- 0. *frog*
- 1. *frogs*
- 2. *toad*
- 3. *litoria*
- 4. *leptodactylidae*
- 5. *rana*
- 6. *lizard*
- 7. *eleutherodactylus*



3. *litoria*



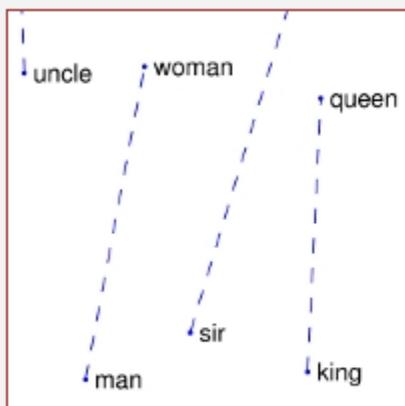
4. *leptodactylidae*



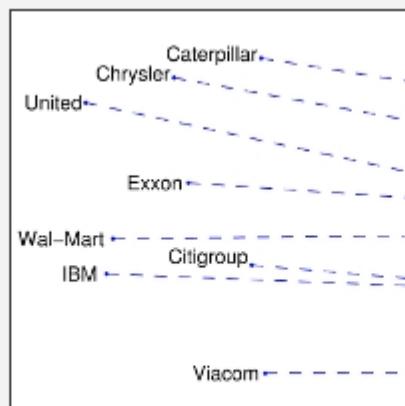
5. *rana*



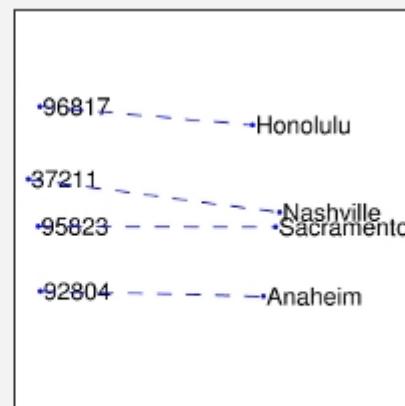
7. *eleutherodactylus*



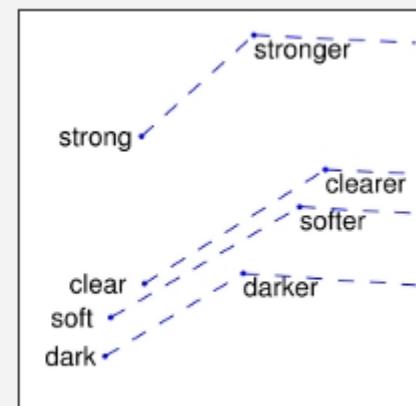
man - woman



company - ceo



city - zip code



comparative - superlative

AGENDA

- 01 Text Mining Overview
- 02 TM Process I: Collection & Preprocessing
- 03 TM Process 2: Transformation
- 04 TM Process 3: Dimensionality Reduction
- 05 TM Process 4: Learning & Evaluation

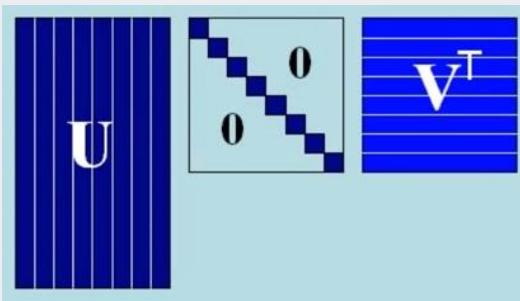
Feature Selection/Extraction

- 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:** $\sum_{F=W,\bar{W}} P(F) \sum_{C=pos,neg} P(C|F) \log \frac{P(C|F)}{P(C)}$
 - **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

$m \times r \quad r \times r \quad r \times n$

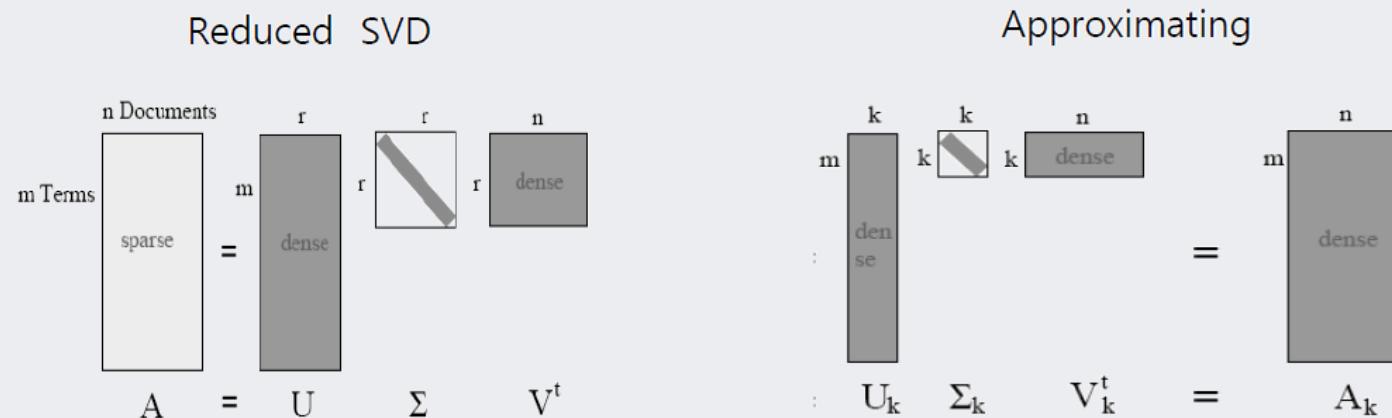


- Rectangular matrix $A(m \times n)$ can be decomposed as $A = U\Sigma V^T$
- All singular vectors of the matrices U and V are orthogonal $U^T U = V^T V = I$
- Eigenvalues of the matrix Σ is positive and sorted in a descending order

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

$$D \approx D_k = U_k \Sigma_k V_k^T$$

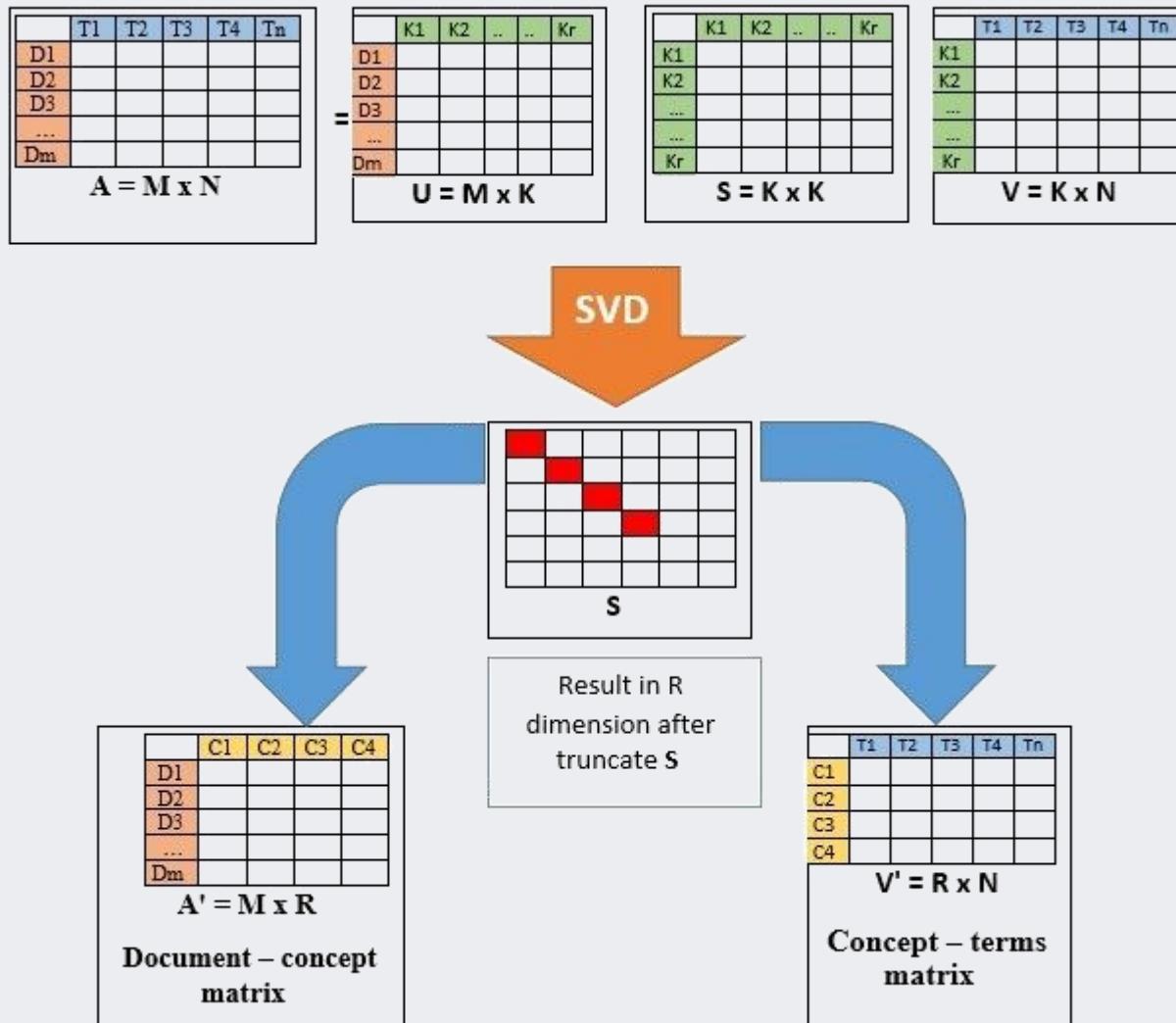
- ✓ Step 2) Multiply the transpose of the matrix U_k

$$U_k^T D_k = \Sigma_k V_k^T$$

- ✓ 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

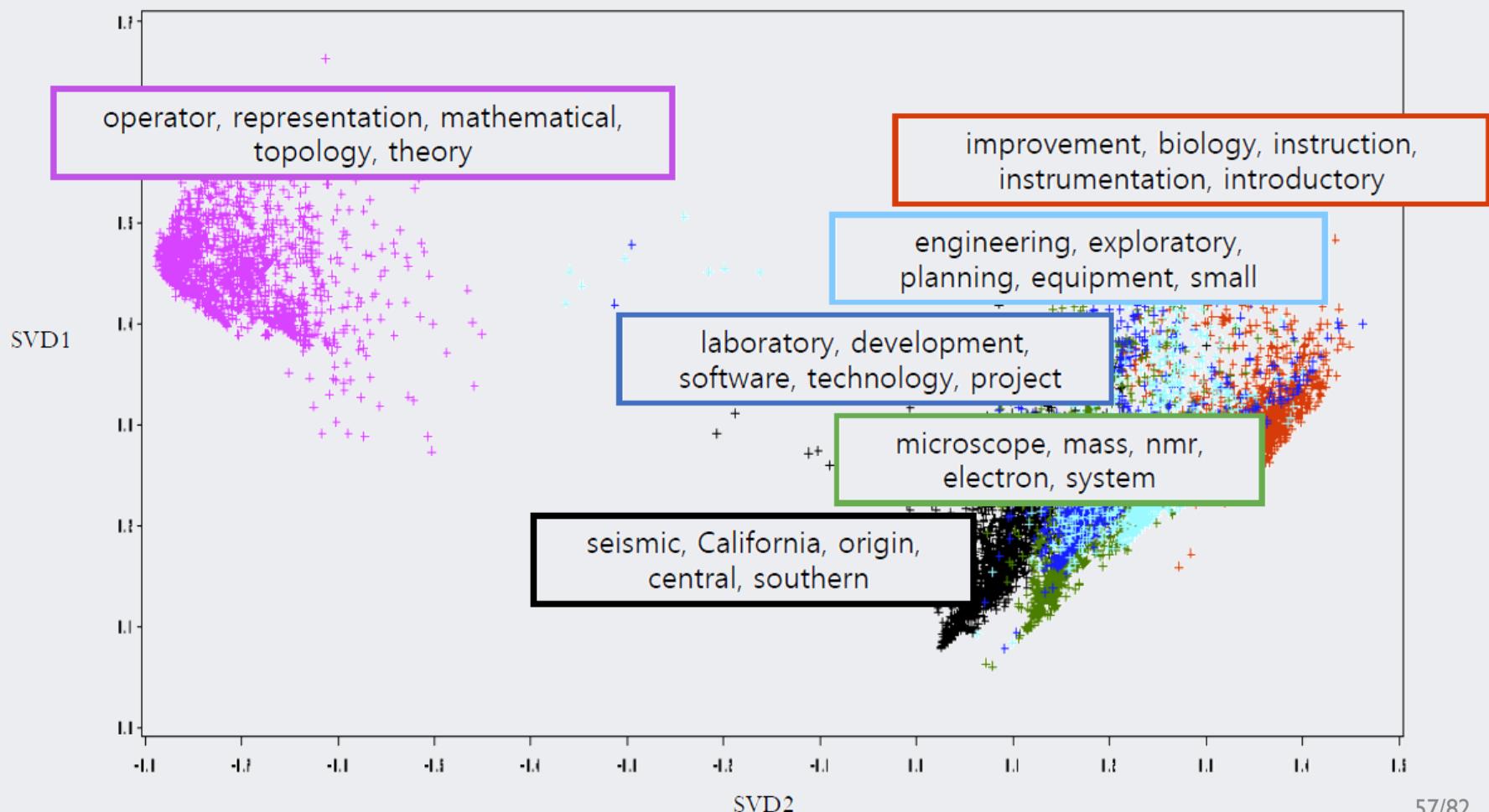
- LSA example

- ✓ 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		minority		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

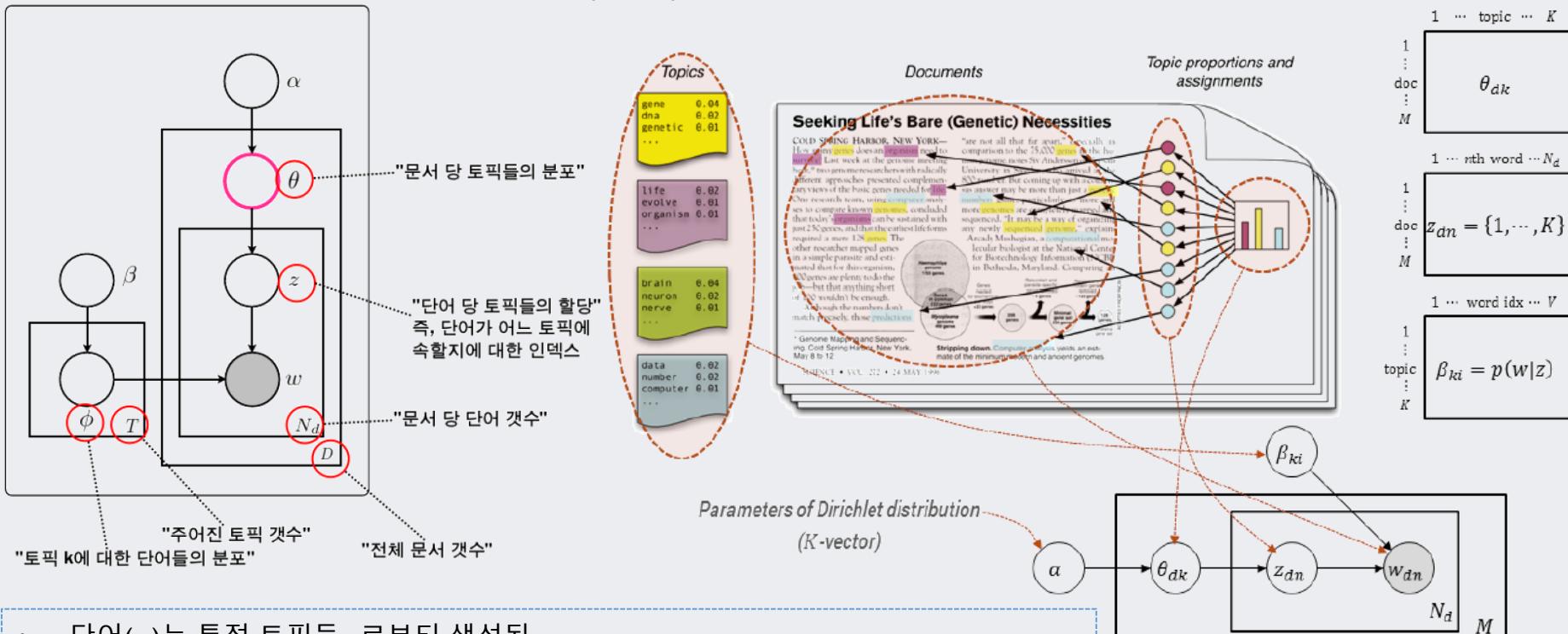
- LSA example
 - ✓ Visualize the project in the reduced 2-D space



Feature Selection/Extraction

- Topic Modeling as a Feature Extractor

- ✓ Latent Dirichlet Allocation (LDA)



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)

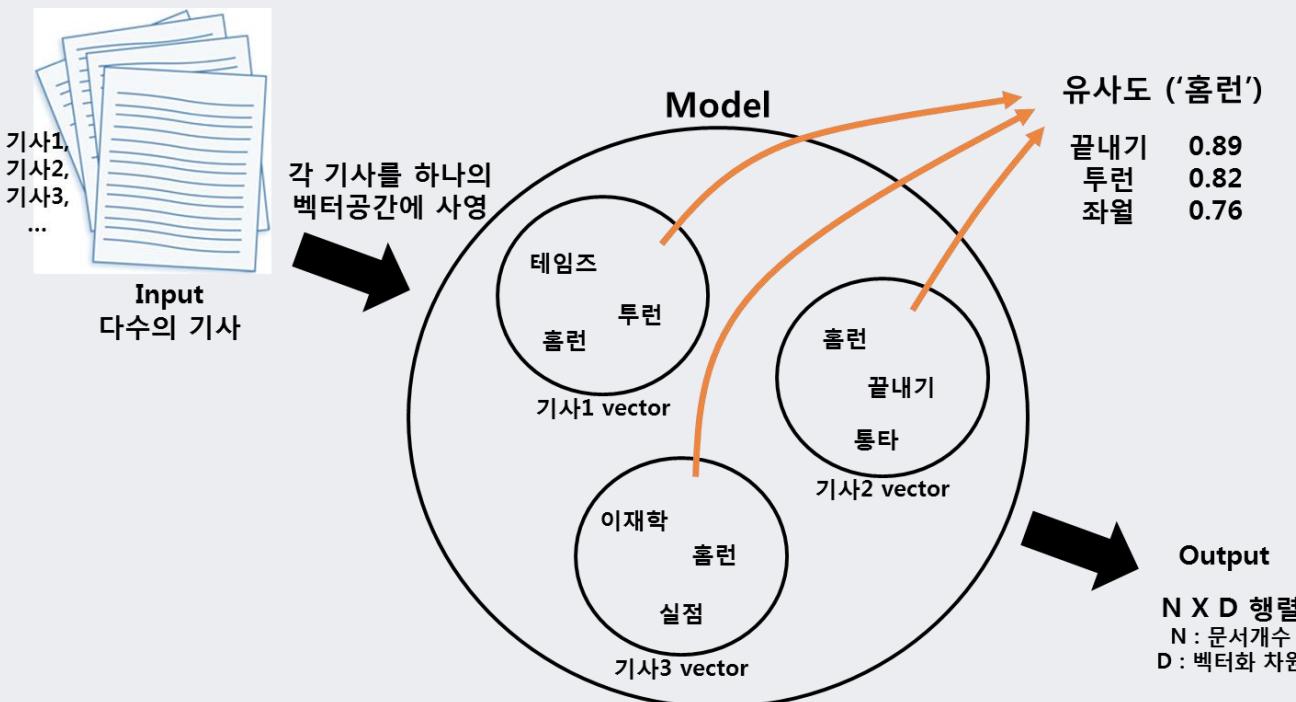
	Topic 1	Topic 2	Topic 3	...	Topic K	Sum
Doc 1	0.20	0.50	0.10	...	0.10	1
Doc 2	0.50	0.02	0.01	...	0.40	1
Doc 3	0.05	0.12	0.48	...	0.15	1
...	1
Doc N	0.14	0.25	0.33	...	0.14	1

(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

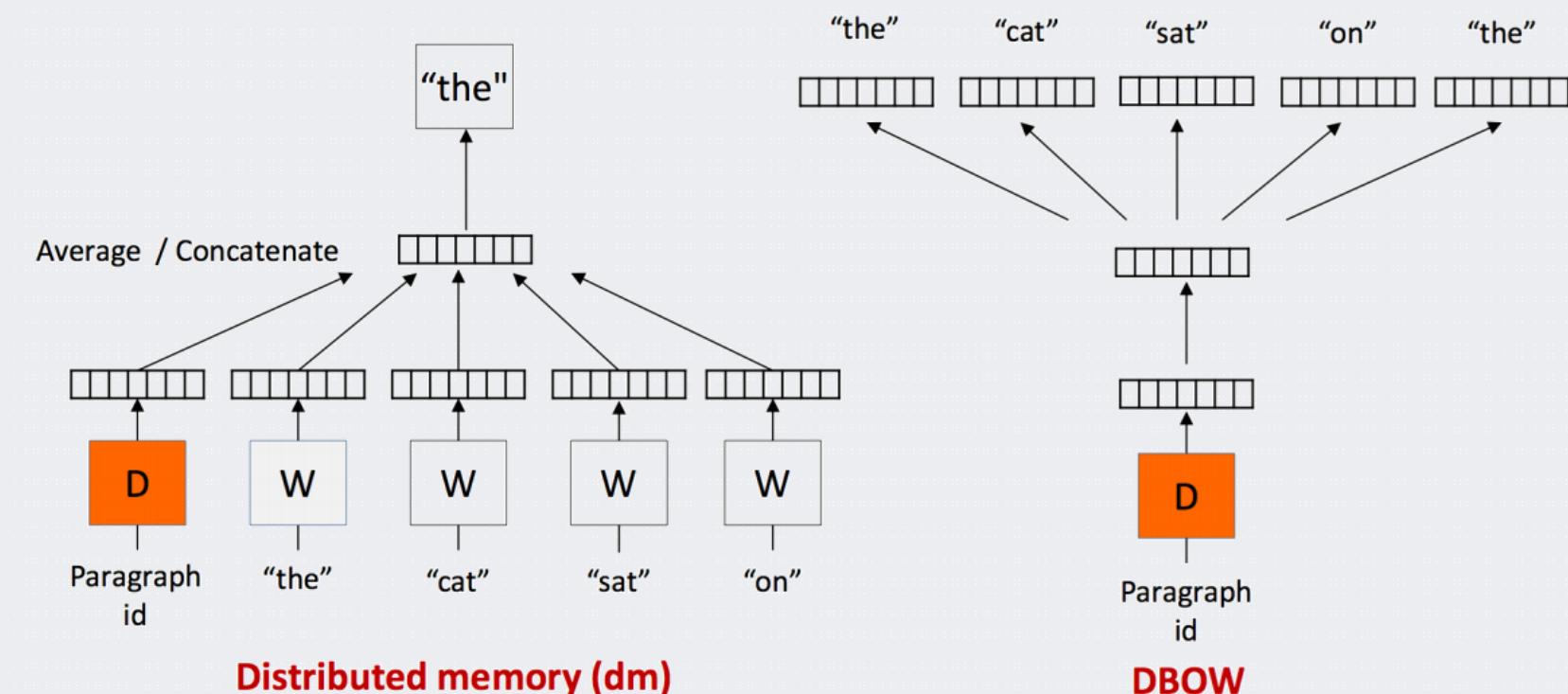
Feature Selection/Extraction

- 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)

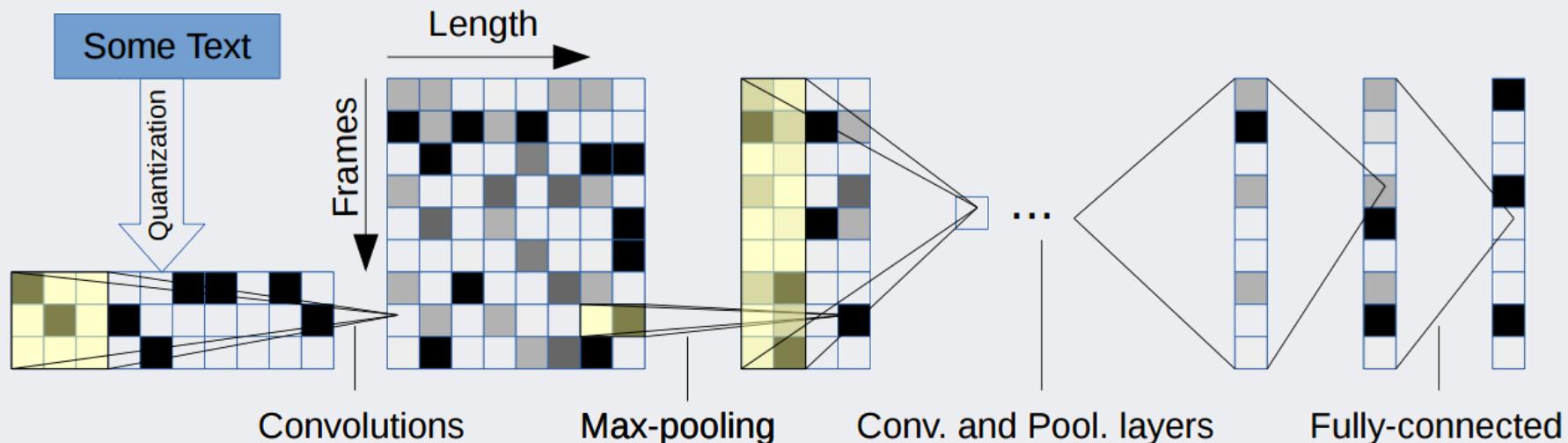
- No needs for feature engineering

- ✓ I-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

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	99.44%	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%

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

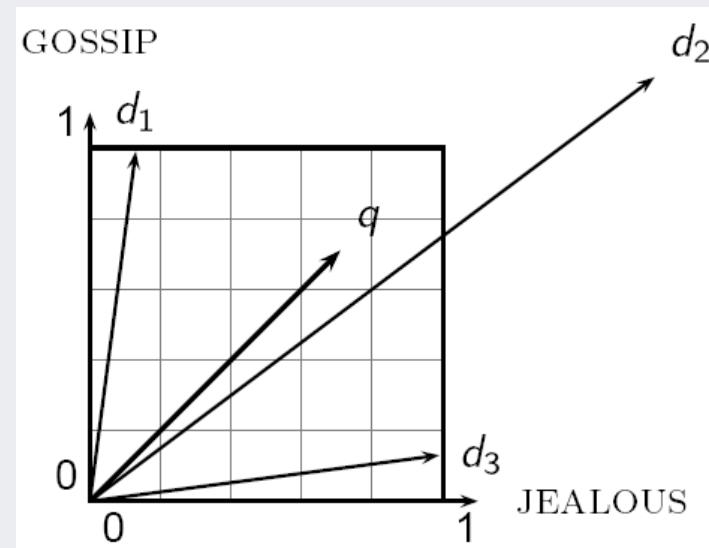
- 01 Text Mining Overview
- 02 TM Process I: Collection & Preprocessing
- 03 TM Process 2: Transformation
- 04 TM Process 3: Dimensionality Reduction
- 05 TM Process 4: Learning & Evaluation

Similarity Between Documents

- Document similarity
 - ✓ Use the [cosine similarity](#) rather than Euclidean distance
 - Which two documents are more similar?

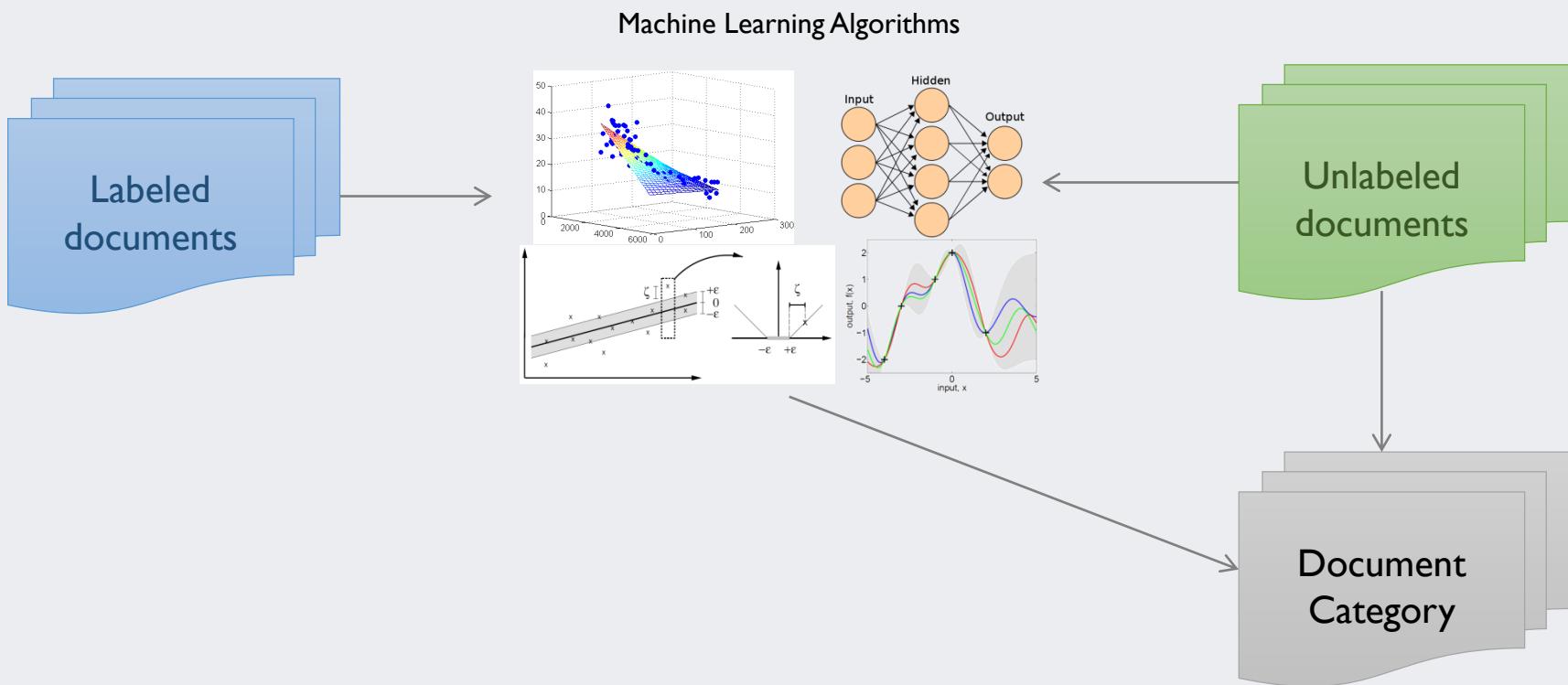
Doc.	Word 1	Word 2	Word 3
Document 1	1	1	1
Document 2	3	3	3
Document 3	0	2	0

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



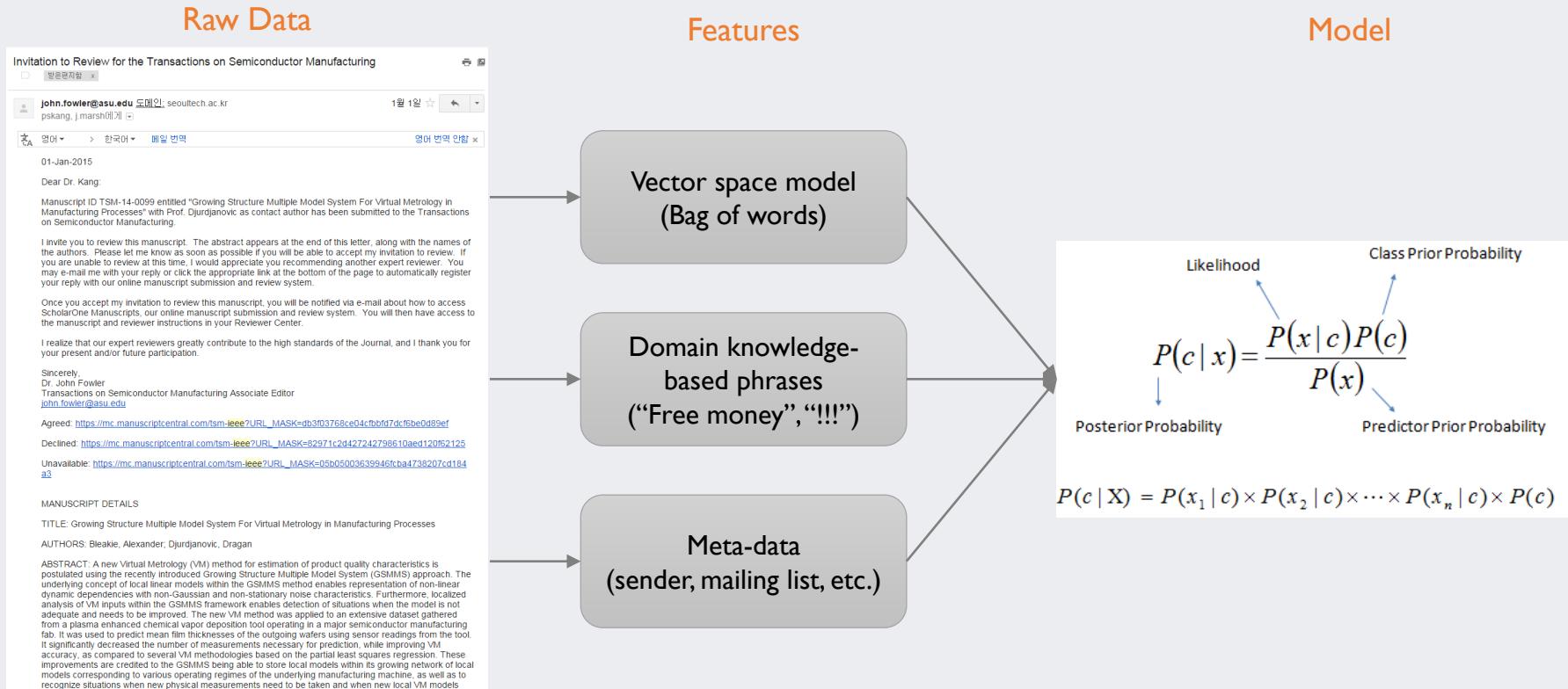
Learning Task I: Classification

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



Learning Task I: Classification

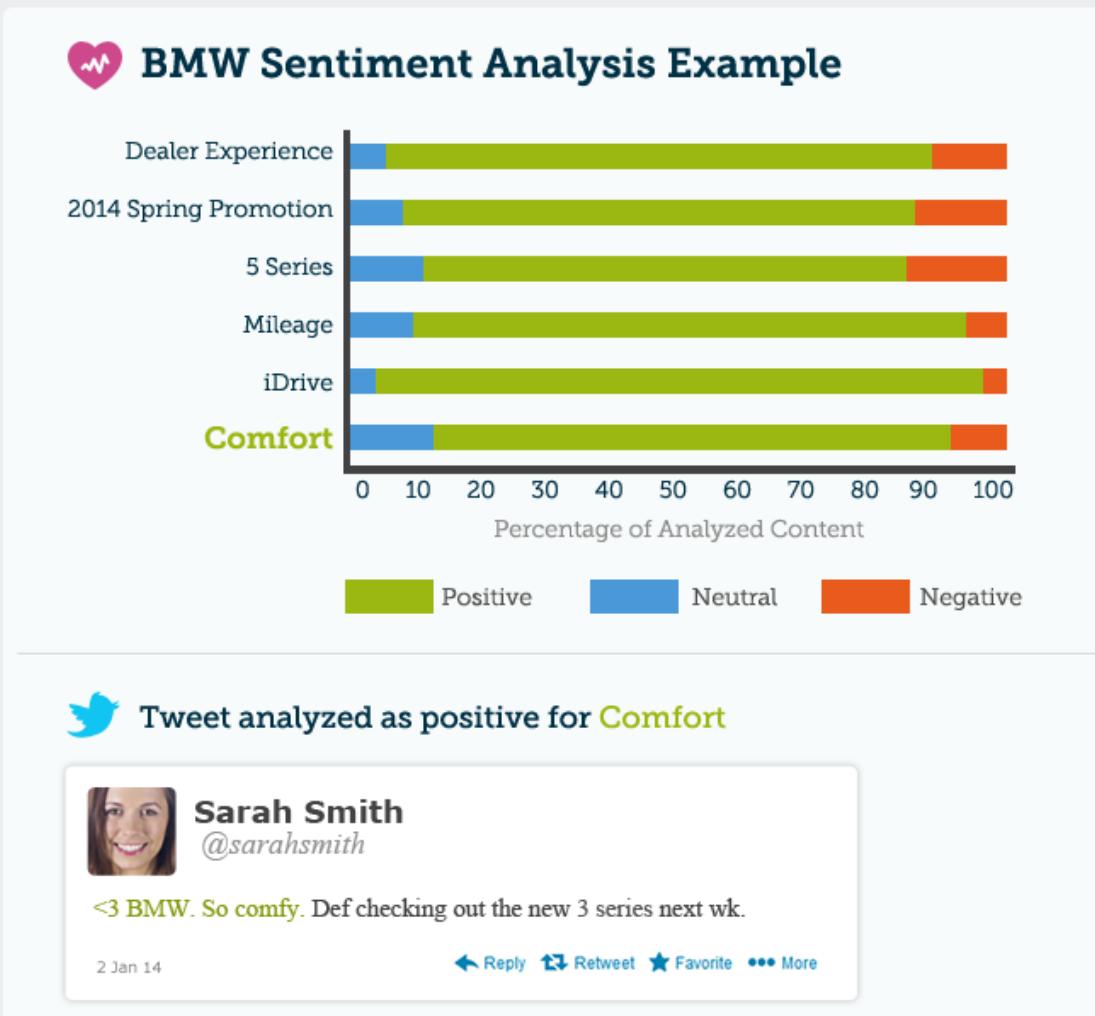
- Spam filtering



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 I: Classification

- Sentiment Analysis



Learning Task I: Classification

Socher et al. (2013)

- Sentiment Analysis

✓ Sentiment tree bank @Stanford NLP Lab

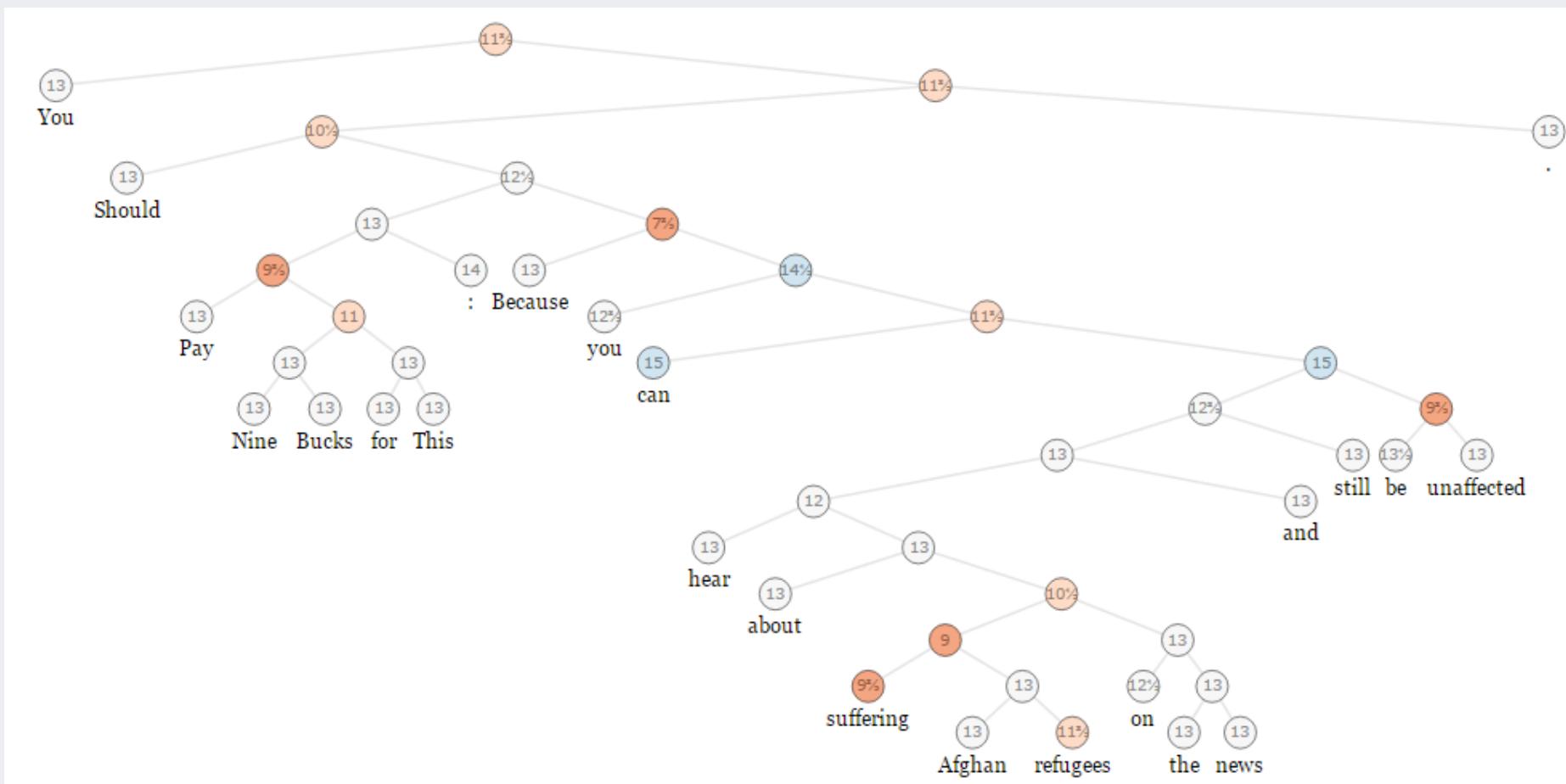


Learning Task I: Classification

Socher et al. (2013)

- Sentiment Analysis

- ✓ Sentiment tree bank @Stanford NLP Lab



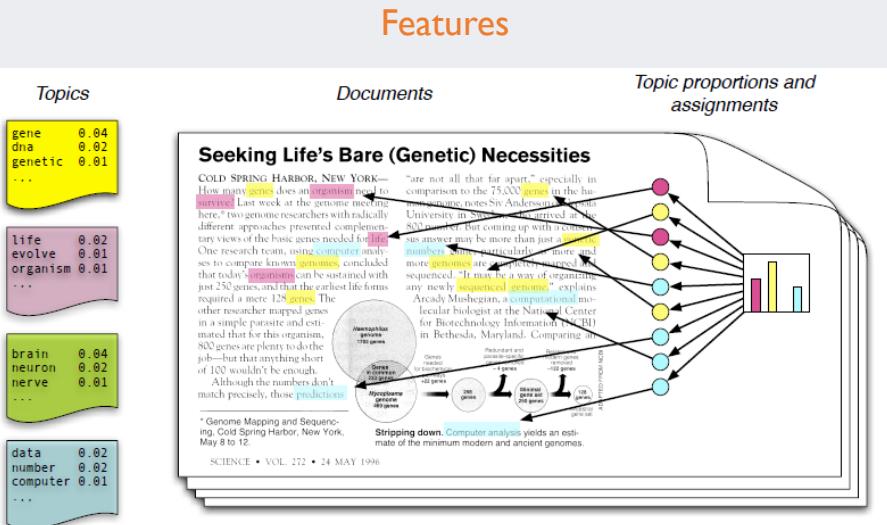
Learning Task 2: Clustering

- 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	Abstract
IEEE Transactions on Engineering Management	Impact of interorganizational relationships on organizational performance	2014	Interorganizational relationships are an integral part of an organization's decision-making process. We explore the boundary conditions of Christensen's theory of disruption. Using a two-stage procurement process, we find that firms that outsource their procurement processes face a challenge in maintaining quality and cost efficiency.
IEEE Transactions on Engineering Management	Exploring the boundary conditions of disruption	2014	We explore the boundary conditions of Christensen's theory of disruption. Using a two-stage procurement process, we find that firms that outsource their procurement processes face a challenge in maintaining quality and cost efficiency.
IEEE Transactions on Engineering Management	Two stage procurement processes with c	2014	This paper considers a sourcing problem faced by a manufacturer who outsources its procurement processes. The manufacturer wants to maintain quality and cost efficiency while outsourcing. We propose a two-stage procurement process to achieve this goal.
IEEE Transactions on Engineering Management	Identifying emerging customer requirements	2014	Rapid changes of new technologies, market dynamics, and swift fluctuation of customer requirements pose significant challenges to organizations. This study identifies emerging customer requirements and proposes a methodology for capturing them.
IEEE Transactions on Engineering Management	Supply chain quality integration: Antecedents and consequences	2014	This study extends quality management from an individual company perspective to supply chain quality integration. Antecedents of quality management in supply chains are identified, and the consequences of quality management on supply chain performance are explored.
IEEE Transactions on Engineering Management	Achieving IT-enabled enterprise agility	2014	An ongoing challenge that enterprises currently face is that of fostering agile mindsets. This study elucidates how a firm accesses external technological knowledge (TK) to achieve IT-enabled enterprise agility. The findings suggest that IT-enabled enterprise agility is achieved through the use of IT to facilitate knowledge sharing and collaboration across different departments.
IEEE Transactions on Engineering Management	Accessing external technological knowledge	2014	This study elucidates how a firm accesses external technological knowledge (TK) to achieve IT-enabled enterprise agility. The findings suggest that IT-enabled enterprise agility is achieved through the use of IT to facilitate knowledge sharing and collaboration across different departments.
IEEE Transactions on Engineering Management	An investigation on centralized and decentralized design for supply chain	2014	Design for supply chain has become an essential consideration while designing a product. This study presents an investigation of the relationship between internalization of design and product design.
IEEE Transactions on Engineering Management	ISO 9000 internalization and organizational design	2014	This study presents an investigation of the relationship between internalization of design and product design.
IEEE Transactions on Engineering Management	Knowledge specialization in ph.D.	2014	Researchers have argued that specialization within groups yields productivity gains. This study explores the relationship between knowledge specialization and productivity gains.
IEEE Transactions on Engineering Management	Overlap-based design structure	2014	To gain competitive leverage, firms that design and develop complex products see the need to overlap-based design structure. This study presents a methodology for developing a complex product using an overlap-based design structure.
IEEE Transactions on Engineering Management	The TOC-based algorithm for solving M	2014	In a previous paper in this journal named "The TOC-Based Algorithm for Solving M
IEEE Transactions on Engineering Management	Toward environmental and quality sustainability	2014	The last decade, industrial research for sustainable development and environmental protection has been a major focus of research.
IEEE Transactions on Engineering Management	Balancing exploration and exploitation	2014	This study investigates the influence of supply chain portfolios on the paradox of exploration and exploitation.
IEEE Transactions on Engineering Management	Managing nuclear spare parts inventory	2014	This paper presents a methodology for developing a spare parts inventory management system.
IEEE Transactions on Engineering Management	Scenario-based modeling of Interdepenc	2014	This paper develops a scenario generation framework for modeling demand and supply uncertainty.
IEEE Transactions on Engineering Management	Governing the portfolio management pr	2014	Past strategy and innovation research has basically ignored potential effects of portfolio management.
IEEE Transactions on Engineering Management	An efficacious operation is safe operat	2014	One of the most fundamental indicators of a facility's social sustainability is the health of its employees.
IEEE Transactions on Engineering Management	Managing highly innovative projects: Th	2014	The climate change debate and economic recovery strategies in various industries require innovative projects. Managing such projects requires a different approach.
IEEE Transactions on Engineering Management	Improved mean and variance estimating	2014	The program evaluation and review technique (PERT) is a popular method for mean and variance estimation. This study proposes a new method for mean and variance estimation.
IEEE Transactions on Engineering Management	Openness and appropriation: Empirical e	2014	The adoption of open innovation creates a dilemma for firms. On one hand, a company can benefit from open innovation, but on the other hand, it can also lead to intellectual property disputes.
IEEE Transactions on Engineering Management	The redundancy queuing-location-alloca	2014	Redundancy queuing-location-allocation problems (RQLAPs) involve the economic allocation of resources to multiple locations.
IEEE Transactions on Engineering Management	Response to buyout options in Internet	2014	A buyout option allows a bidder to acquire an auctioned product immediately at a discounted price.
IEEE Transactions on Engineering Management	Contracting a development supplier in th	2014	The development of important components for a new product is frequently outsourced to third-party suppliers.
IEEE Transactions on Engineering Management	Exploration, exploitation, and growth in	2014	This study investigates how firm age and environmental adversity influence exploration, exploitation, and growth.
IEEE Transactions on Engineering Management	Innovation and performance: The role of	2014	Innovation has become the cornerstone for achieving competitive advantage and improving performance.
IEEE Transactions on Engineering Management	Spotting lemons in platform markets: A	2014	This study addresses the understudied question of how content integrity for digital platforms is affected by user behavior.
IEEE Transactions on Engineering Management	Business models and tactics in new pro	2014	Effectuation and causation are two contrasting approaches to new business development.
IEEE Transactions on Engineering Management	Technology change and capacity	2014	This paper explores the role of learning in managing the capacities of existing and future technologies.



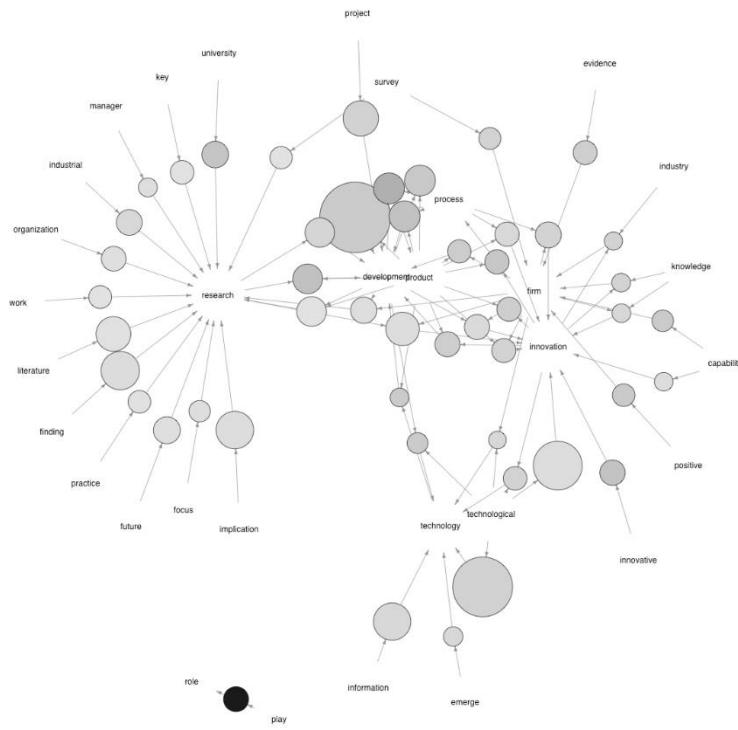
- 8,850 articles from 11 Journals for the recent 10 years
- 21,434 terms after preprocessing

- 50 topics from Latent Dirichlet Allocation (LDA)

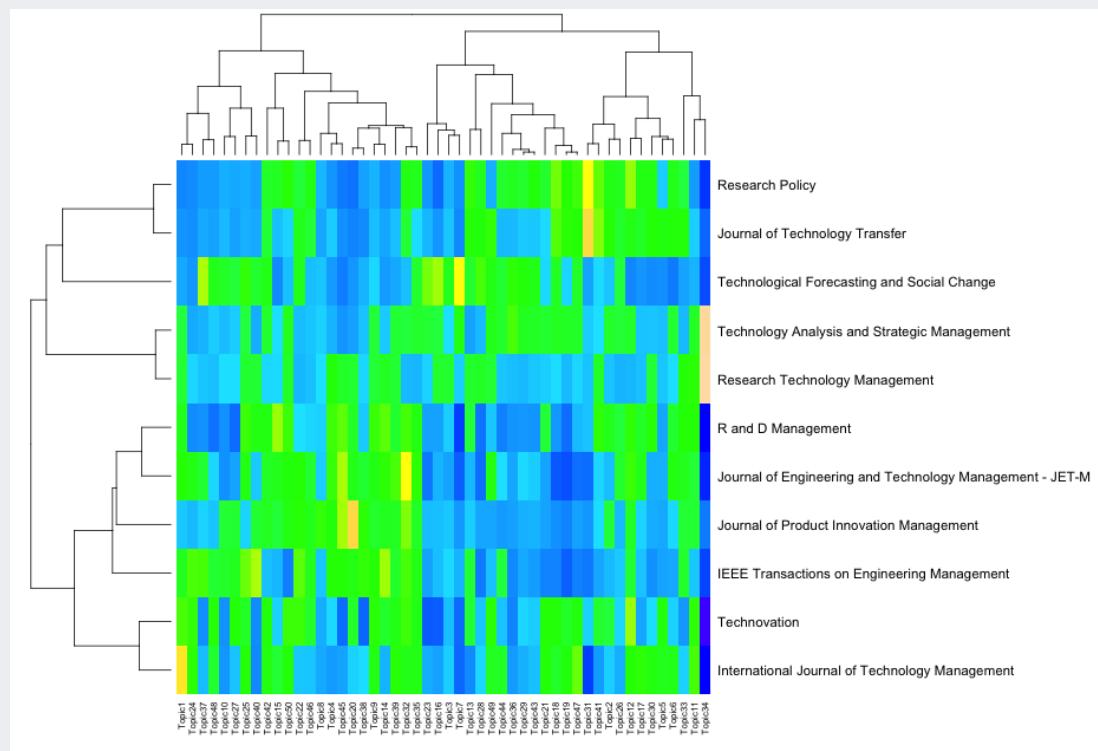
Learning Task 2: Clustering

- Document Clustering & Visualization

Keywords association



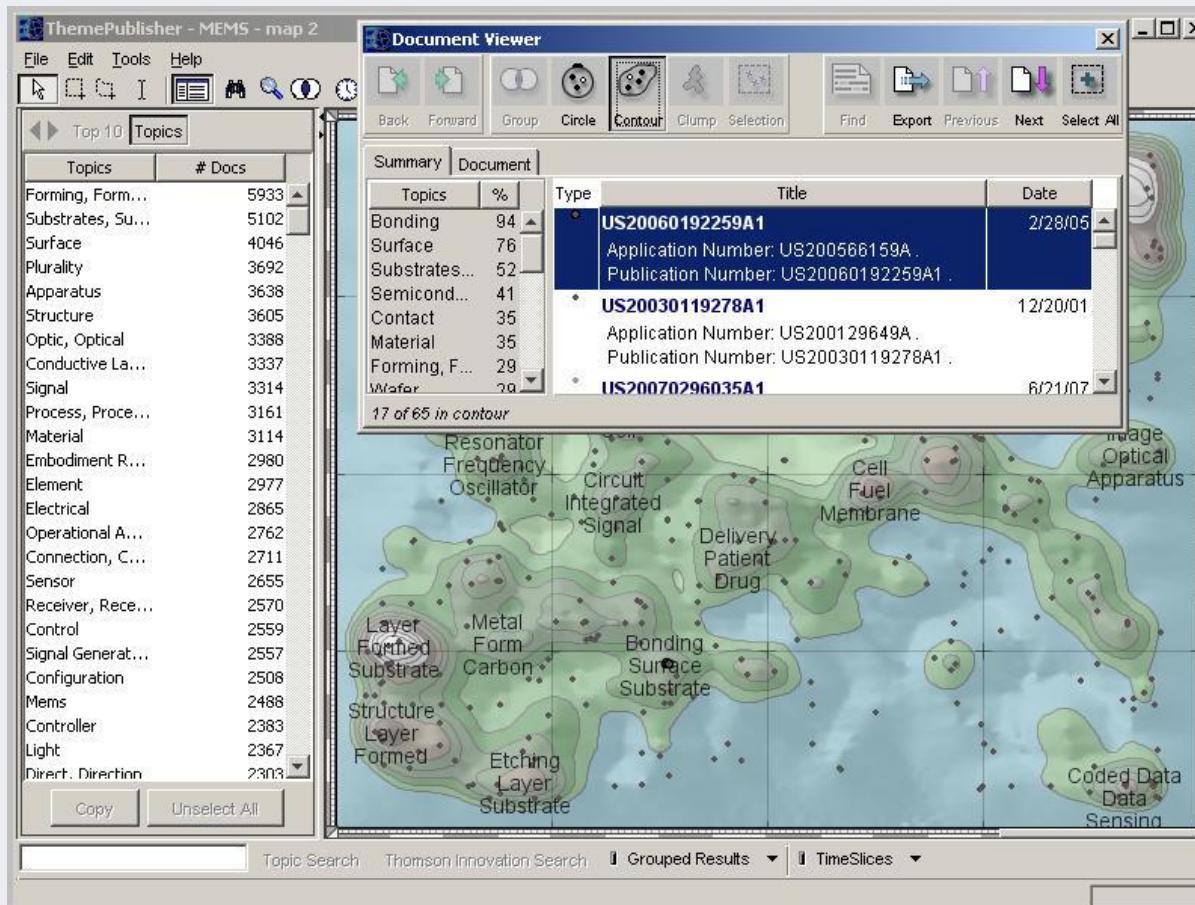
Journal/Topic Clustering



Learning Task 2: Clustering

- Document Clustering & Visualization

- ✓ Themescape: contents maps from Thomson innovation full text patent data



Learning Task 3: Information Extraction/Retrieval

- 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
Headache	Cinnamomi Cortex Spissus	1.2	23.5	1.8
	Moutan Cortex	1.0	20.6	3.6
	Schizandreae Fructus	1.0	20.6	3.4

Learning Task 3: Information Extraction/Retrieval

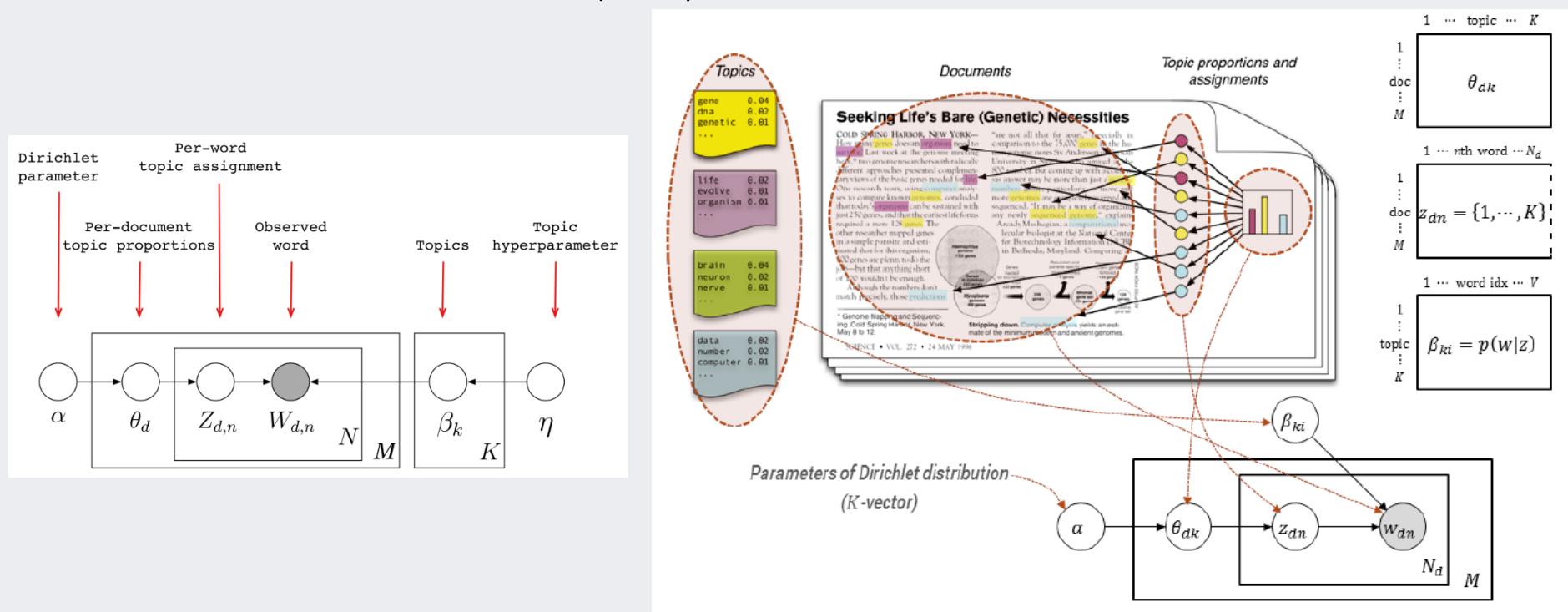
Topic Modeling

- ✓ 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

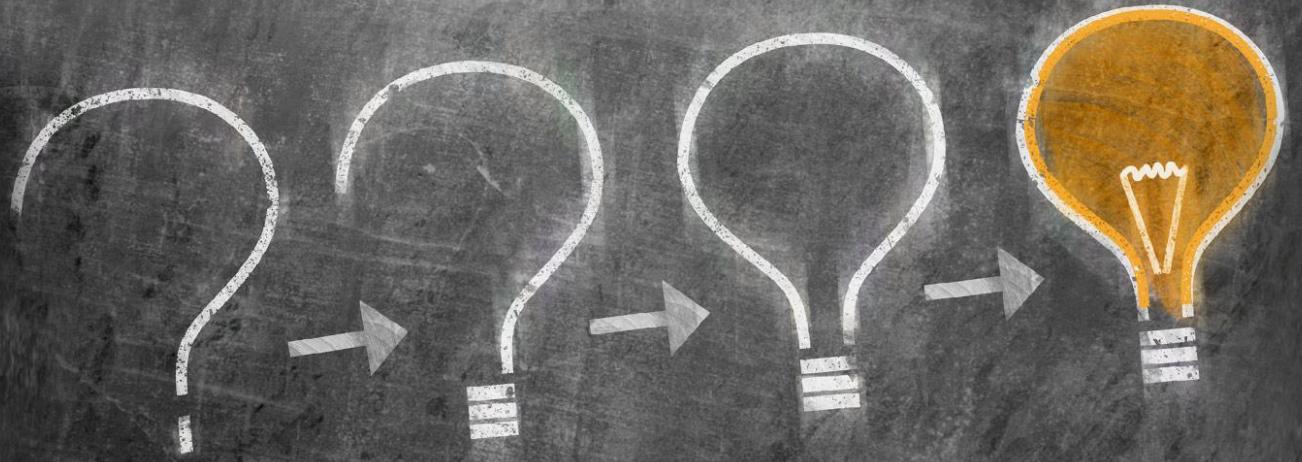


Learning Task 3: Information Extraction/Retrieval

- 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 $W; \theta, z, \Phi$ are latent variables (hidden, cannot be observed)
- Document generation process: (1) Estimate the topic proportions θ , (2) estimate the word probability w from θ



References

Research Papers

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.
- Lee, H. and Kang, P. (2015). Research trends analysis in Technology Management based on Topic Modeling, working paper.
- 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).
- Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

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>