

Introduction to Text Analytics

Pilsung Kang

School of Industrial Management Engineering
Korea University

AGENDA

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

Text Analytics: 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 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



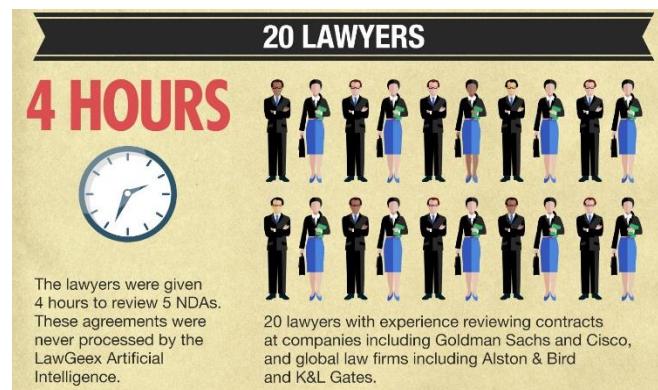
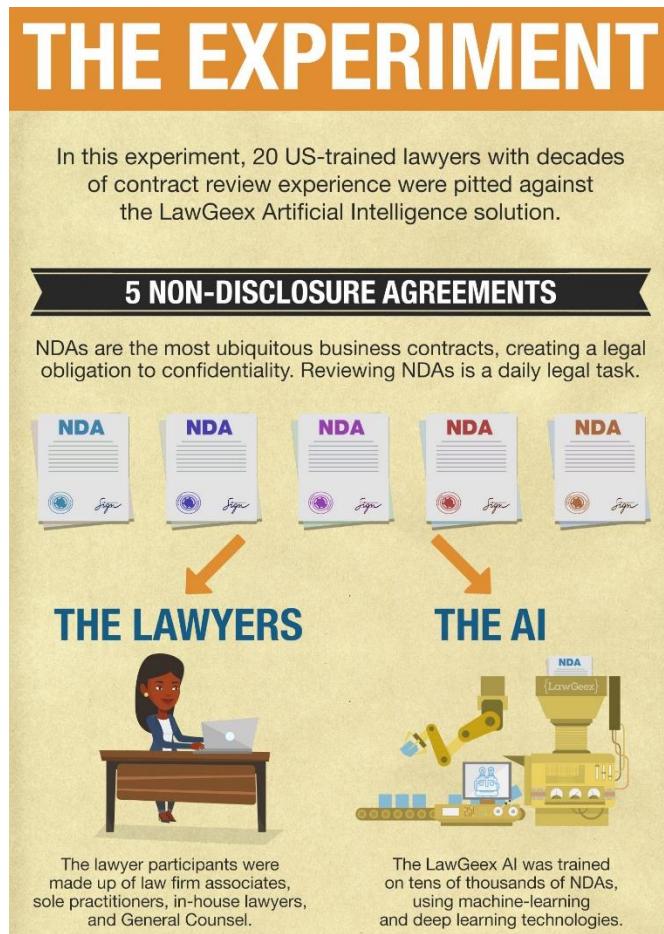
This raises some issues for organisations that need to process and exploit unstructured data. Some big data tools, primarily those based on Hadoop, are designed from the ground up to manage and analyse unstructured information. Other, more conventional business intelligence (BI) and data warehousing 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>

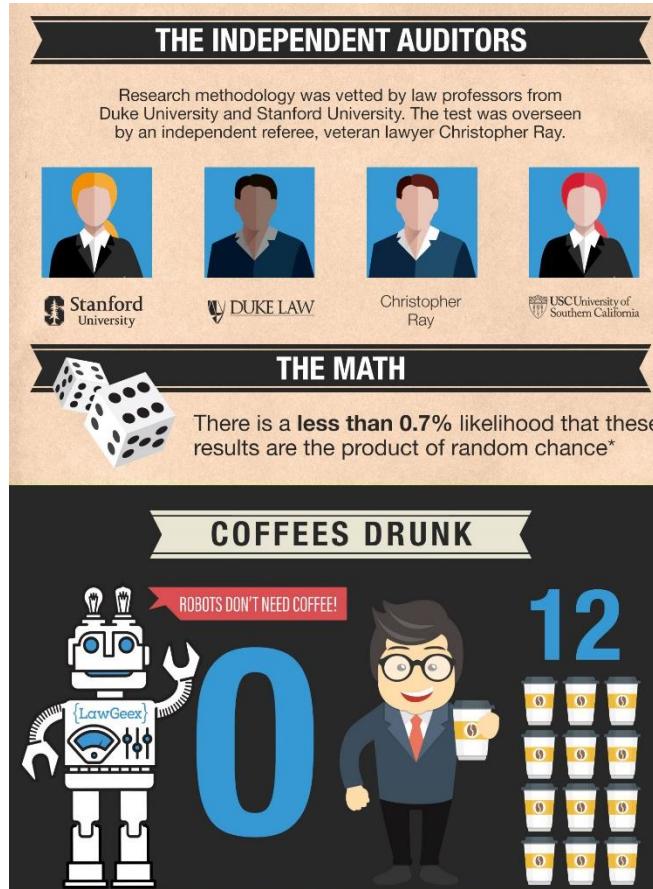
Text Analytics: Background

- AI vs. Lawyers: The ultimate showdown
 - ✓ Task: to spot issues in five Non-Disclosure Agreements (NDAs)



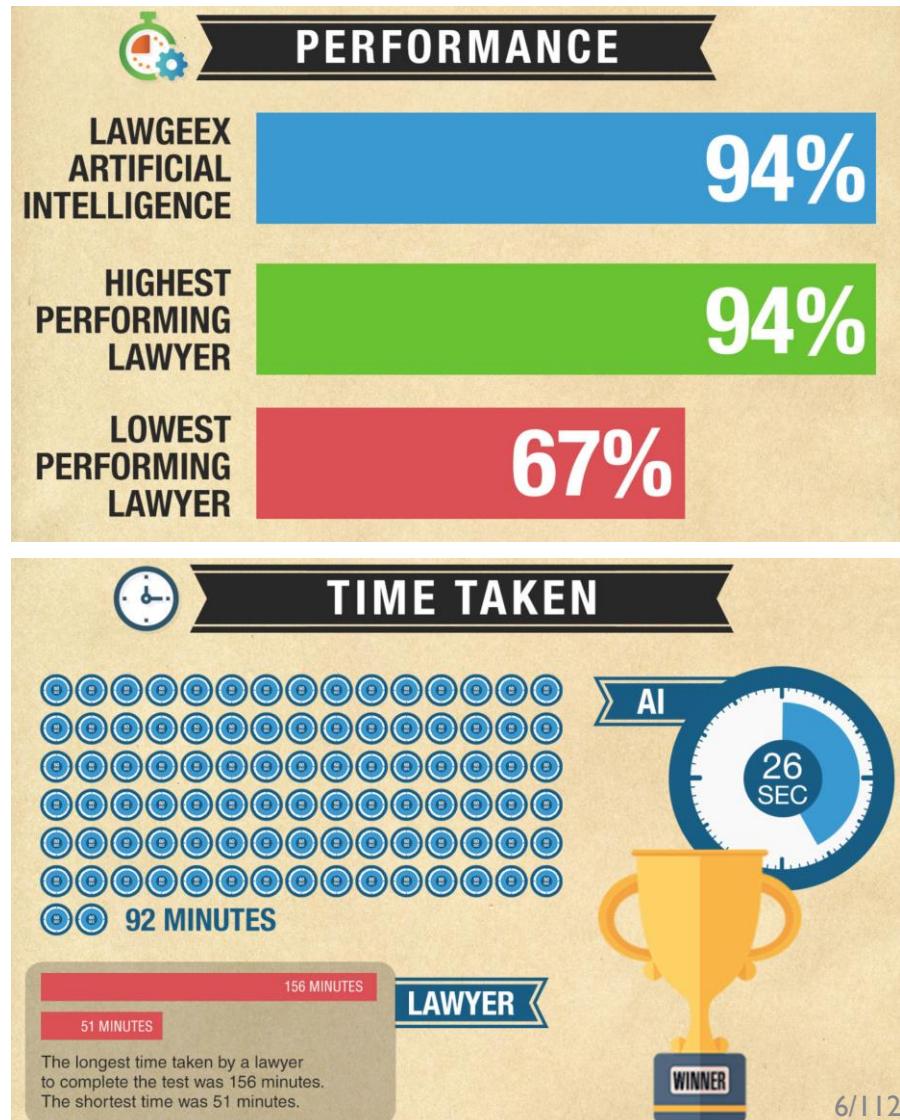
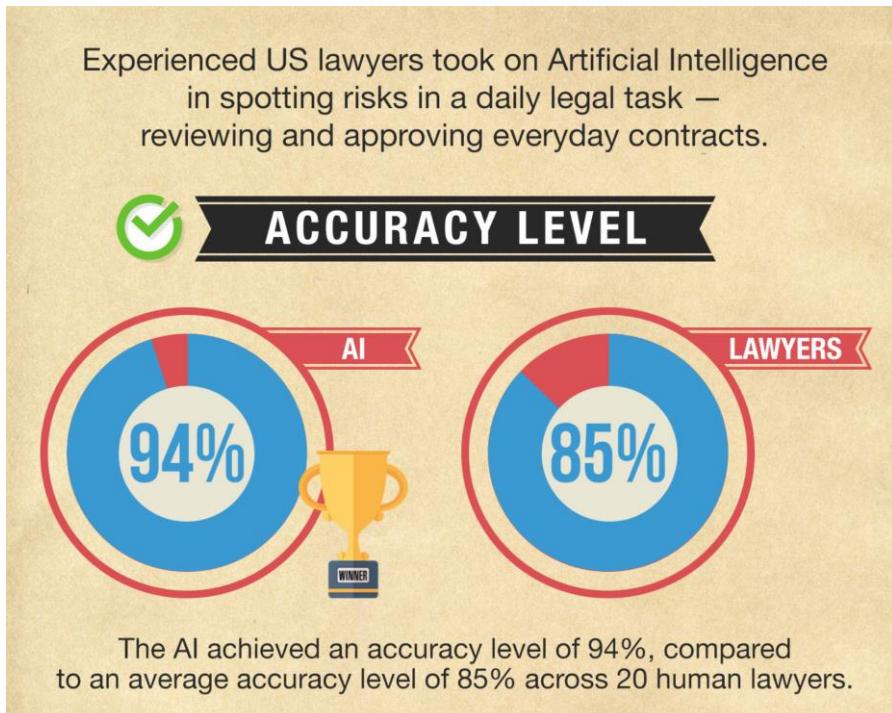
Text Analytics: Background

- AI vs. Lawyers: The ultimate showdown



Text Analytics: Background

- AI vs. Lawyers: The ultimate showdown



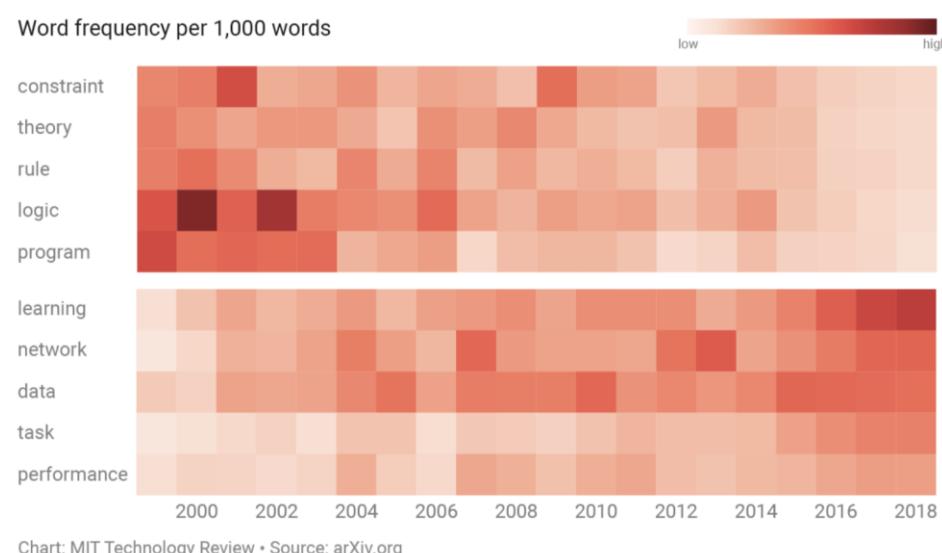
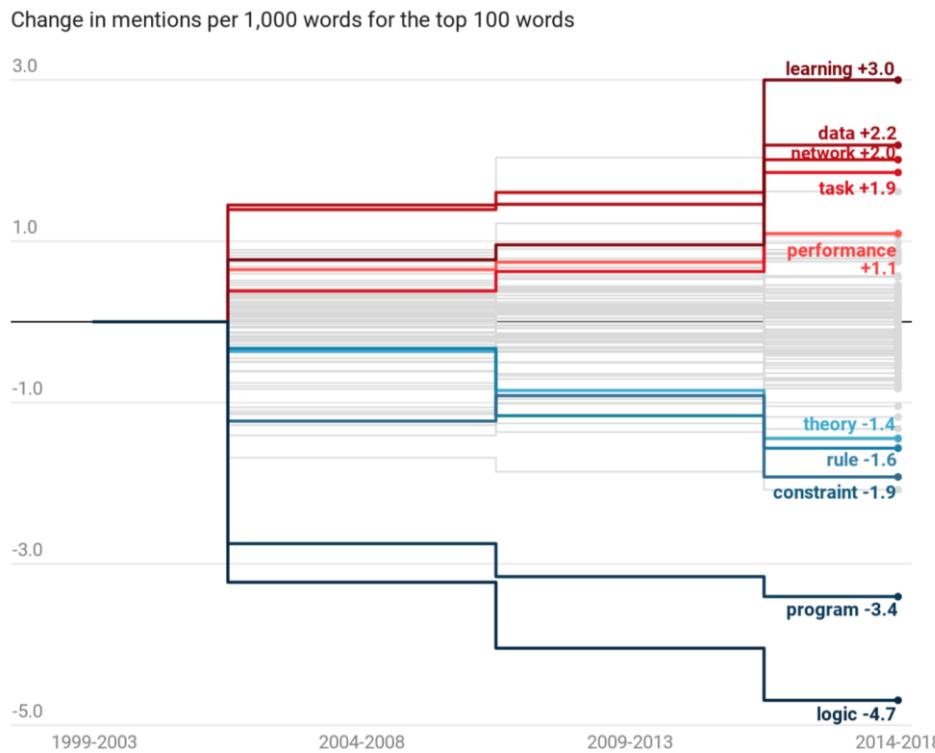
Example: AI papers in arXiv

- The number of papers in the “artificial intelligence” section
- ✓ Can you read them all?



Example: AI papers in arXiv

- Let's do some text mining!
- ✓ Actually, it was just a simple word frequency analysis
- Discovery I: Machine learning eclipses knowledge-based reasoning

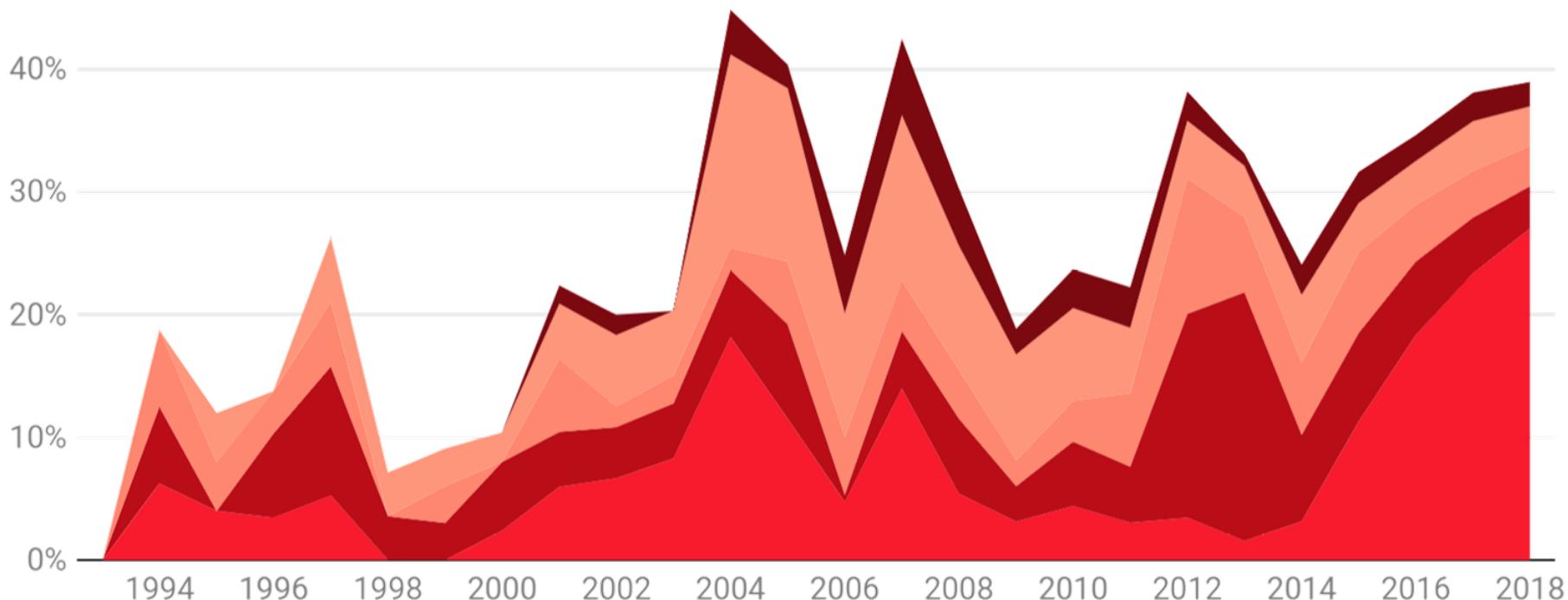


Example: AI papers in arXiv

- Discovery 2: The Neural-Network Boom

Percentage of papers that mention each method

neural networks bayesian networks markov methods evolutionary algorithms
support vector machines



Hover over the chart areas to see their labels.

Chart: MIT Technology Review • Source: arXiv.org • Created with Datawrapper

Example: AI papers in arXiv

- Discovery 3: The rise of reinforcement learning

Share of papers that mention it compared to any type of machine learning

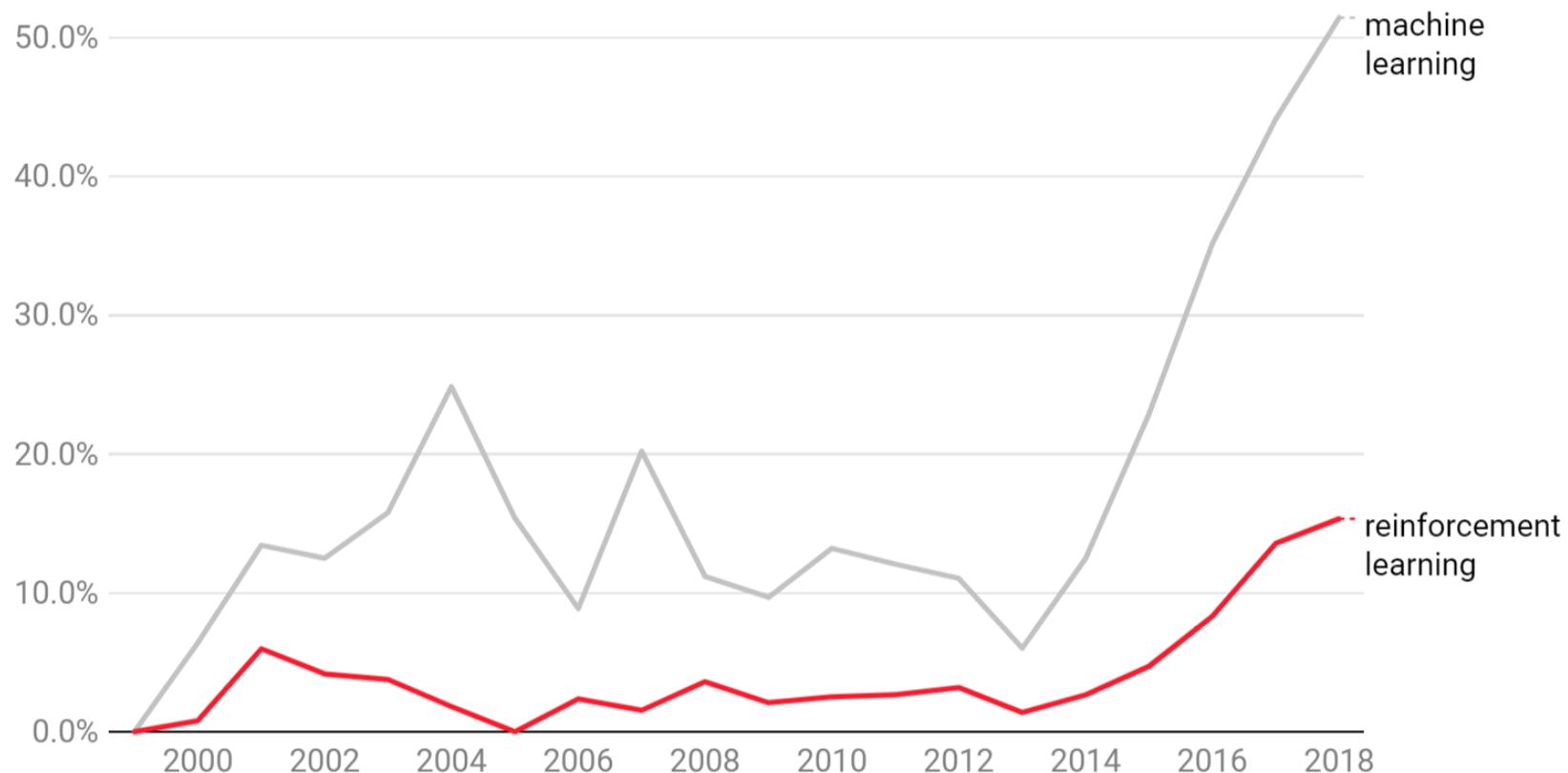
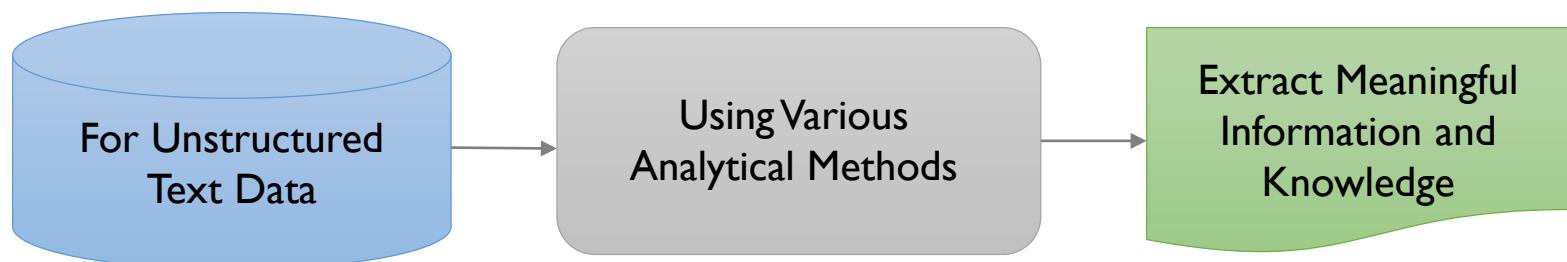


Chart: MIT Technology Review • Source: [arXiv.org](https://arxiv.org) • Created with Datawrapper

Text Analytics: 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.



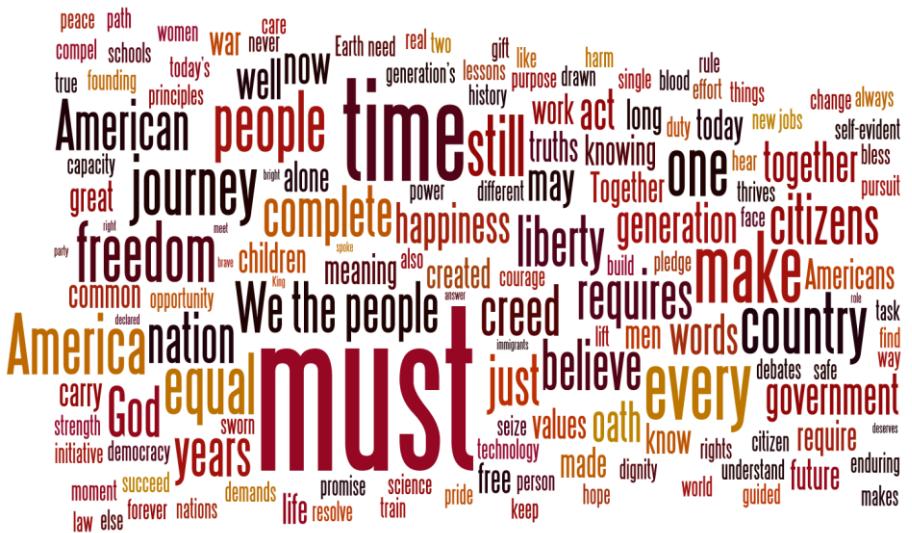
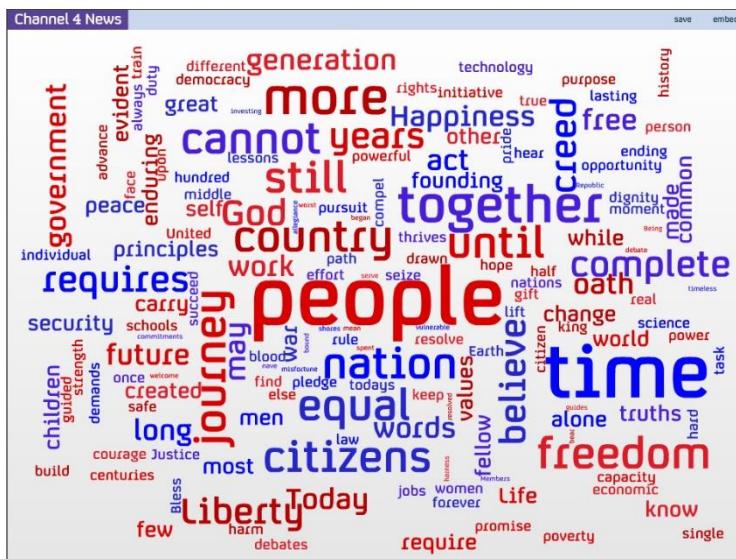
Text Analytics: Applications

- Information Abstraction/Summarization/Visualization



Text Analytics: Applications

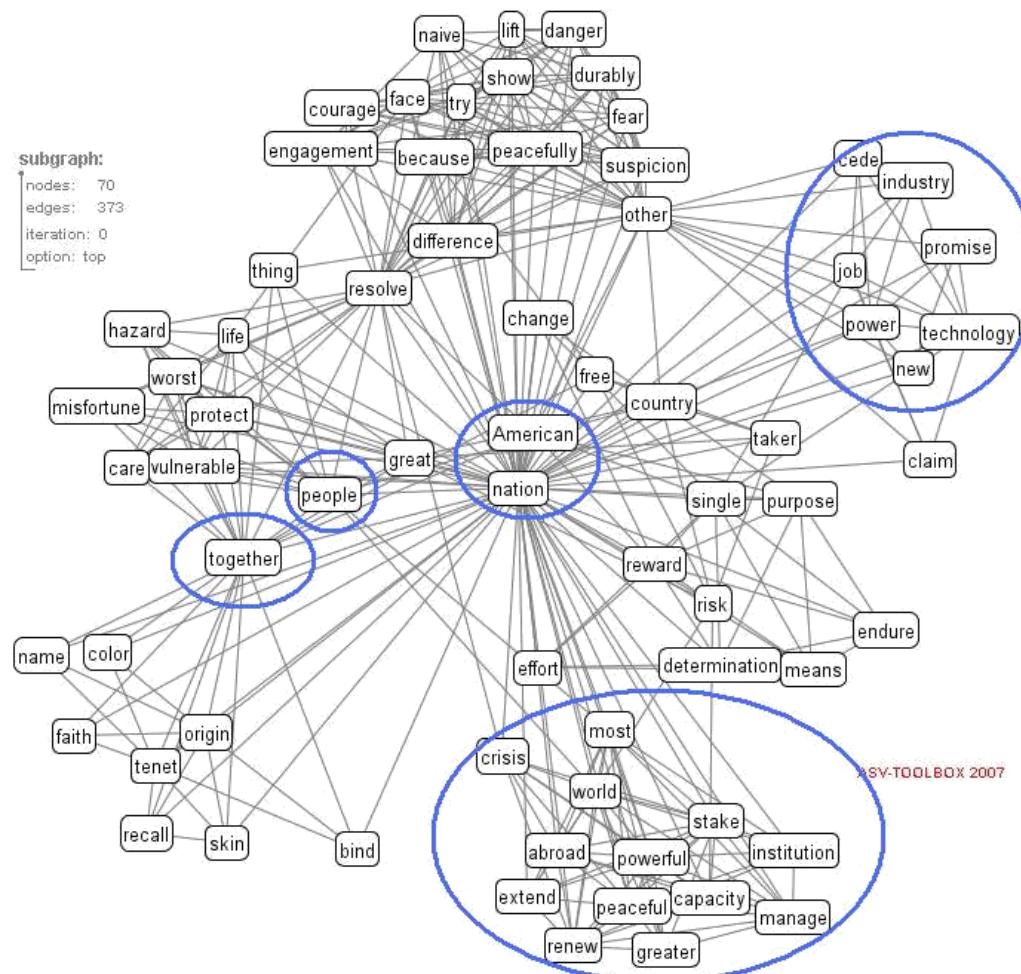
- Information Abstraction/Summarization/Visualization



Text Analytics: Applications

- Information Abstraction/Summarization/Visualization

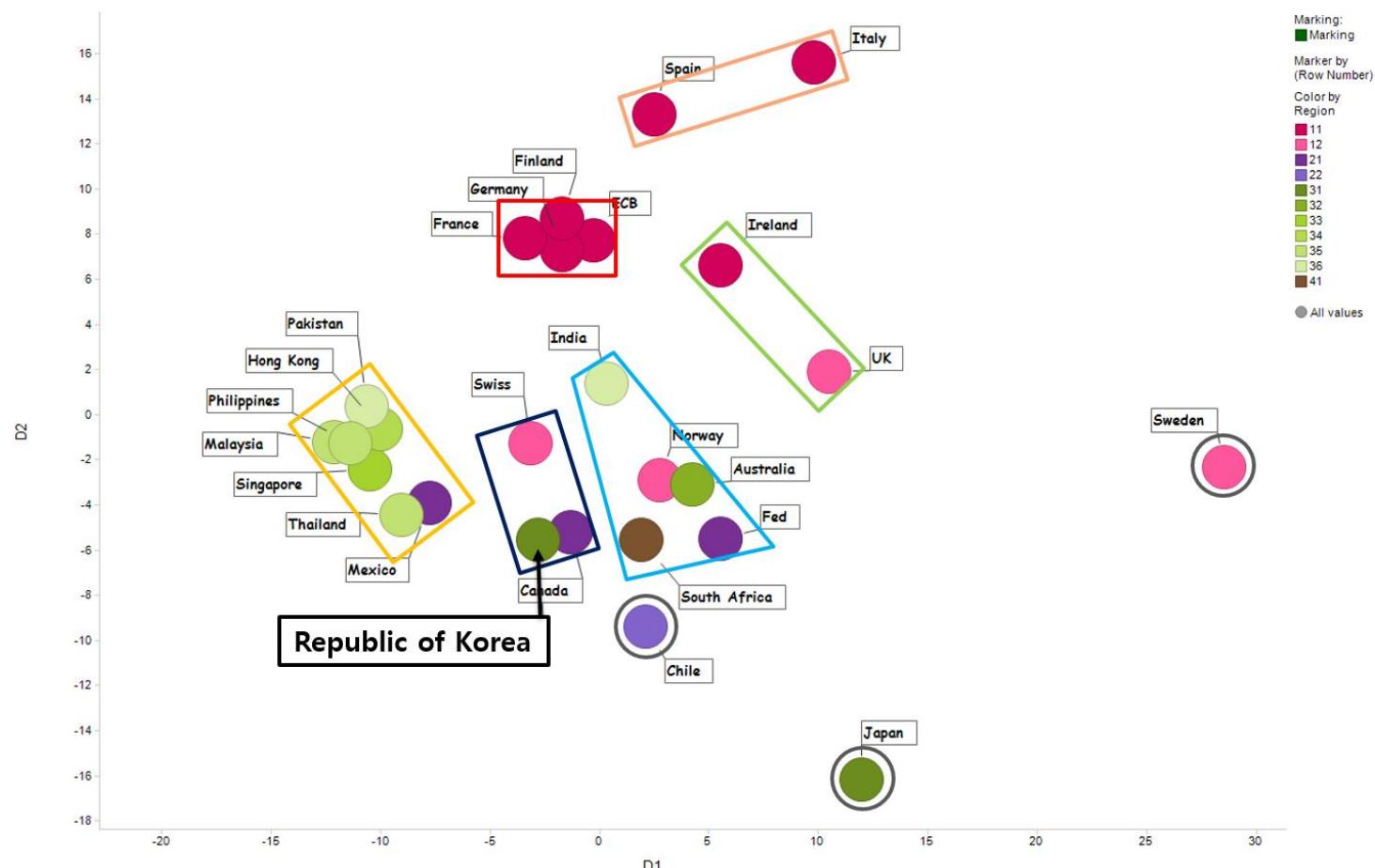
2013년 오바마 취임 연설 주제어 연결망 - nation



Text Analytics: Applications

- Information Abstraction/Summarization/Visualization

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



Text Analytics: 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 Analytics: Applications

Seo et al. (2020)

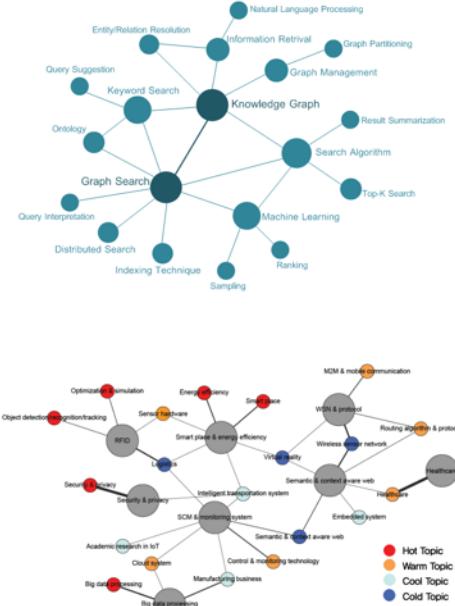
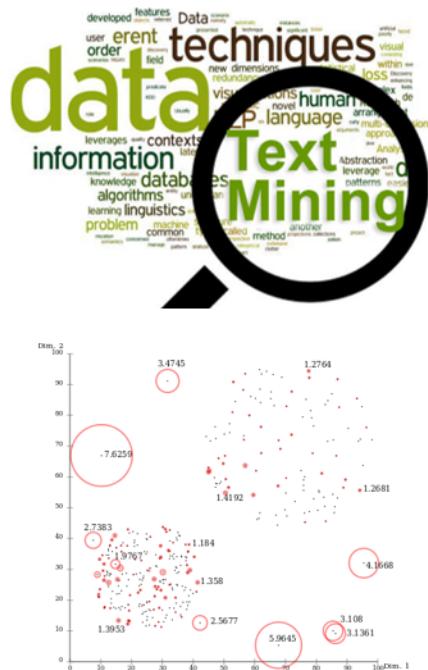
- Information Abstraction/Summarization/Visualization

- ✓ Unusual customer response identification and visualization



Doc	Comment
Doc 1	Comment1
Doc 2	Comment2
Doc 3	Comment3

VDS 주관식 문항 추출



데이터 마이닝 기법 적용

효율적인 소비자 의견 파악

Text Analytics: Applications

Seo et al. (2020)

• Information Abstraction/Summarization/Visualization

✓ Unusual customer response identification and visualization

1. 데이터 전처리 및 벡터화

데이터 전처리

- 속어, 구어체 등을 고려하지 않음
- Space → Apostrophe로 변환
- Na, 중복 문서 제거
- 이상 단어 탐지 및 검출



▪ 속어, 구어체 등을 고려 x



▪ Space → apostrophe



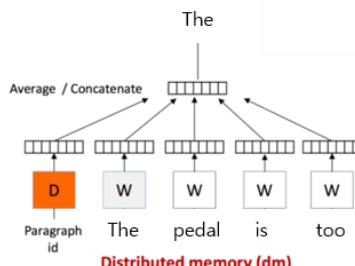
▪ Na, 중복 문서 제거



▪ 이상 단어 탐지 및 검출

문서 벡터화

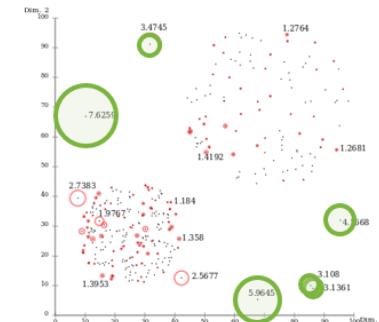
- Tokenization
- Doc2Vec



2. 특이 소비자 의견 탐지

Novelty Detection

- Doc2vec으로 벡터화 된 문서를 바탕으로 Local Outlier Factor(LOF) 기법 적용



키워드 추출

- TF-IDF 방법에 기반하여 일반적인 문서와 중복되지 않는 특이 응답 문서의 중요 단어를 추출

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

$$W_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

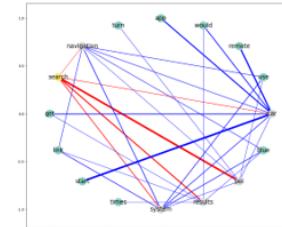
3. 키워드 및 관계도 분석

연관 규칙 관계 분석

- 키워드와 통계적 연관성이 높은 단어 분석

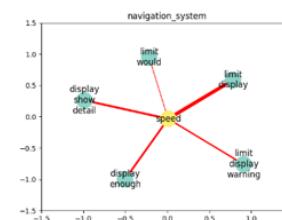
빈도 기반 관계 분석

- 키워드와 함께 등장한 단어 빈도 분석
- 키워드와 관련성이 높은 단어를 모두 포함하고 있는 문서 자동 추출



거리 기반 관계 분석

- 키워드의 단어 벡터와 N-gram의 벡터 사이에 거리 분석



Text Analytics: Applications

Seo et al. (2020)

- **Information Abstraction/Summarization/Visualization**

- ✓ **Unusual customer response identification and visualization**

Response comments	Ranking
had to be reprogrammed to give expressway routes....stopped doing this w/o my making changes.	1/17,430
This problem has persisted the entire length I have owned the vehicle; however, it only happens on 1 out of 10 start ups so it is not something that causes much grief.	1/17,430
Also, bell comes at a specific place near my home which address is deleted.	1/17,430

(a) Responses with the highest LOF scores for Navigation System

Response comments	Ranking
I drove a Lexus for many years. The seats were soft cushioning and soft leather covered.	1/8,687
THIS IS MY SECOND HYUNDAI. SAME PROBLEM. ON 2012 VERA CRUZ.	1/8,687
I have not had any trouble with the seat covers but I do think that they should be treated with something before they are sold, that would make them easier to keep clean.	1/8,687

(b) Responses with the highest LOF scores for Seat Material

Text Analytics: Applications

Seo et al. (2020)

- **Information Abstraction/Summarization/Visualization**

Response comments	Ranking
The app takes forever to load up, forever to send signal, and more often than not it's useless because my car is far from warmed up. Perhaps if I could set it to warm for longer than 10 min.	1/20,575
I ABSOLUTELY hate the fact that the hazards flash the entire time when remote starting the car. We live in NH, have had over 4 feet of snow this past month, temperatures haven't been above freezing in weeks and I still won't use the auto start because of the flashing lights. We will definitely be canceling the service because of this. And judging by all of the comments from other owners I read online, we are in the majority on this.	1/20,575
Whether connected to WiFi or mobile data, the app takes forever to log in (application to database coding flaws I'm sure) then, when in the app, when I press "lock" or "unlock" it takes a good 30 seconds to say "request sent" then, once the request is sent, it takes almost 3-5 minutes for the door to unlock! The app, by concept is brilliant, however the team working on the coding clearly is inexperienced and doesn't know how to write scripts to compile and run efficiently!	1/20,575

Text Analytics: Applications

Seo et al. (2020)

- Information Abstraction/Summarization/Visualization

- ✓ Unusual customer response identification and visualization

Table 4: Keywords that are more frequently used in unusual customer response comments for each category

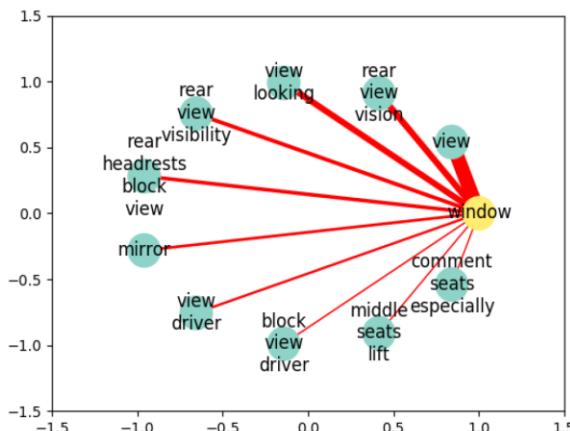
Category	Keywords
Adjustments and Controls	air, leather, window, cars, belt, setting, cooling, door, view
BlueLink Telematic System	money, information, area, month, right, auto, miles, oil, driving, user
Bluetooth	update, names, help
Navigation System	entering, iphone, dealership, locations, speed, manual
Ride Quality	engine, firm, genesis, body, driver, seats
Seat Belt	leg, extenders, thing, plastic, road, wheel, leather, arm, care, miles end, frame, years
Seat material	warranty, type, seating
Wind noise	bumper, seat, paint, water, trunk, mirror, rain, mileage

Text Analytics: Applications

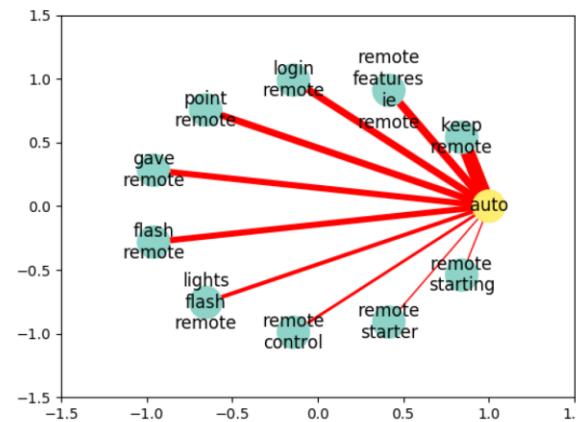
Seo et al. (2020)

- Information Abstraction/Summarization/Visualization

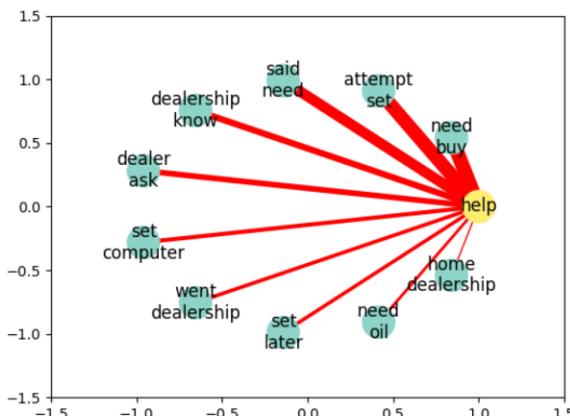
- ✓ Unusual customer response identification and visualization



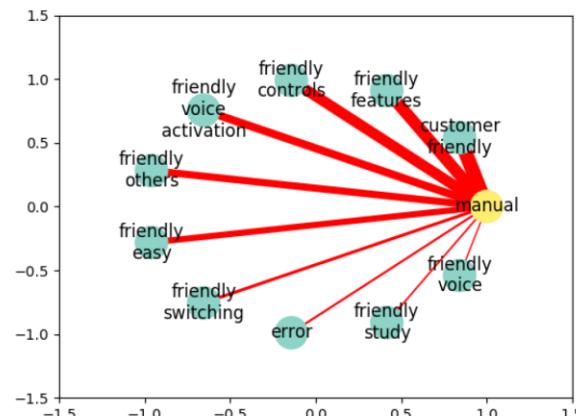
(a) Adjustments and controls



(b) BlueLink Telematic System



(c) Bluetooth



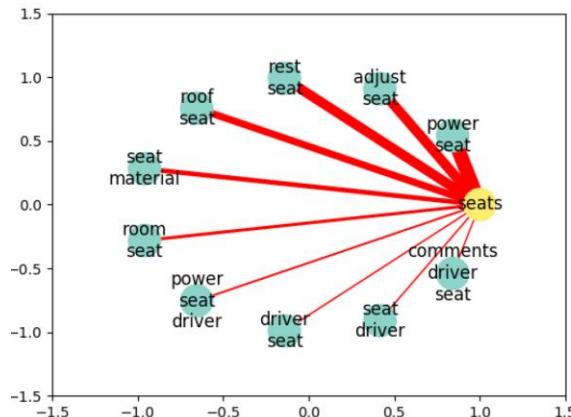
(d) Navigation system

Text Analytics: Applications

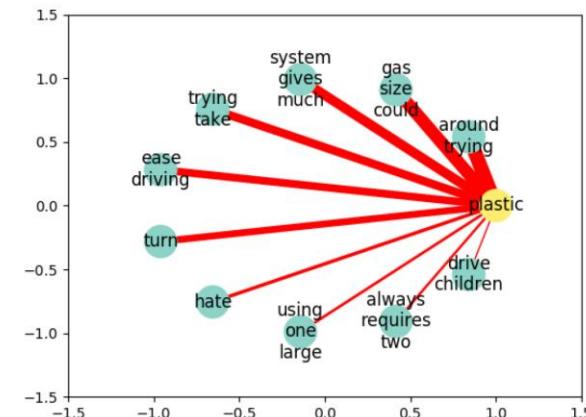
Seo et al. (2020)

- Information Abstraction/Summarization/Visualization

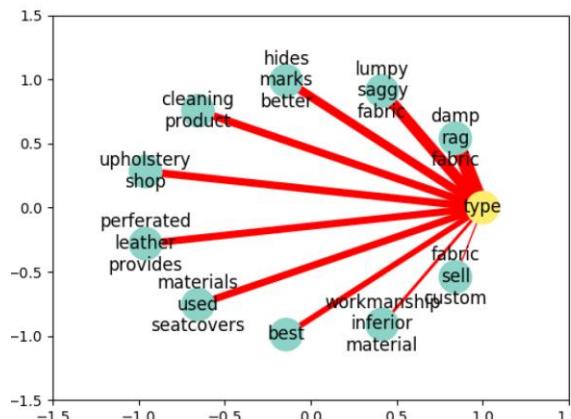
- ✓ Unusual customer response identification and visualization



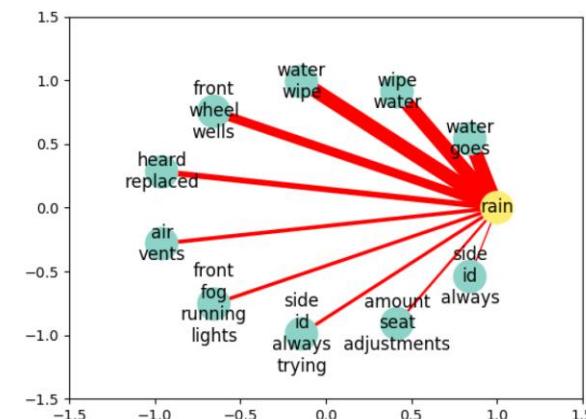
(e) Ride quality



(f) Seat belt



(g) Seat material

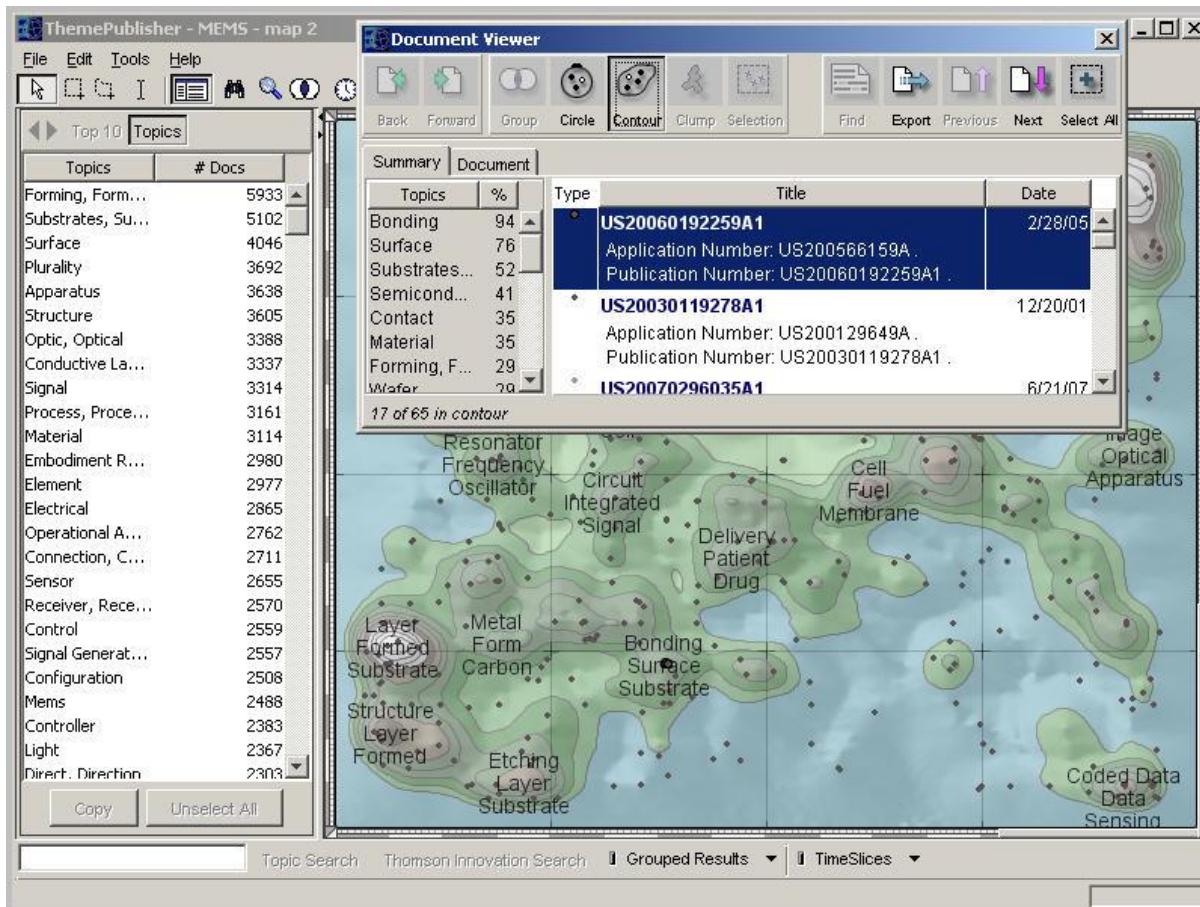


(h) Wind noise

Text Analytics: Applications

- Document Clustering

- ✓ Cluster documents and extract representative keywords for each cluster



Text Analytics: Applications

- Document Clustering

**10 YEARS TED TALKS
2006 – 2016**

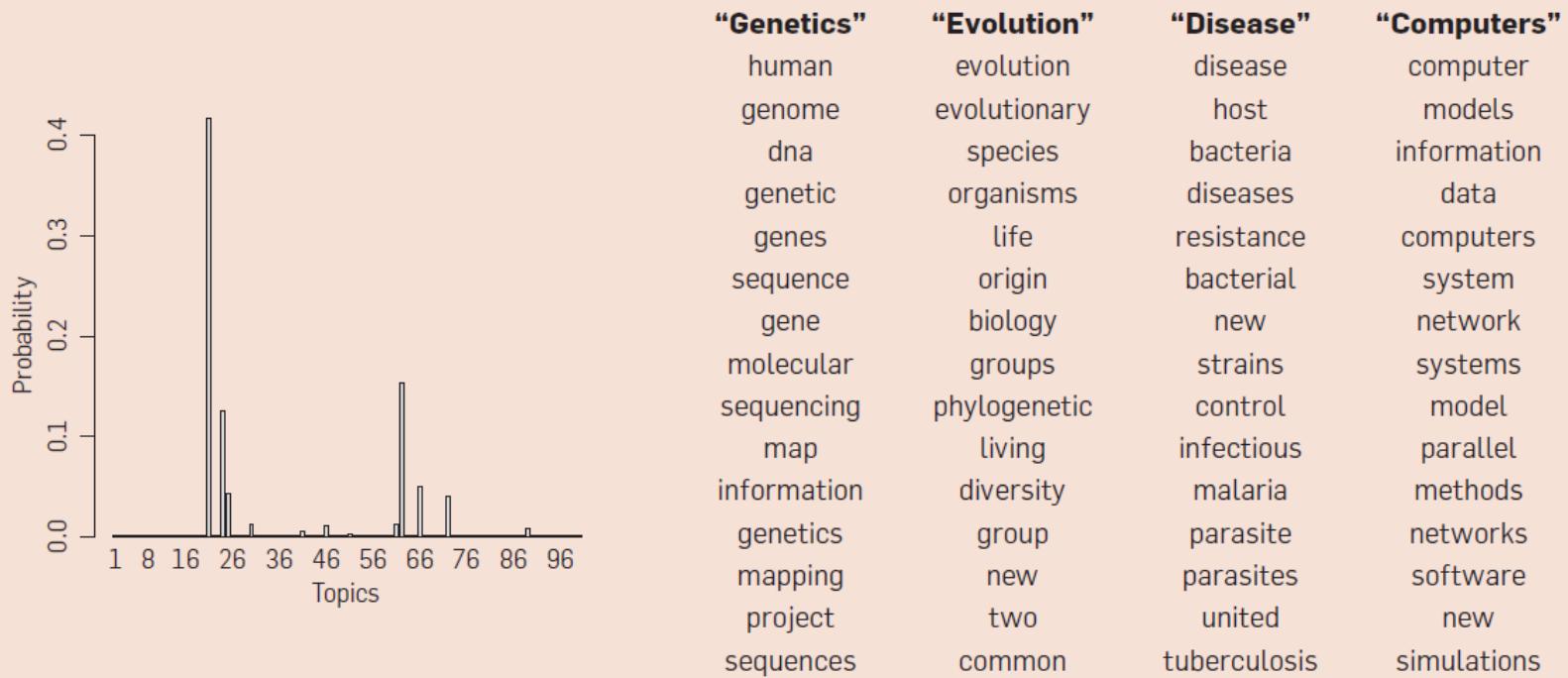
2224 TALKS
>30.000 minutes >500 hours

Text Analytics: 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 Analytics: Applications

Kim et al. (2016)

- Topic Extraction

✓ 30 Topics discovered by LDA

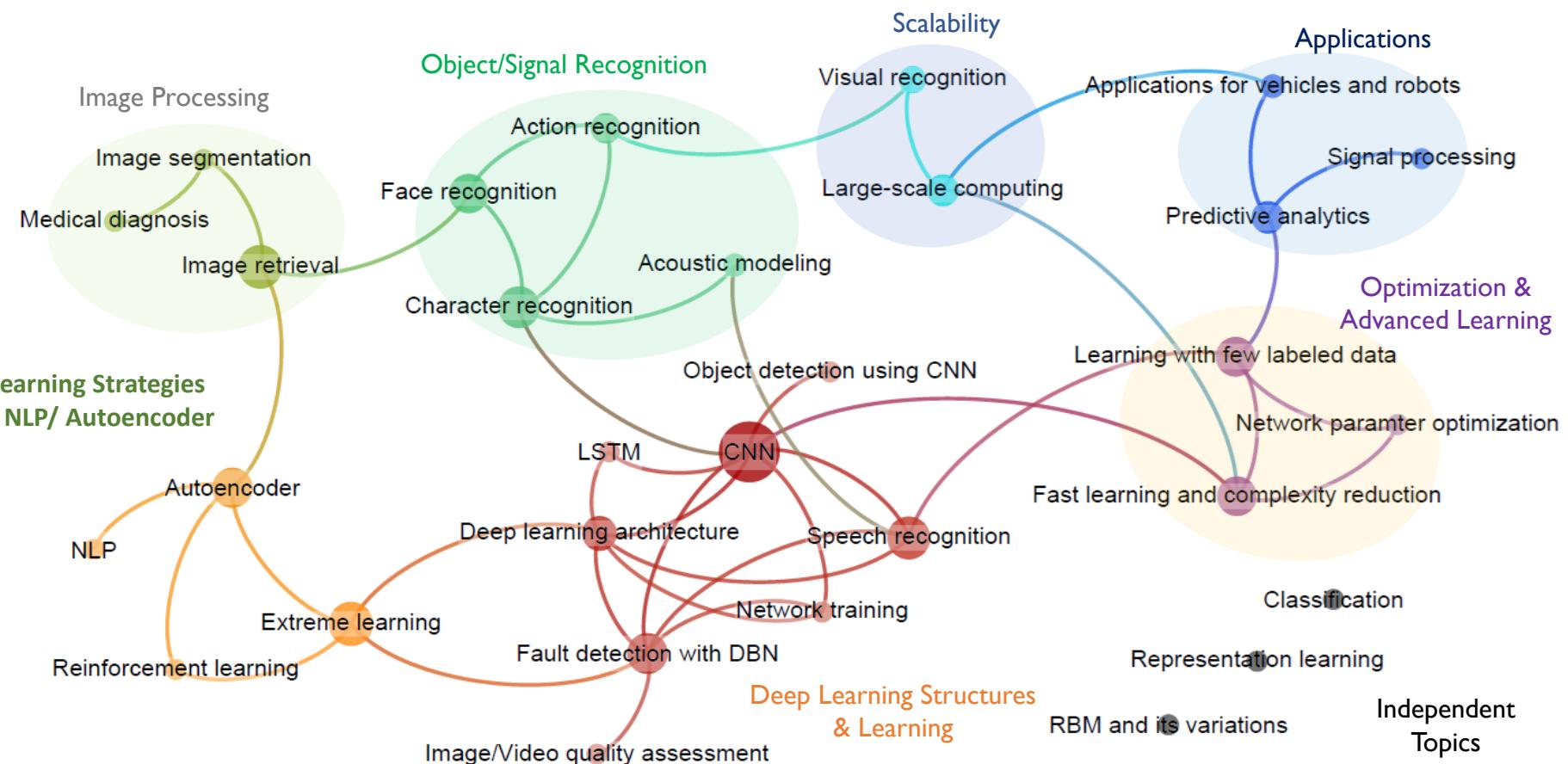
Fault detection with DBN	Convolutional neural network	Network Learning	Representation learning	Face Recognition	Speech Recognition	Acoustic Modeling	Extreme Learning	Deep learning architecture	Image Segmentation
deep belief network dbn fault	neural convolutional pool convolution convnet	layer input output 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 Analytics: Applications

Kim et al. (2016)

- Topic Extraction

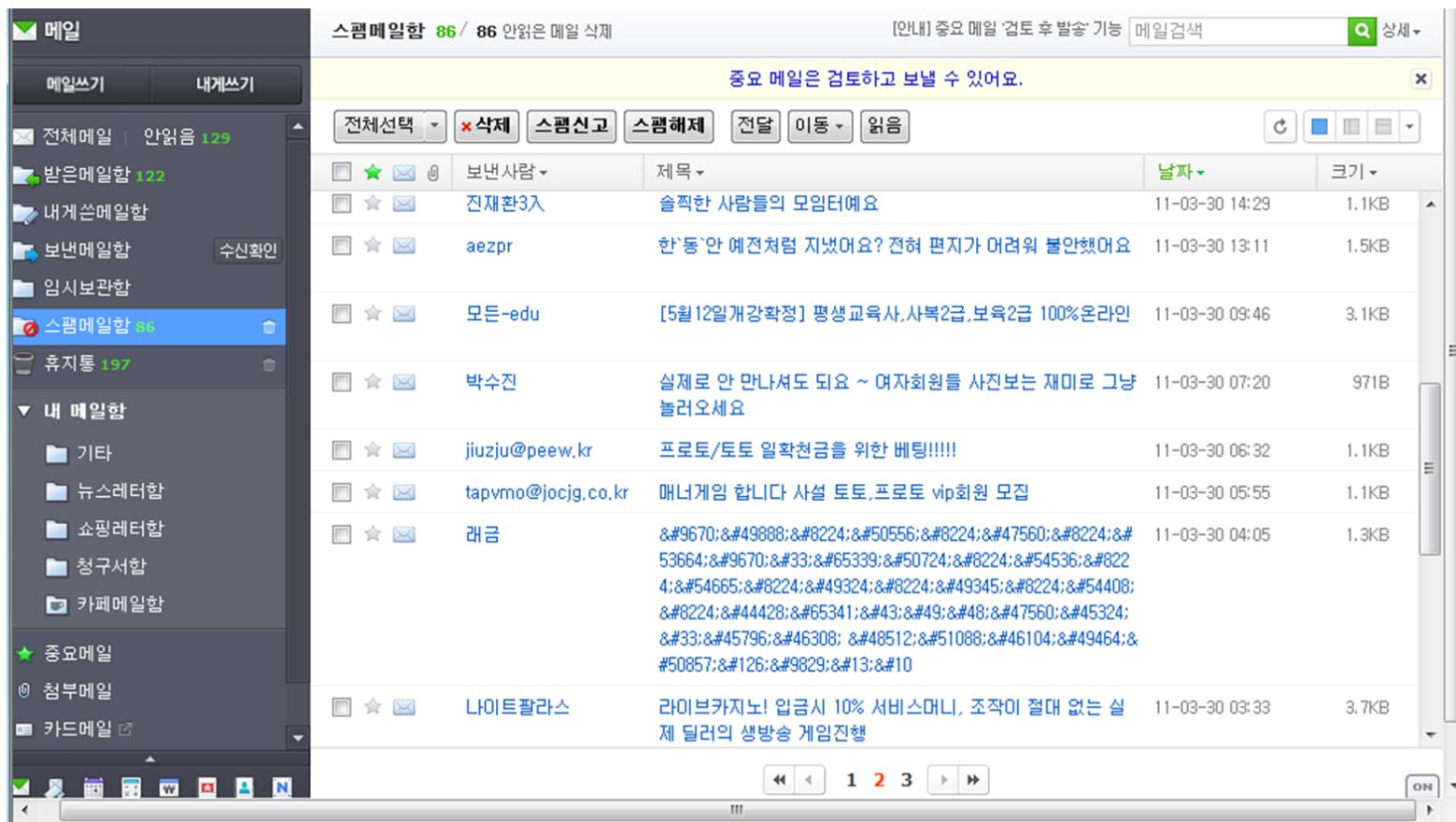
- ✓ Relations between topics



Text Analytics: Applications

- Document Categorization/Classification

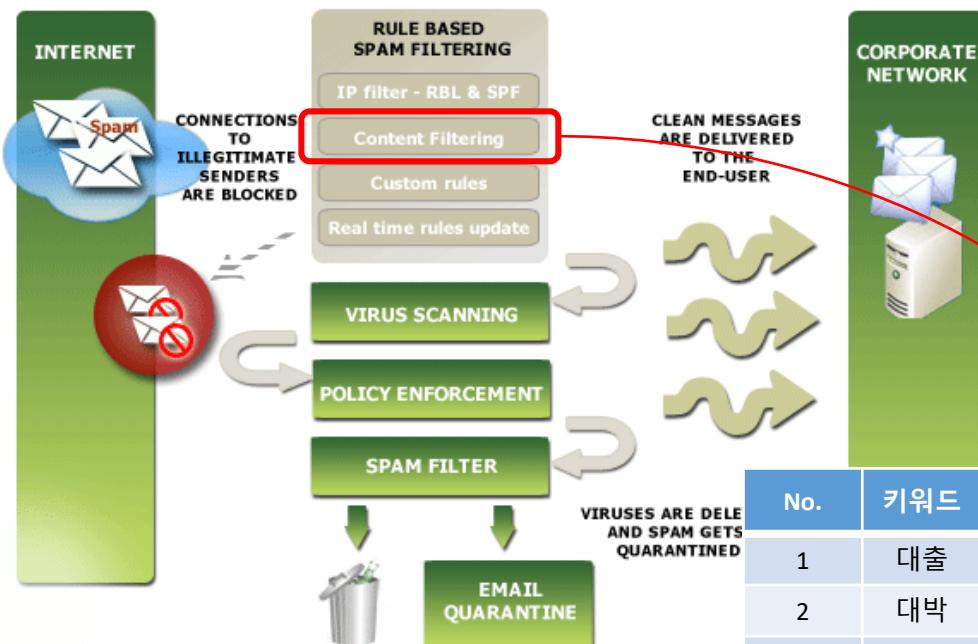
- ✓ Spam mail filtering



Text Analytics: Applications

- Document Categorization/Classification

- ✓ Spam mail filtering



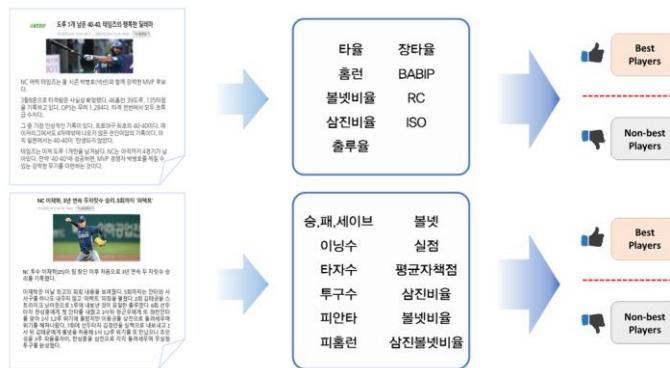
No.	키워드	Mail 1	Mail 2	Mail 3	...	Mail N
1	대출	0	2	0	...	0
2	대박	0	0	0	...	0
3	미팅	0	0	2	...	0
4	이상형	0	0	2	...	0
5	머니	0	2	0	...	0
6	외로	0	0	3	...	1
스팸 여부		N	Y	Y	...	N

Text Analytics: Applications

Kim et al. (2016)

- Document Categorization/Classification

- ✓ Sport player evaluation



Best Player Case



<2014.05.29 NC 테일즈 관련 기사>

이전 경기기록									
날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-28	테일즈	3	2	0	0.315	0	1	0	0
볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당	RC	BABIP
0.33	0	Inf	0.75	1.00	1.75	0.68	0	2.25	0.67

당일 경기기록									
날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-29	테일즈	6	5	7	4	0.333	2	0	0
볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당	RC	BABIP
0	0	NA	0.83	2.00	2.83	1.67	0.333	10.00	0.75

Target										Target
타수	연타	타점	득점	타율	홈런	볼넷	삼진	타석당	RC	BABIP
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
불꽃 삼진 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당	RC	BABIP	0.25
0	0	0	0	0	0	0	0	0	0	0.25

Best player	1
-------------	---

Non-best Player Case

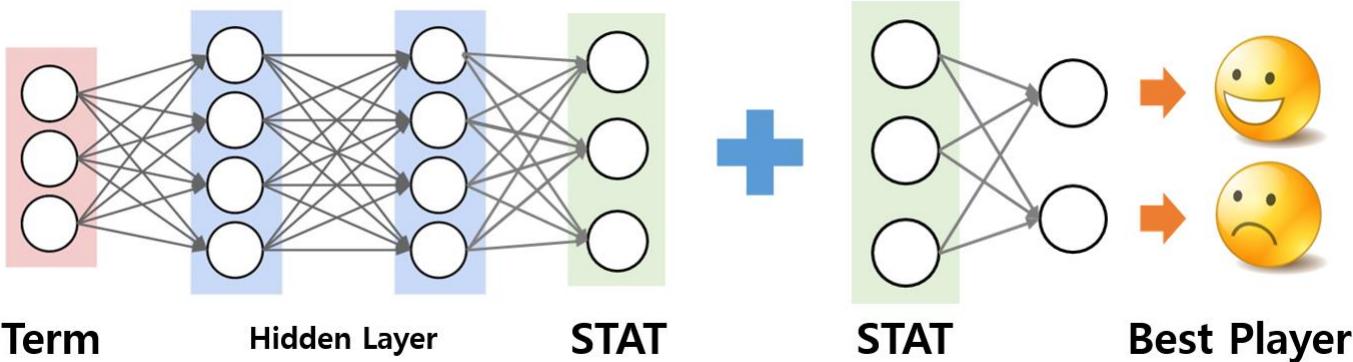


<2014.05.29 NC 손시현 관련 기사>

이전 경기기록									
날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-28	손시현	4	1	0	0	0.292	0	0	0
볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당	RC	BABIP
0	0	0	0	0.25	0.5	-0.042	0	0.25	0.25

당일 경기기록									
날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-29	손시현	4	1	0	1	0.291	0	1	0
볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당	RC	BABIP
0.25	0	Inf	0.4	0.25	0.65	-0.041	0	0.40	0.25

Best player	0
-------------	---



Text Analytics: Applications

Lee et al. (2017)

- Document Categorization/Classification

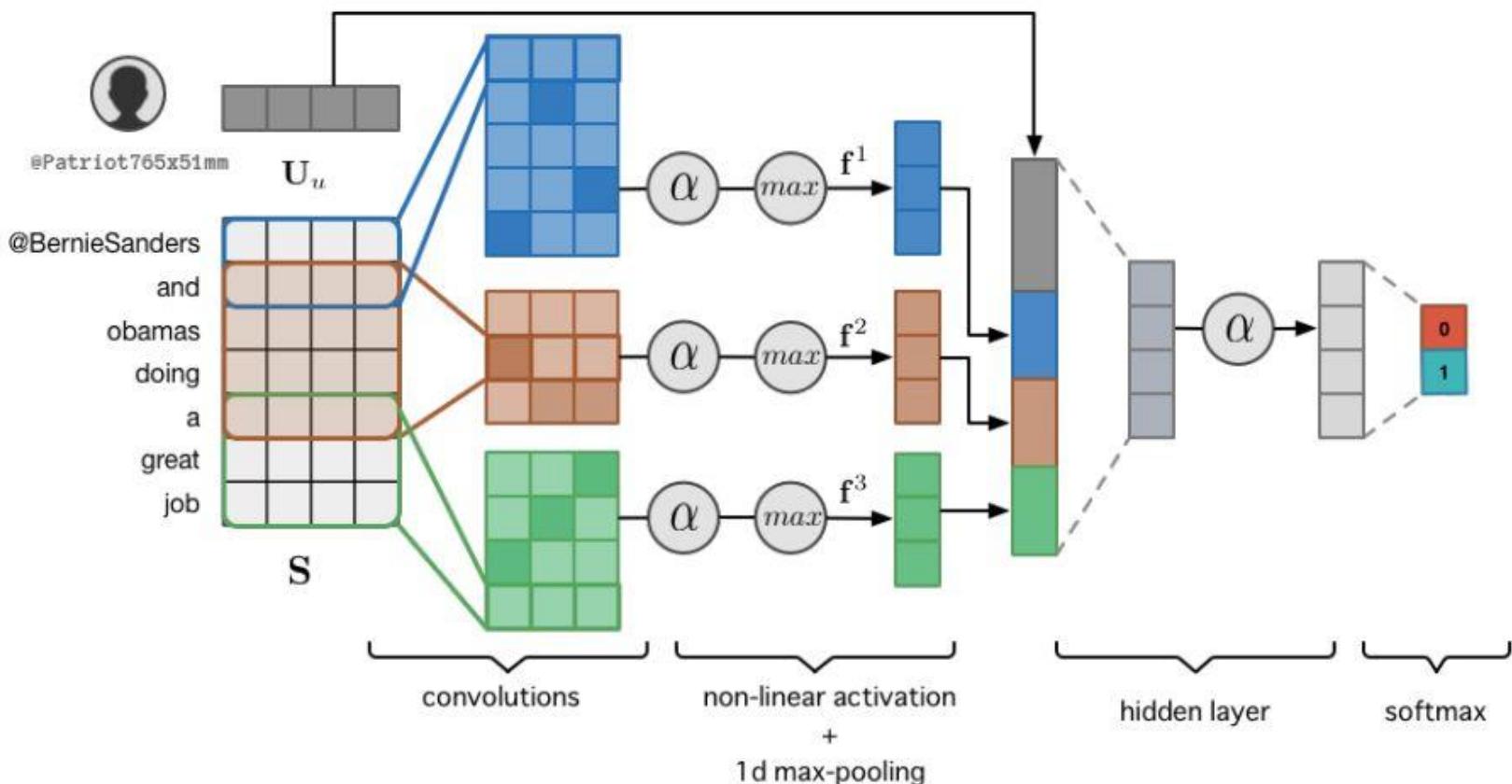
- ✓ Sentiment Analysis

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' NCLG 상대로 창단 첫 스크립트	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 1사후 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 Analytics: Applications

- Document Categorization/Classification

- ✓ Sentiment Analysis

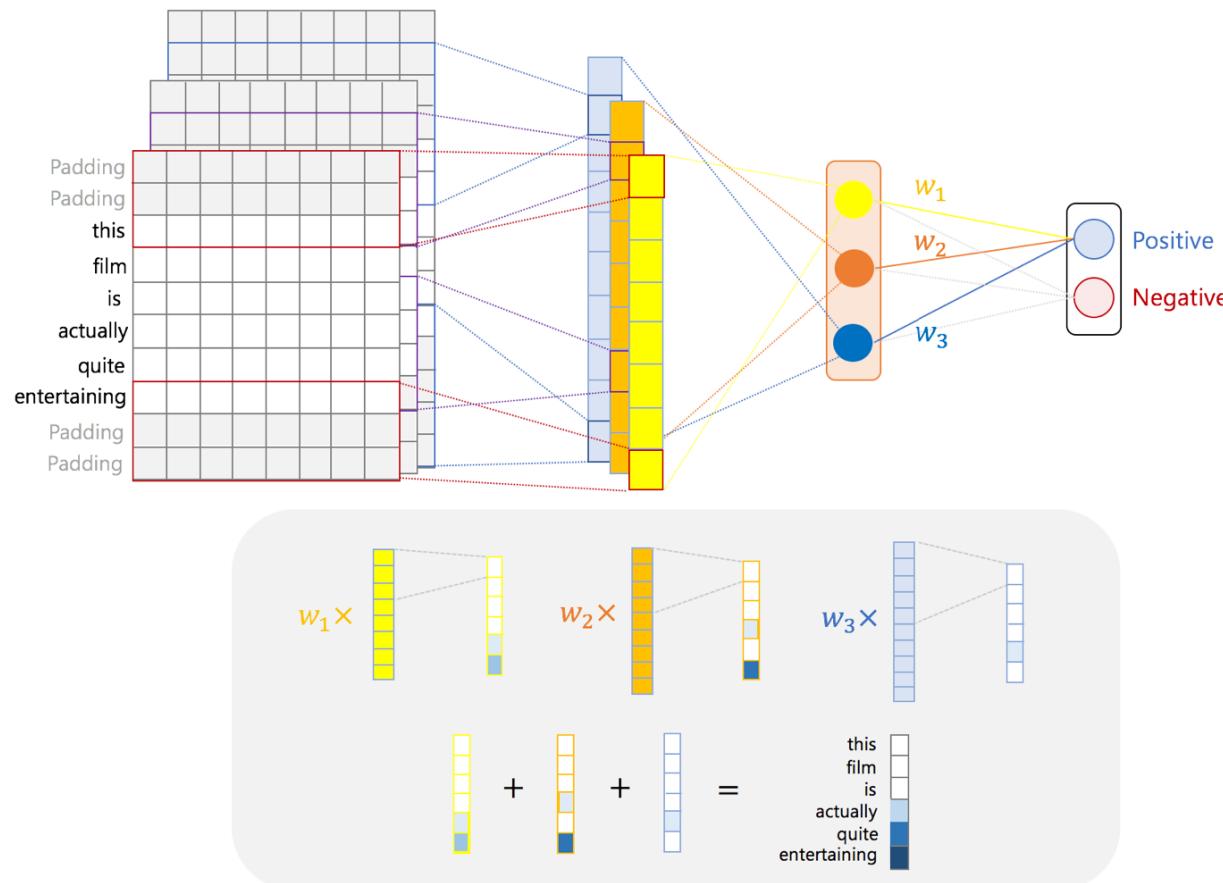


Text Analytics: Applications

Lee et al. (2017)

- Document Categorization/Classification

- ✓ Sentiment Analysis

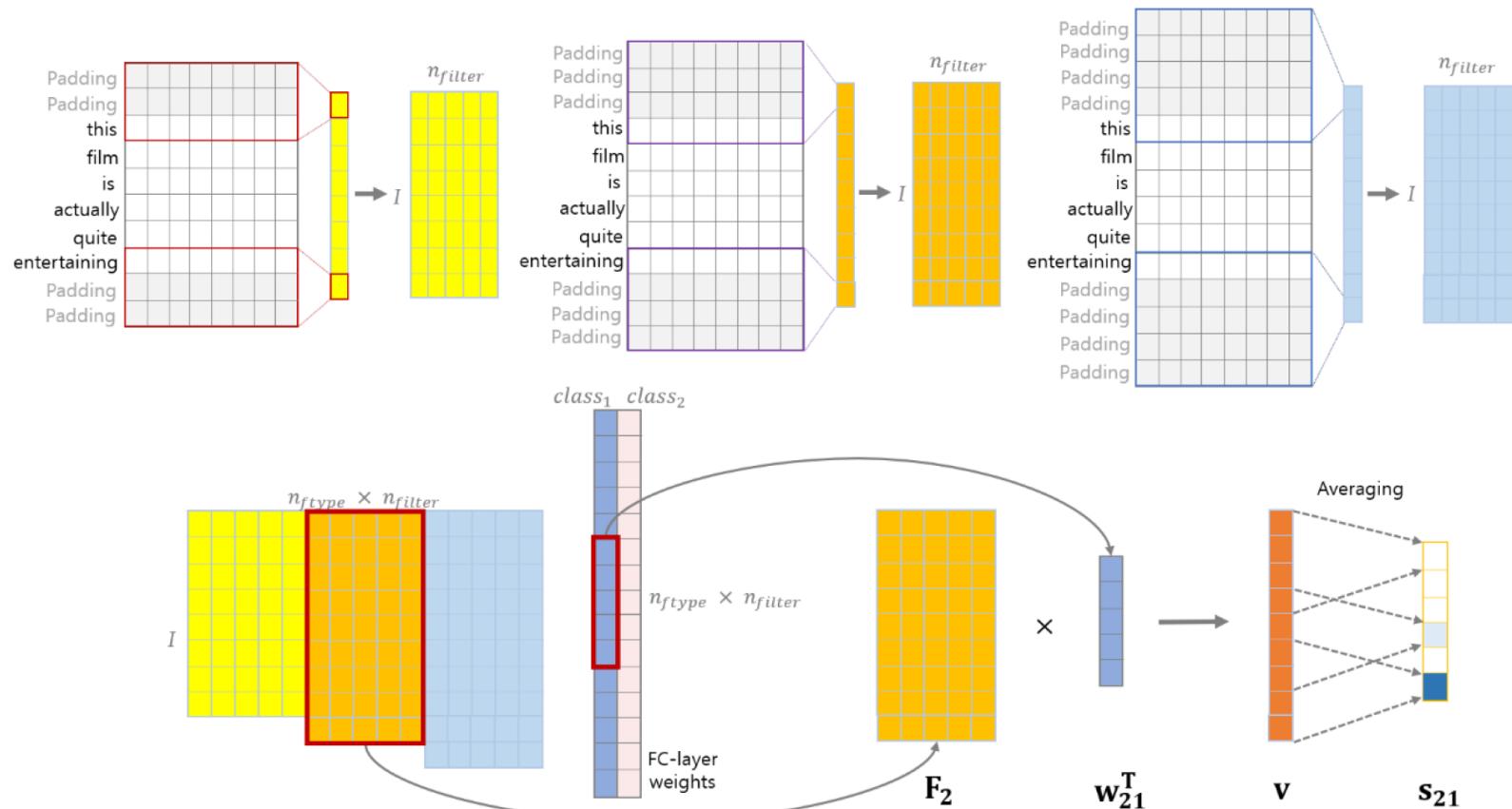


Text Analytics: Applications

Lee et al. (2017)

- Document Categorization/Classification

- ✓ Sentiment Analysis



Text Analytics: Applications

Lee et al. (2017)

- Document Categorization/Classification

- ✓ Sentiment Analysis

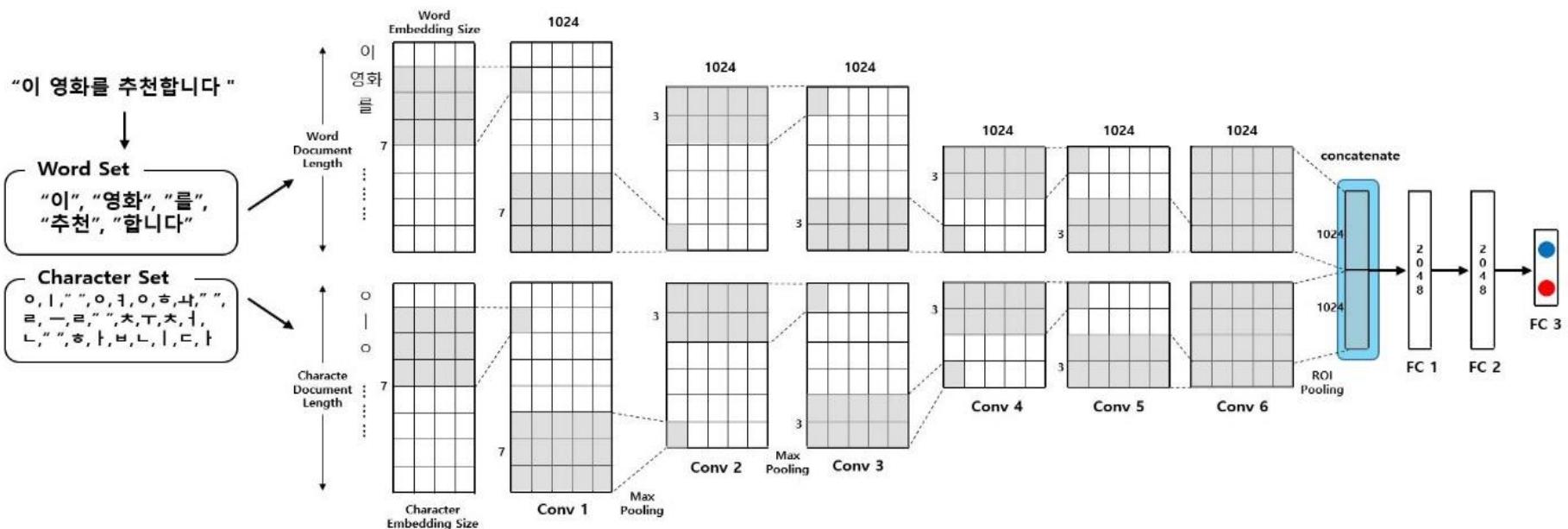
Methodology	Sentence
Raw text	I hate this movie. It is a horrid movie. Sean Young's character is completely unsympathetic. Her performance is wooden at best. The storyline is completely predictable and completely uninteresting. I would never recommend this film to anyone. It is one of the worst movies I have ever had the misfortune to see. (1 / 10 points)
CAM ² -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 uninteresting I would never recommend this film to anyone It is one of the worst movies I have ever had the misfortune to see Negative
CAM ² -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 uninteresting I would never recommend this film to anyone It is one of the worst movies I have ever had the misfortune to see Negative
CAM ² -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 uninteresting I would never recommend this film to anyone It is one of the worst movies I have ever had the misfortune to see Negative
CAM ² -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 uninteresting I would never recommend this film to anyone It is one of the worst movies I have ever had the misfortune to see Negative
CAM ² -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 uninteresting 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 uninteresting 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 Analytics: Applications

Mo et al. (2017)

- Document Categorization/Classification

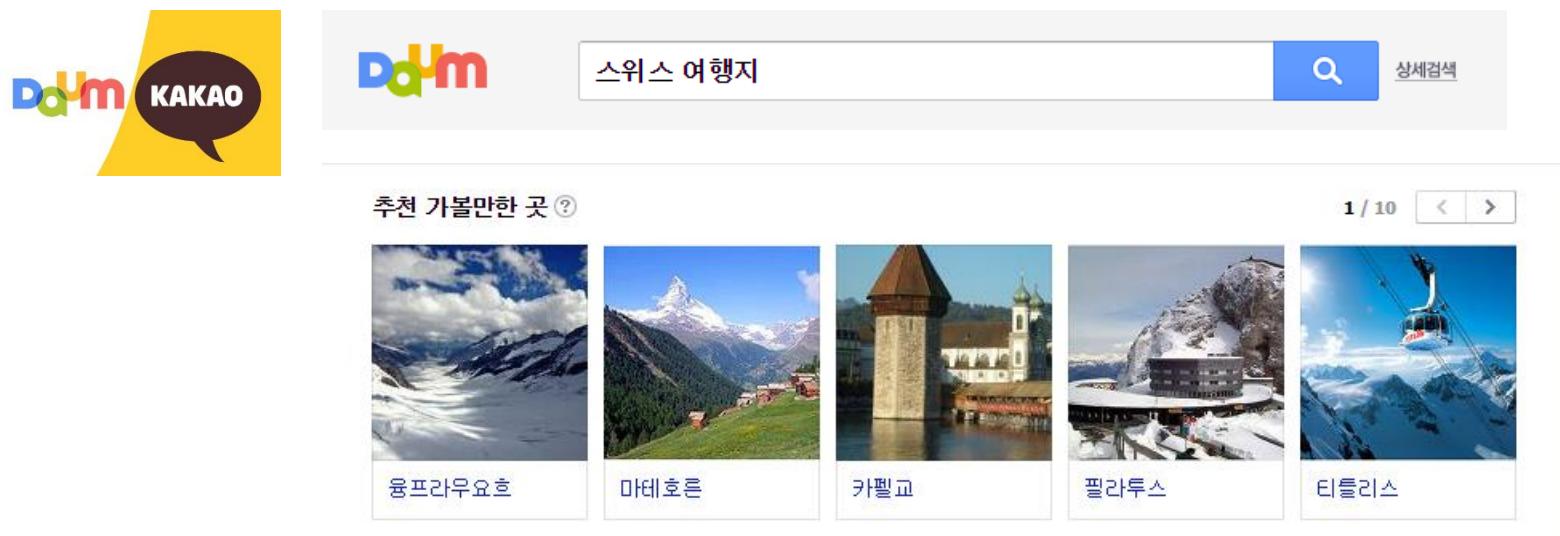
- ✓ Sentiment Analysis



Text Analytics: 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 Analytics: 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

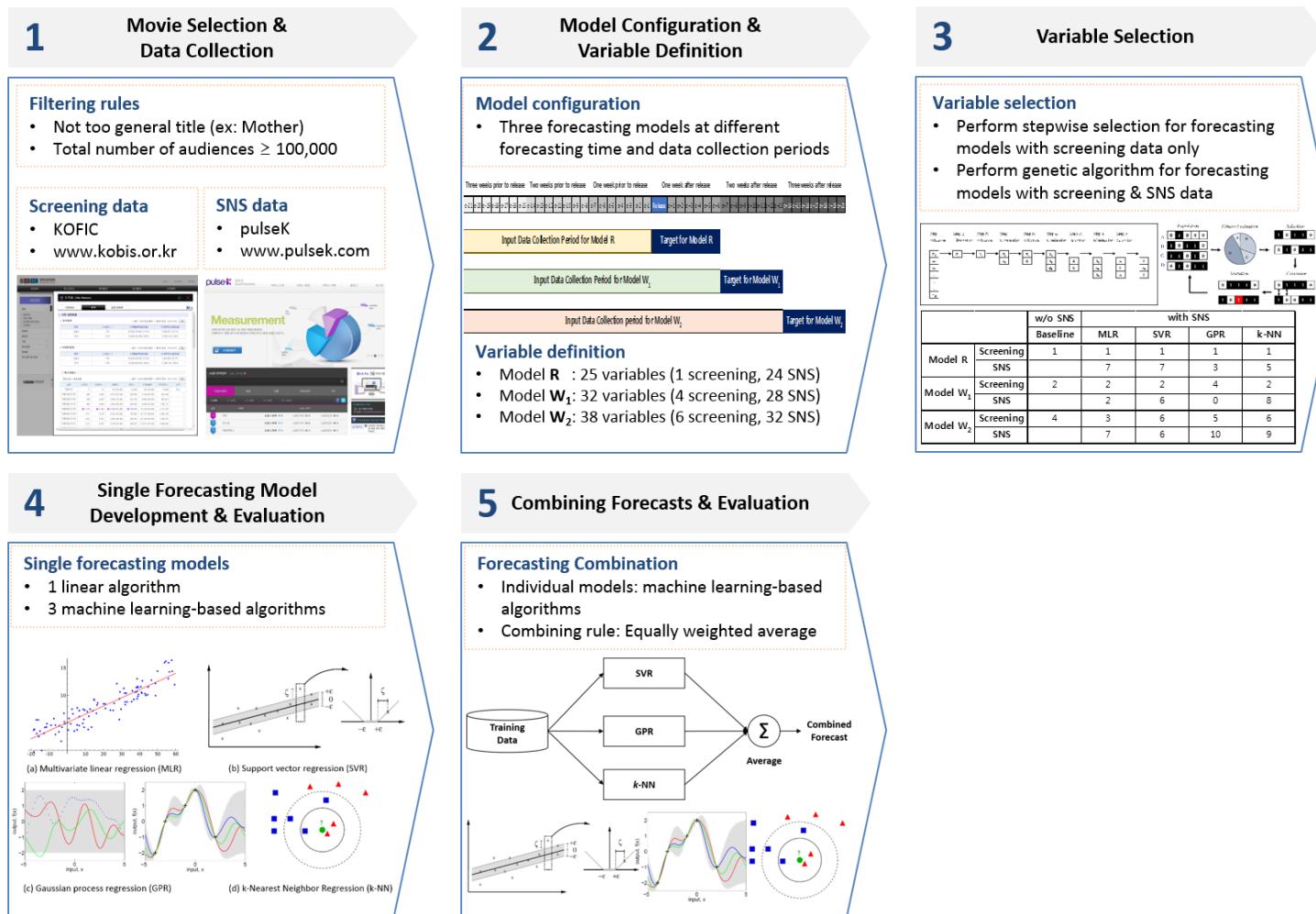
The screenshot shows the Dining Code website interface. At the top, there is a search bar with the query '고려대'. Below the search bar, there is a sidebar with categories: 배달, 수요미식회, 박종원의 3대천왕, 생활의달인, TV맛집, 한식, 중식, 일식, and 카페. A '전체보기' button is also present. The main content area displays a list of restaurants under the heading "'고려대' 상위 랭킹 맛집 1339곳의 맛집". The first result is '1. 어머니대성집' (어머니대성집), which is a Korean barbecue restaurant located in Gwangju, South Korea. It has a rating of 4.5 stars, 238 likes, and 14 reviews. The second result is '2. 윤휘식당' (윤휘식당), a Japanese restaurant in Seoul, with a rating of 4.5 stars, 244 likes, and 19 reviews. The third result is '3. 애정 마라샹궈 1호점' (애정 마라샹궈 1호점), a Ma La Xiang Guo restaurant in Seoul, with a rating of 4.5 stars, 35 likes, and 11 reviews. At the bottom of the page, there is a footer with contact information: TEL 02-928-0683, SCORE 139, and a note about the restaurant's specialty: '아는한 모던한분위기, 고급스러운 세련된 분위기, 여자들끼리 런치, 조식'.

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

Text Analytics: Applications

Kim et al. (2015)

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



Text Analytics: Applications

Kim et al. (2015)

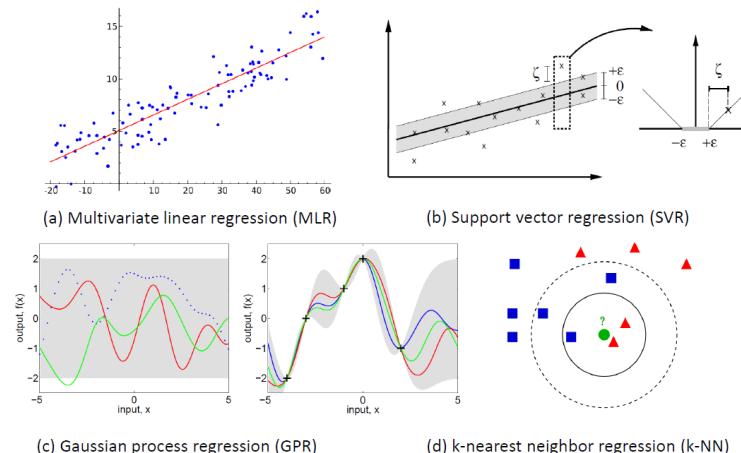
- Improve forecasting accuracy combined with structured data
 - ✓ Forecasting the box office scores based on the polarity of SNS posts

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	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 Analytics: Applications

Kim et al. (2015)

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

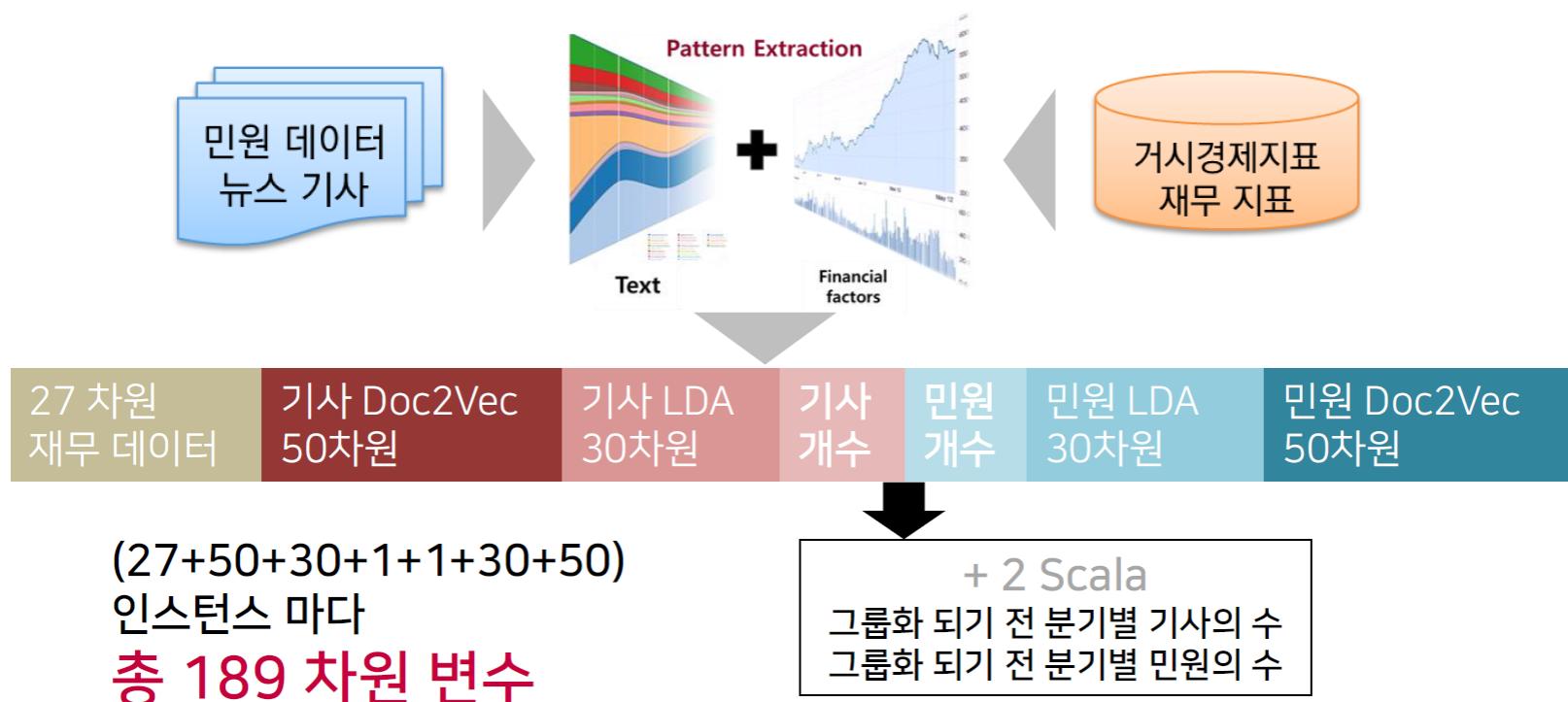


Model	Linear Algorithm		Machine Learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.8626	0.5958 (30.93%)	0.4278 (50.40%)* (28.19%)+	0.5087 (41.02%)* (14.62%)+	0.4605 (46.61%)* (22.71%)+	0.4657 (46.01%) (21.84%)	0.4496 (47.88%)* (24.54%)+ (3.45%)
Model W ₁	0.1664	0.1535 (7.76%)	0.1251 (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%)

Text Analytics: Applications

송서하 외 (2019)

- Improve forecasting accuracy combined with structured data
 - ✓ Early warning model for financial firms



Text Analytics: Applications

송서하 외 (2019)

- Improve forecasting accuracy combined with structured data
 - ✓ Early warning model for financial firms



27개 재무 변수	
BIS자기자본비율	총자산경비율
BIS기본자본비율	자기자본이익률
총여신대비외화대출비율	수지비율
순이자마진율	원화예대금리차
단순자기자본비율	외화예대금리차
고정이하여신비율	총자산순이익율_1
연체대출채권비율	총자산순이익률_2
대손충당금적립비율	단기대출비율
커버리지비율	예대율
순요주의이하여신비율	부채총계
부채총계대비외화부채비율	자산총계
총여신대비고객예수금	자본총계
고정이하여신비율	순이자마진율
대손충당금비율	

실제 로지스틱 회귀
조기 경보 모형에서 사용중인 재무 변수

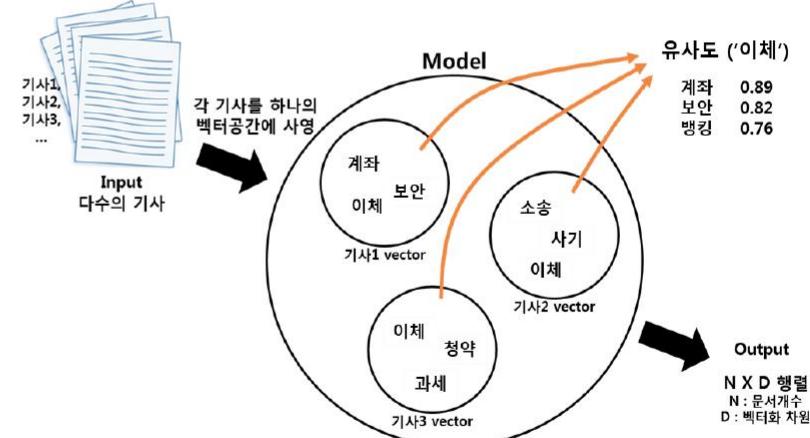
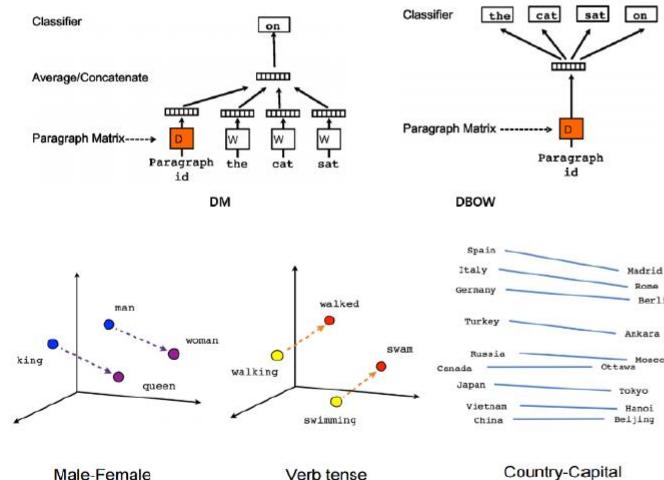
사.코드	회사명	날짜	C001.BIS지	C002.BIS가	A011.총여	E012.순이	C003.단순
10001	우리	200212	11.5927	6.580368	9.567668	3.136833	4.736041
10001	우리	200303	11.25434	6.543492	9.626936	3.149197	4.804393
10001	우리	200306	11.93472	7.269505	8.343038	3.306897	4.925109
10001	우리	200309	11.83714	7.351041	7.785889	3.2216	6.153432
10001	우리	200312	11.22903	6.821693	8.040745	3.19587	5.533509
10001	우리	200403	11.39602	6.975554	7.571517	3.170896	5.818036
10001	우리	200406	12.06747	7.559141	7.553917	3.000672	6.220084
10001	우리	200409	12.08542	7.152281	7.333561	3.041082	5.867916
10001	우리	200412	12.20036	7.814226	7.264898	2.987705	7.257877
10001	우리	200503	12.27278	8.625104	7.760409	2.915894	7.155567
10001	우리	200506	12.49602	8.935698	7.881458	2.90306	7.530039
10001	우리	200509	12.32116	9.035741	8.025136	2.820672	7.546571
10001	우리	200512	11.64708	8.085056	7.725328	2.967578	7.742147
10001	우리	200603	11.20014	7.847749	7.518282	2.904957	7.951988
10001	우리	200606	11.47836	7.582419	7.224173	2.776941	7.116087
10001	우리	200609	12.11703	7.700332	7.255265	2.666571	6.956471

Text Analytics: Applications

송서하 외 (2019)

- Improve forecasting accuracy combined with structured data

- ✓ Early warning model for financial firms



〈 단어들 사이의 의미론적 관계의 보존 〉

〈 Doc2Vec을 통한 뉴스기사 및 민원 데이터 표현 〉

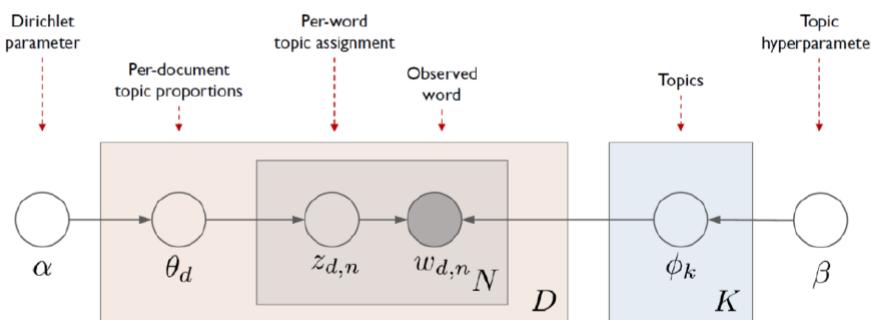
Text Analytics: Applications

송서하 외 (2019)

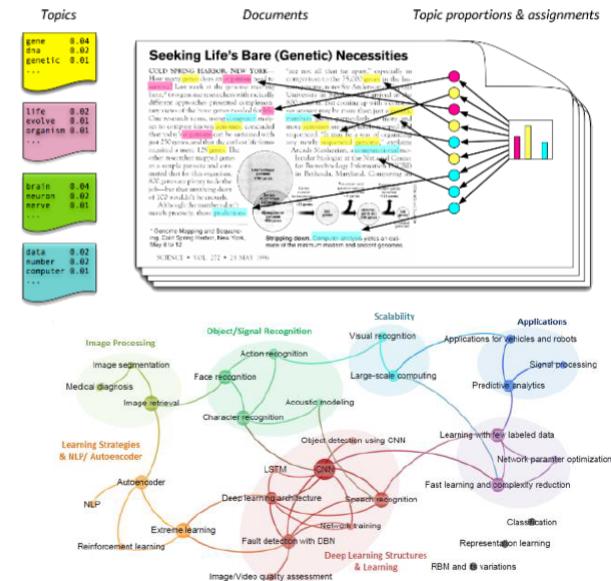
- Improve forecasting accuracy combined with structured data
 - ✓ Early warning model for financial firms



Topic Modeling(Latent Dirichlet Allocation;LDA)



⟨ Topic Modeling (LDA) 프로세스 ⟩



⟨ 토픽 모델링 수행 결과 예시 ⟩

Text Analytics: Applications

송서하 외 (2019)

- Improve forecasting accuracy combined with structured data
- ✓ Early warning model for financial firms

Table 4. Model performance evaluation

		without text data			with text data		
		Precision	Recall	F1	Precision	Recall	F1
recent 1 quarter	logistic regression	0.5557 (0.2117)	0.90 (0.2)	0.6636 (0.1837)	0.6162 (0.2239)	0.950 (0.15)	0.722 (0.1703)
	random forest	0.8889 (0.2733)	0.7667 (0.3091)	0.8044 (0.2758)	0.9556 (0.1133)	1.0 (0)	0.9733 (0.0680)
	recurrent neural net	/	/	/	/	/	/
	logistic regression	0.8544 (0.1871)	1.0 (0)	0.9070 (0.1340)	0.6289 (0.232)	1.0 (0)	0.7483 (0.1692)
recent 2 quarters	random forest	0.8833 (0.2693)	0.80 (0.3055)	0.8189 (0.2686)	0.9556 (0.1133)	1.0 (0)	0.9733 (0.0680)
	recurrent neural net	0.8733 (0.1868)	0.8111 (0.1863)	0.8252 (0.1539)	0.9228 (0.1462)	0.8667 (0.1846)	0.8775 (0.1426)
	logistic regression	0.8633 (0.1871)	1.0 (0)	0.9146 (0.1206)	0.5967 (0.2466)	0.950 (0.150)	0.7083 (0.1955)
	random forest	0.9278 (0.2094)	0.90 (0.2380)	0.9033 (0.2147)	0.9556 (0.1133)	1.0 (0)	0.9733 (0.0680)
recent 3 quarters	recurrent neural net	0.9444 (0.1242)	0.9833 (0.0898)	0.9556 (0.0916)	0.9167 (0.1537)	0.9833 (0.0898)	0.9488 (0.1074)
	logistic regression	0.6911 (0.2016)	1.0 (0)	0.8013 (0.1354)	0.6144 (0.1979)	0.9833 (0.0898)	0.7384 (0.1475)
	random forest	0.9444 (0.1449)	0.8833 (0.2115)	0.8978 (0.1655)	0.950 (0.1302)	0.9833 (0.0898)	0.9633 (0.1048)
	recurrent neural net	0.9222 (0.1595)	0.9833 (0.0898)	0.940 (0.1133)	0.9111 (0.1652)	0.9667 (0.1247)	0.9222 (0.1242)

Text Analytics: Applications

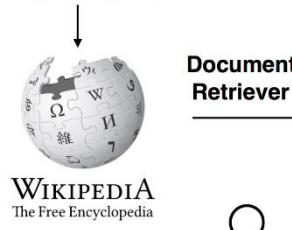
- Natural Language Understanding: Question Answering



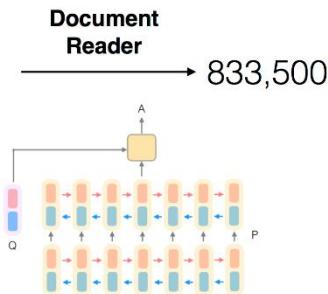
Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



Document
Retriever



<https://github.com/facebookresearch/DrQA/blob/master/img/drqa.png>

Text Analytics: Applications

- Natural Language Understanding: Question Answering

Q: When did Kendrick Lamar's first album come out?

Article [Talk](#) Read [View source](#) [View history](#) Search Wikipedia [Q](#)

Kendrick Lamar

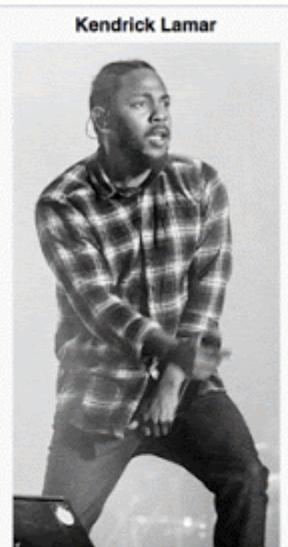
From Wikipedia, the free encyclopedia

An accepted version of this page, accepted on 28 February 2018, was based on this revision.

Kendrick Lamar Duckworth (born June 17, 1987) is an American rapper and songwriter. Raised in Compton, California, Lamar embarked on his musical career as a teenager under the stage name K-Dot, releasing a mixtape that garnered local attention and led to his signing with indie record label Top Dawg Entertainment (TDE). He began to gain recognition in 2010, after his first retail release, *Overly Dedicated*. The following year, he independently released his first studio album, *Section.80*, which included his debut single, "Hill PoWeR". By that time, he had amassed a large online following and collaborated with several prominent artists in the hip hop industry, including The Game, Busta Rhymes, and Snoop Dogg. Lamar's major label debut album, *good kid, m.A.A.d city*, was released in 2012 by TDE, Aftermath, and Interscope Records to critical success. It debuted at number two on the US *Billboard 200* chart and was later certified platinum by the Recording Industry Association of America (RIAA). The record contained the top 40 singles "Swimming Pools (Drank)", "Bitch, Don't Kill My Vibe", and "Poetic Justice". His critically acclaimed third album *To Pimp a Butterfly* (2015) comprised funk, soul, and spoken word, debuted atop the charts in the US and the UK, and won the Grammy Award for Best Rap Album at the 58th ceremony. In 2016, Lamar released *Untitled Unmastered*, a collection of unreleased demos that originated during the recording sessions for *Butterfly*. He released his fourth album *Damn* in 2017 to further acclaim; its lead single "Humble" topped the US *Billboard Hot 100* chart. Lamar has received a number of accolades over the course of his career, including twelve Grammy Awards. In early 2013, MTV named him the number one "Hottest MC in the Game", on their annual list.^[1] *Time* named him one of the 100 most influential people in the world in 2016.^[2] Aside from his solo career, Lamar is also known as a member of the West Coast hip hop supergroup Black Hippy, alongside his TDE label-mates and fellow South Los Angeles-based rappers Ab-Soul, Jay Rock, and Schoolboy Q.

Contents [\[hide\]](#)

1 [Edit this page](#)



Lamar in 2016

Born
Kendrick Lamar Duckworth
June 17, 1987 (age 31)
Compton, California, U.S.

Text Analytics: Applications

- Natural Language Understanding: Question Answering

Browse > Natural Language Processing > Question Answering

Passage Sentence
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question
What causes precipitation to fall?

Answer Candidate
gravity

Question Answering

476 papers with code - [Natural Language Processing](#)

Edit

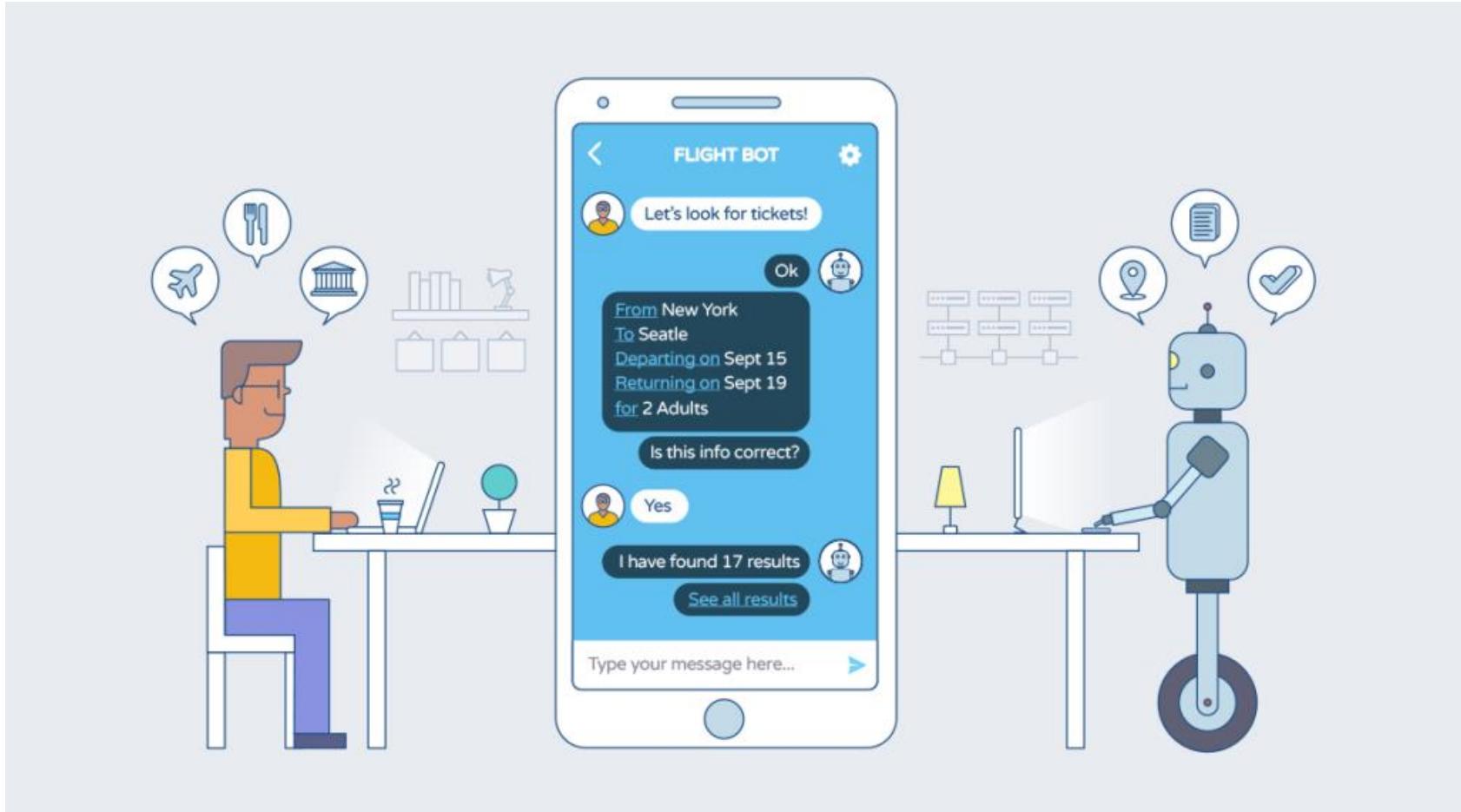
(Image credit: SQuAD)

Leaderboards

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	SQuAD2.0	🏆 Retro-Reader on ALBERT (ensemble)				See all
	SQuAD1.1	🏆 XLNet (single model)	XLNet: Generalized Autoregressive Pretraining for Language Understanding			See all
	SQuAD1.1 dev	🏆 T5-11B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			See all
	WikiQA	🏆 TANDA-RoBERTa (ASNQ, WikiQA)	TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection			See all
	CNN / Daily Mail	🏆 GA+MAGE (32)	Linguistic Knowledge as Memory for Recurrent Neural Networks			See all
	Quora Question Pairs	🏆 ALBERT	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations			See all
	SQuAD2.0 dev	🏆 RoBERTa (no data aug)	RoBERTa: A Robustly Optimized BERT Pretraining Approach			See all

Text Analytics: Applications

- Doing Conversation like Human Beings: ChatBot (Dialogue system)



<https://chatbotslife.com/chatbots-are-the-future-of-marketing-31fd285f37d9>

Text Analytics: Challenges

- Challenges

- ✓ High number of possible “dimensions” (word, phrases, etc.)

List of dictionaries by number of words

From Wikipedia, the free encyclopedia

This list is incomplete; you can help by expanding it.

This is a list of dictionaries considered authoritative or complete by approximate number of total words, or headwords, included. These figures do not take account of entries with senses for different word classes (such as noun and adjective) and homographs. Although it is possible to count the number of entries in a dictionary, it is not possible to count the number of words in a language.^{[1][2]} In compiling a dictionary, a lexicographer decides whether the evidence of use is sufficient to justify an entry in the dictionary. This decision is not the same as determining whether the word exists.^[citation needed]

Language	Approx. no. of words
Korean	1,100,373
Turkish	616,767
Swedish	600,000
Icelandic	560,000
Japanese	500,000
Lithuanian	500,000
Norwegian	500,000
English	470,000
Dutch	400,000
English	350,000
German	330,000
Gujarati	281,377
Italian	260,000
Czech	250,000
Belarusian	223,000
English	207,016
Danish	200,000
English	171,476



Text Analytics: Challenges

- Challenges
 - ✓ Complex and subtle relationship between concepts in texts

“장명준은 즐겁게 오버워치를
하다가 지도교수에게 들켰다 ”

“강필성 교수는 우연히 들른
신공학관 220호에서 게임을
하는 한 학생을 목격했다 ”

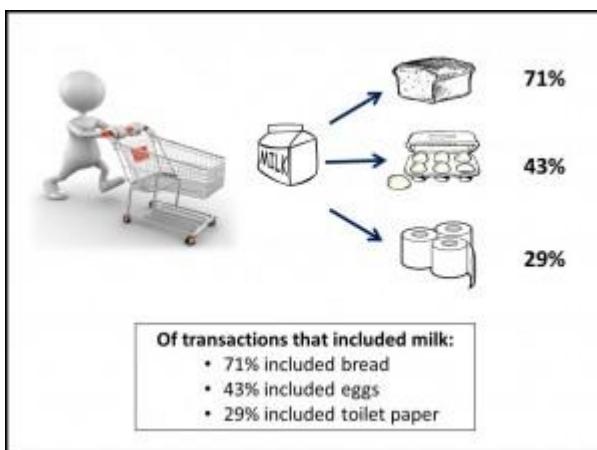


Text Analytics: Challenges

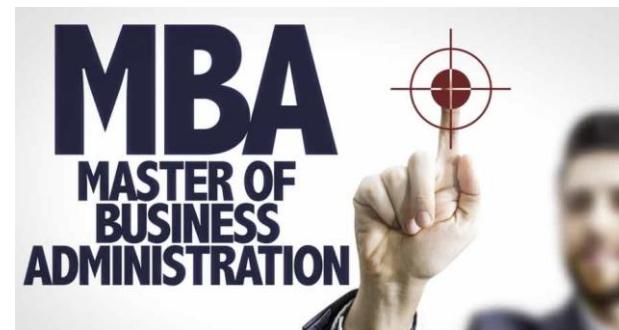
- Challenges
 - ✓ Ambiguity and context sensitivity
 - automobile = car = vehicle = Hyundai



vs.



vs.



Text Analytics: Text Structures

- Structure of text data

Data Mining / Knowledge Discovery



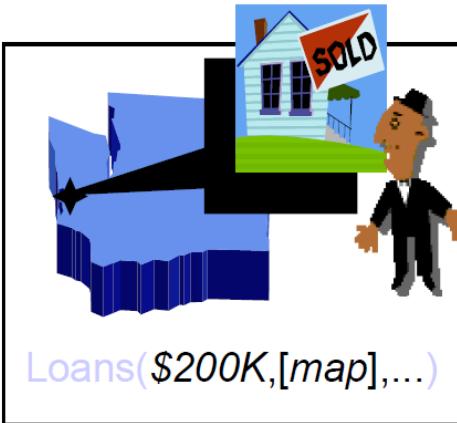
Structured Data

Multimedia

Free Text

Hypertext

HomeLoan (
Loanee: *Frank Rizzo*
Lender: *MWF*
Agency: *Lake View*
Amount: *\$200,000*
Term: *15 years*)

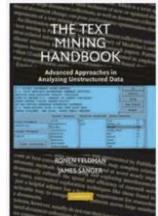


Frank Rizzo bought his home from Lake View Real Estate in 1992. He paid \$200,000 under a 15-year loan from MW Financial.

<a href>Frank Rizzo Bought <a href>this home from <a href>Lake View Real Estate In 1992. <p>...

Text Analytics: Text Structures

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?

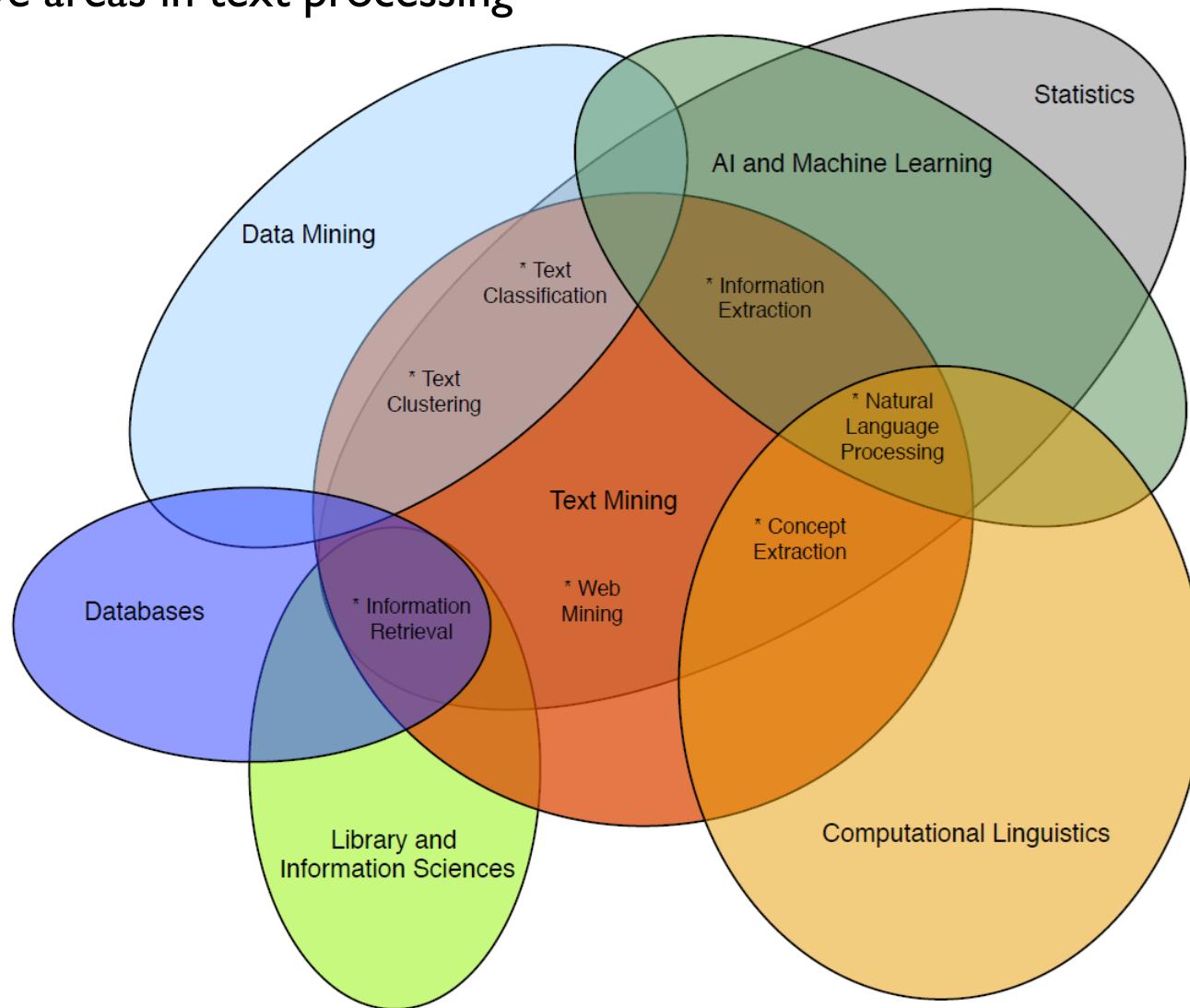
- ✓ Language itself is ambiguous

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

Text Analytics: Areas

Abbott (2013)

- Active areas in text processing



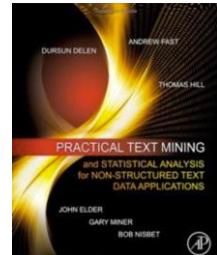
Types of Text Analytics

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 Analytics

Source of text data

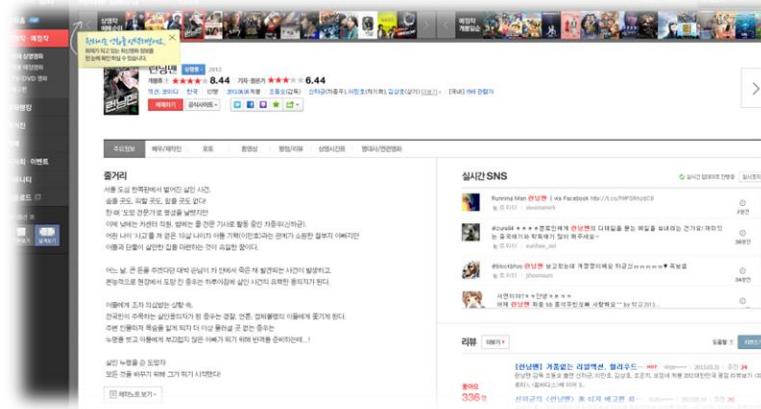
Digital library



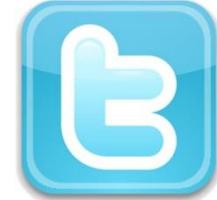
Corporate document archive



Word Wide Web (WWW)



SNS



Step 1:
Decide what to mine
& Collect text data

A Simplified Process of Text Analytics

From unstructured to structured!

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

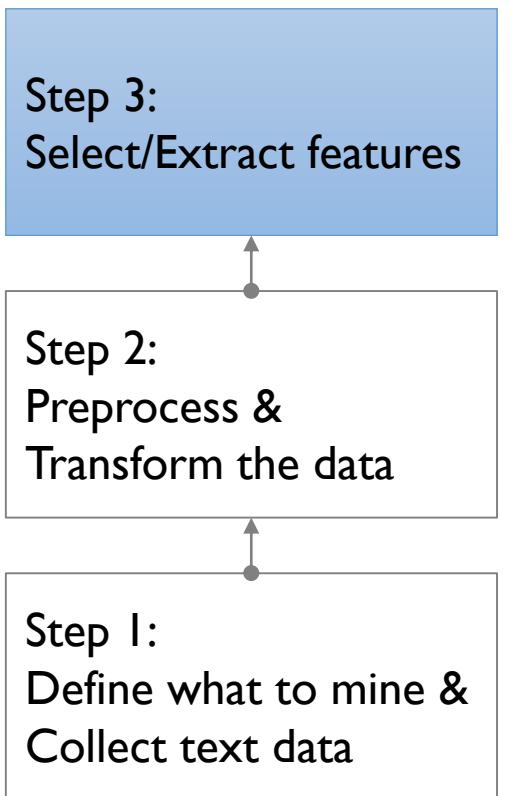
S2: John also likes to watch football game.

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 Analytics



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 Analytics

Step 4:
Algorithm Learning &
Evaluation

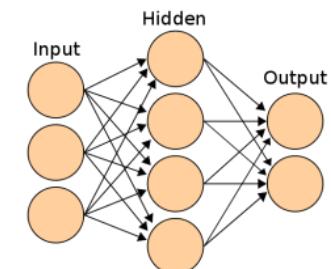
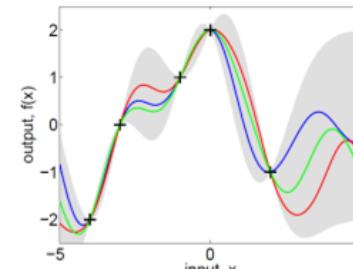
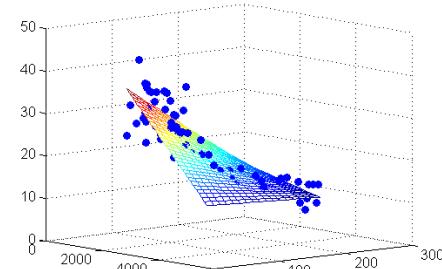
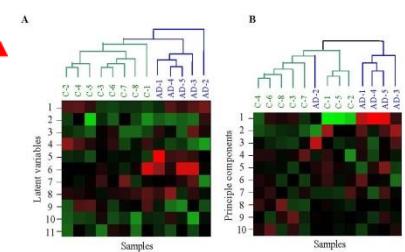
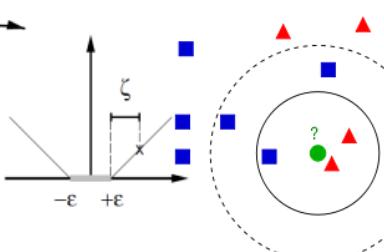
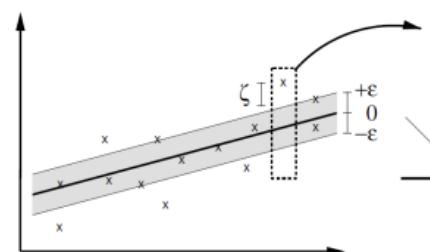
Step 3:
Select/Extract features

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 Analytics: Overview
- 02 TA Process I: Collection & Preprocessing
- 03 TA Process 2: Transformation
- 04 TA Process 3: Dimensionality Reduction
- 05 TA 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")

Decide What to Mine

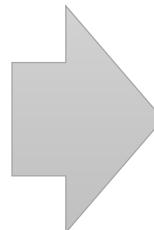
- Open datasets/repositories
 - ✓ The best 25 datasets for NLP (2018.06.07)
 - <https://gengo.ai/datasets/the-best-25-datasets-for-natural-language-processing/>
 - ✓ Alphabetical list of free/public domain datasets with text data for use in Natural Language Processing (NLP)
 - <https://github.com/niderhoff/nlp-datasets>
 - ✓ 50 Free Machine Learning Datasets: Natural Language Processing
 - <https://blog.cambridgespark.com/50-free-machine-learning-datasets-natural-language-processing-d88fb9c5c8da>
 - ✓ 25 Open Datasets for Deep Learning Every Data Scientist Must Work With
 - <https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/>

Text Preprocessing Level 0:Text

- Remove unnecessary information from the collected data

The screenshot shows the Bits section of The New York Times website. The main headline is "Questions for IBM's Watson" by Steve Lohr, dated August 28, 2014, 3:04 PM. The article discusses W. Scott Spangler, a data scientist at IBM, demonstrating how Watson's cognitive technology can visually display connections in scientific data. It also mentions Uber Strengthening Driver Background Checks in India and Facebook's Privacy Policy Reviewed by Hamburg Data Regulars. A sidebar for an upcoming info session at Chicago Booth is visible.

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



The screenshot shows a Microsoft Word document with the title "제목 없음 - 메모장". The text discusses Watson's claim to fame in beating human champions in Jeopardy!, its work in the marketplace, and its role as a cloud service. It also mentions Watson's triumph in its current version, Discovery Advisor, and its potential for strategy and product commercialization. The text concludes with a quote from W. Scott Spangler about the challenge of finding the right question for the scientist to ask.

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

*New York City Public Schools
Should Be Evaluated Based on
Diversity, Not Just Tests, Panel Says*



Mayor Bill de Blasio, flanked by New York City Schools Chancellor Richard Carranza, right, delivers remarks about a diversity plan in Brooklyn. Dave Sanders for The New York Times

By Eliza Shapiro

Feb. 12, 2019



[What you need to know to start the day: [Get New York Today in your inbox.](#)]

A high-level panel commissioned by Mayor Bill de Blasio called on the city to adopt a sweeping measure to address entrenched segregation in education: create diversity targets for all 1,800 schools so that their population reflects the racial and economic makeup of the surrounding areas.

Over the next five years, the panel recommended, elementary and middle schools should reflect the racial makeup of their local school district, and high schools should look as much like their local borough as possible, in terms of race, income level, disability and proficiency in English.

New York's schools have become increasingly divided along racial lines over the last two decades, and the city is currently home to one of the most segregated urban public school systems in America.

Mr. de Blasio, now in his second term, ran on a promise to reduce inequality in all aspects of city life. But, when it comes to school segregation, he has been stymied by the same quandary that previous mayors faced: How to redistribute resources so that black and Hispanic students have more access to high-quality schools without alienating middle class, white families.

Text Preprocessing Level I: 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)

Text Preprocessing Level I: Sentence

- Sentence Splitting ([페이지 링크](#))

강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다. 회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 쉐쉑버거 골목길로 쭉 올라가야 하는데 다들 쉐쉑버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습. 강남 토끼정은 4층 건물 독채로 이루어져 있습니다. 역시 토끼정 본점 담죠? ㅎㅅㅎ 건물은 크지만 간판이 없기 때문에 지나칠 수 있으니 조심하세요 강남 토끼정의 내부 인테리어. 평일 저녁이었지만 강남역 맛집 답게 사람들이 많았어요 전체적으로 편안하고 아늑한 공간으로 꾸며져 있었습니다 ㅎㅎ 한 가지 아쉬웠던 건 조명이 너무 어두워 눈이 침침했던.. 저희는 3층에 자리를 잡고 음식을 주문했습니다. 총 5명이서 먹고 싶은 음식 하나씩 골라 다양하게 주문했어요 첫 번째 준비된 메뉴는 토끼정 고로케와 깻잎불고기 사라다를 듬뿍 올려 먹는 맛있는 밥입니다. 여러가지 메뉴를 한 번에 시키면 준비되는 메뉴부터 가져다 주더라구요, 토끼정 고로케 금방 튀겨져 나와 같은 바삭하고 속은 촉촉해 맛있었어요! 깻잎불고기 사라다는 불고기, 양배추, 버섯을 볶아 깻잎을 듬뿍 올리고 우엉 튀김을 곁들여 밥이랑 함께 먹는 메뉴입니다. 사실 전 고기를 안 먹어서 무슨 맛인지 모르겠지만.. 다들 엄청 잘 드셨습니다ㅋㅋ 이건 제가 시킨 축축한 고로케와 크림스튜우동. 강남 토끼정에서 먹은 음식 중에 이게 제일 맛있었어요!!! 크림소스를 원래 좋아하기도 하지만, 느끼하지 않게 부드럽고 달달한 스튜와 풀깃한 우동면이 너무 잘 어울려 계속 손이 가더러구요. 사진을 보니 또 먹고 싶습니다 간사이 풍 연어 지라시입니다. 일본 간사이 지방에서 많이 먹는 멸치는 조밥(지라시스사)이라고 하네요. 밑에 와사비 마요밥 위에 연어들이 담겨져 있어 코끝이 찡할 수 있다고 적혀 있는데, 난 와사비 맛 1도 모르겠는데...? 와사비를 안 좋아하는 저는 불행인지 다행인지 연어 지라시를 매우 맛있게 먹었습니다ㅋㅋㅋ 다음 메뉴는 달짝지근한 숯불 갈비 덮밥입니다. 간장 양념에 구운 숯불 갈비에 양파, 깻잎, 달걀 반숙을 터트려 비벼 먹으면 그 맛이 크.. (물론 전 안 먹었지만.. 다른 분들이 그렇다고 하더라구요ㅋㅋㅋㅋㅋㅋㅋ) 마지막 메인 메뉴 양송이 크림수프와 숯불떡갈비 밥입니다. 크림리조또를 베이스로 위에 그루통과 숯불로 구운 떡갈비가 올라가 있어요! 크림스튜 우동 만큼이나 대박 맛있습니다..ㅠㅠㅠㅠㅠㅠ(크림 소스면 다 좋아하는 거 절대 아닙니다ㅋㅋㅋㅋㅋㅋ) 강남 토끼정 요리는 다 맛있지만 크림소스 요리를 참 잘하는 거 같네요 요건 물만 마시기 아쉬워 시킨 뉴자동과 밀키소다 딸기통통! 유자와 자몽의 맛을 함께 느낄 수 있는 뉴자동은 상큼함 그 자체였어요. 하지만 저는 딸기통통 밀키소다가 더 맛있었습니다 ㅎㅎ 밀키소다는 토끼정에서만 만나볼 수 있는 메뉴라고 하니 한 번 드셔보시길 추천하게요!! 강남 토끼정은 강남역 맛집답게 모든 음식들이 대체적으로 맛있었어요! 건물 위지도 강남 대로변에서 조금 떨어져 있어 내부 인테리어처럼 아늑한 느낌도 있었구요 ㅎㅎ 기회가 되면 다들 꼭 들러보세요~ 😊

순번	정답	자바	오픈소스 한국어 처리기	한나눔
1	강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다.			
2	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 쉐쉑버거 골목길로 쭉 올라가야 하는데 다들 쉐쉑버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 쉐쉑버거 골목길로 쭉 올라가야 하는데 다들 쉐쉑버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 쉐쉑버거 골목길로 쭉 올라가야 하는데 다들 쉐쉑버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 쉐쉑버거 골목길로 쭉 올라가야 하는데 다들 쉐쉑버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.
3	다만, 강남 토끼정이 강남 쉐쉑버거 골목길로 쭉 올라가야 하는데 다들 쉐쉑버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.			
4	강남역 맛집 토끼정의 외부 모습.			
5	강남 토끼정은 4층 건물 독채로 이루어져 있습니다.			

Text Preprocessing Level I: Sentence

- Sentence Splitting with Rule-based Model

	구분	전체	정답	정답률
리뷰포함전체	KOALA	17199	14981	87.1
리뷰포함전체	WIGO	17199	16077	93.47
리뷰포함전체	MERGE	17199	16398	95.34
etri 원본	KOALA	1084	1044	96.31
etri 원본	WIGO	1084	1065	98.24
etri 원본	MERGE	1087	1074	99.07
etri. 제거	KOALA	1084	250	23.06
etri. 제거	WIGO	1084	865	79.79
etri. 제거	MERGE	1084	881	81.27
리뷰만	KOALA	15031	13687	91.05
리뷰만	WIGO	15031	14147	94.11
리뷰만	MERGE	15031	14443	96.08

Text Preprocessing Level 2: 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 2: Token

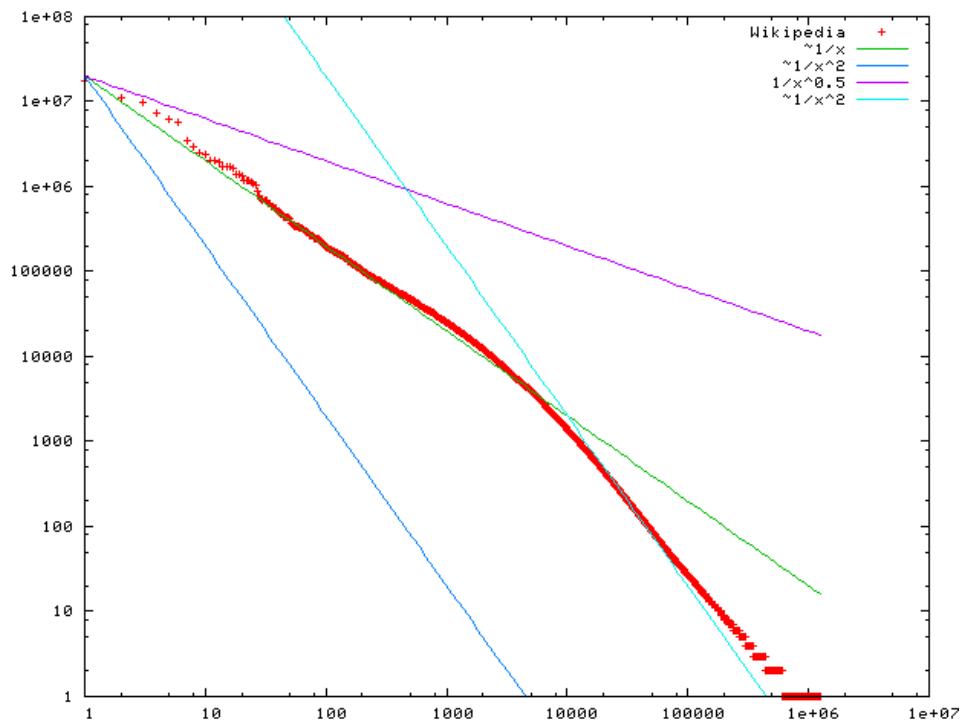
- Power distribution in word frequencies
 - ✓ It is **not true** that more frequently appeared words (tokens) are more important for text mining tasks.

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 2: Token

- 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 2: Token

- **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 2: Token

- 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

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.

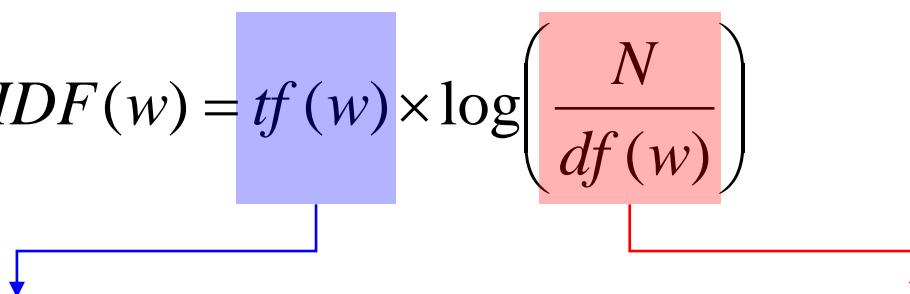
Word	S ₁	S ₂
John	1	1
Likes	2	1
To	1	1
Watch	1	1
Movies	1	0
Also	0	1
Football	0	1
Games	0	1
Mary	1	0
too	1	0



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)

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


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

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

- ✓ 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

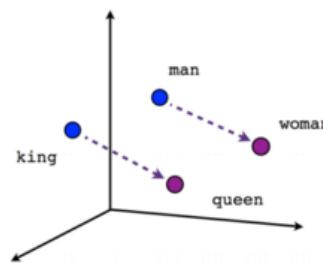
- Word vectors: distributed representation
 - ✓ A parameterized function mapping words in some language to a certain dimensional vectors

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

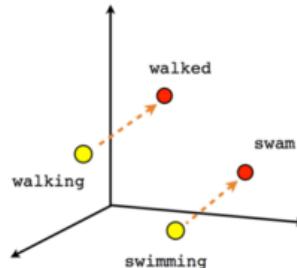
$$W(\text{"cat"}) = (0.2, -0.4, 0.7, \dots)$$

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

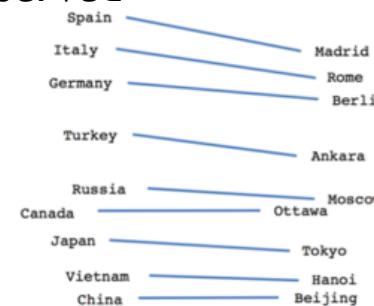
- Interesting feature of word embedding
 - ✓ Semantic relationship between words can be preserved



Male-Female



Verb tense

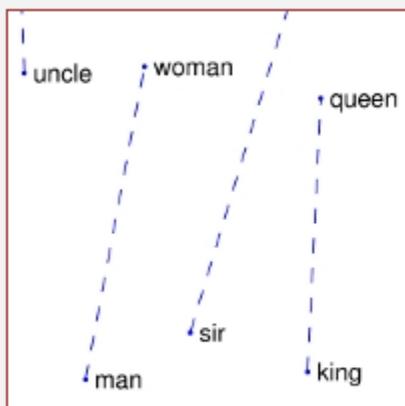


Country-Capital

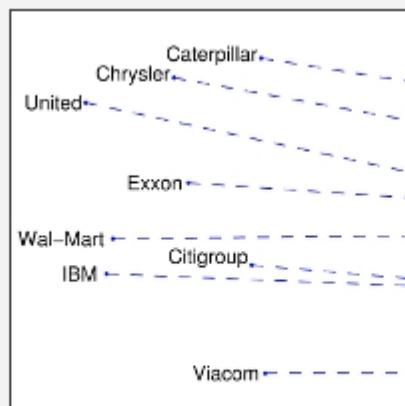
Text Transformation

- Word vectors: distributed representation

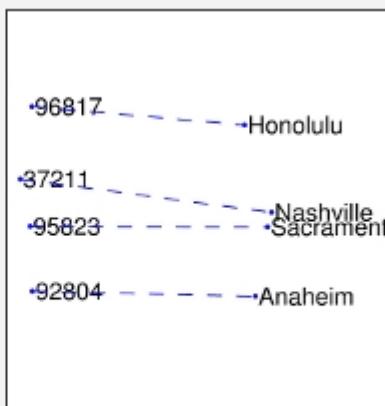
- 0. *frog*
- 1. *frogs*
- 2. *toad*
- 3. *litoria*
- 4. *leptodactylidae*
- 5. *rana*
- 6. *lizard*
- 7. *eleutherodactylus*

3. *litoria*4. *leptodactylidae*5. *rana*7. *eleutherodactylus*

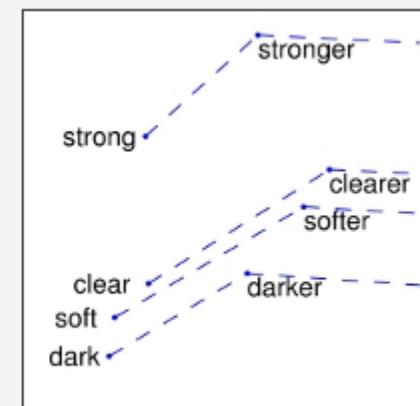
man - woman



company - ceo



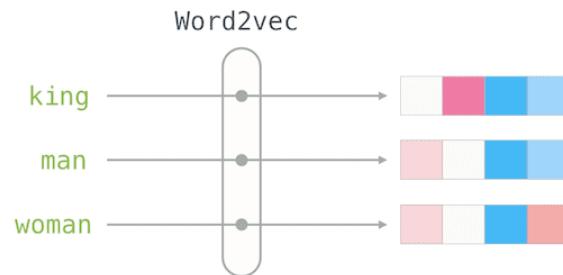
city - zip code



comparative - superlative

Text Transformation

- Pre-trained Word Models



fastText

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

- Pre-trained Language Models



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

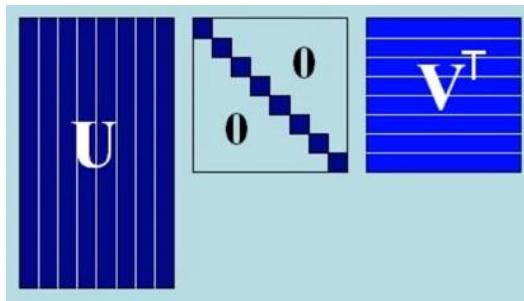
- Feature subset selection
 - ✓ Select only **the best features** for further analysis
 - The most frequent
 - The most informative relative to the all class values, ...
 - ✓ Scoring methods for individual feature (for supervised learning tasks)
 - Information gain: $\sum_{F=W,\bar{W}} P(F) \sum_{C=pos,neg} P(C|F) \log \frac{P(C|F)}{P(C)}$
 - 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

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

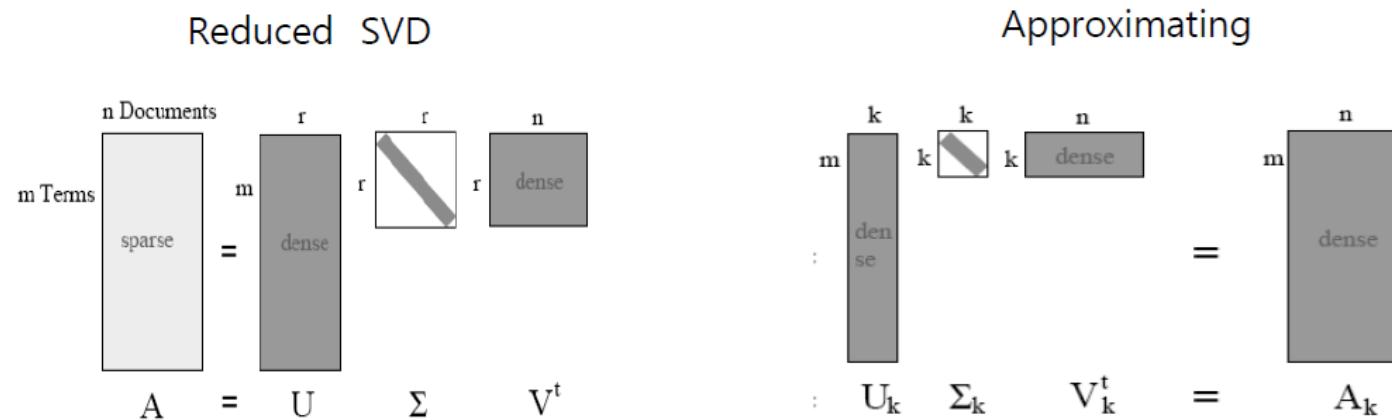


- Rectangular matrix $A(m \times n)$ can be decomposed as $A = U\Sigma V^T$
- All singular vectors of the matrices U and V are orthogonal $U^T U = V^T V = I$
- Eigenvalues of the matrix Σ is positive and sorted in a descending order

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

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

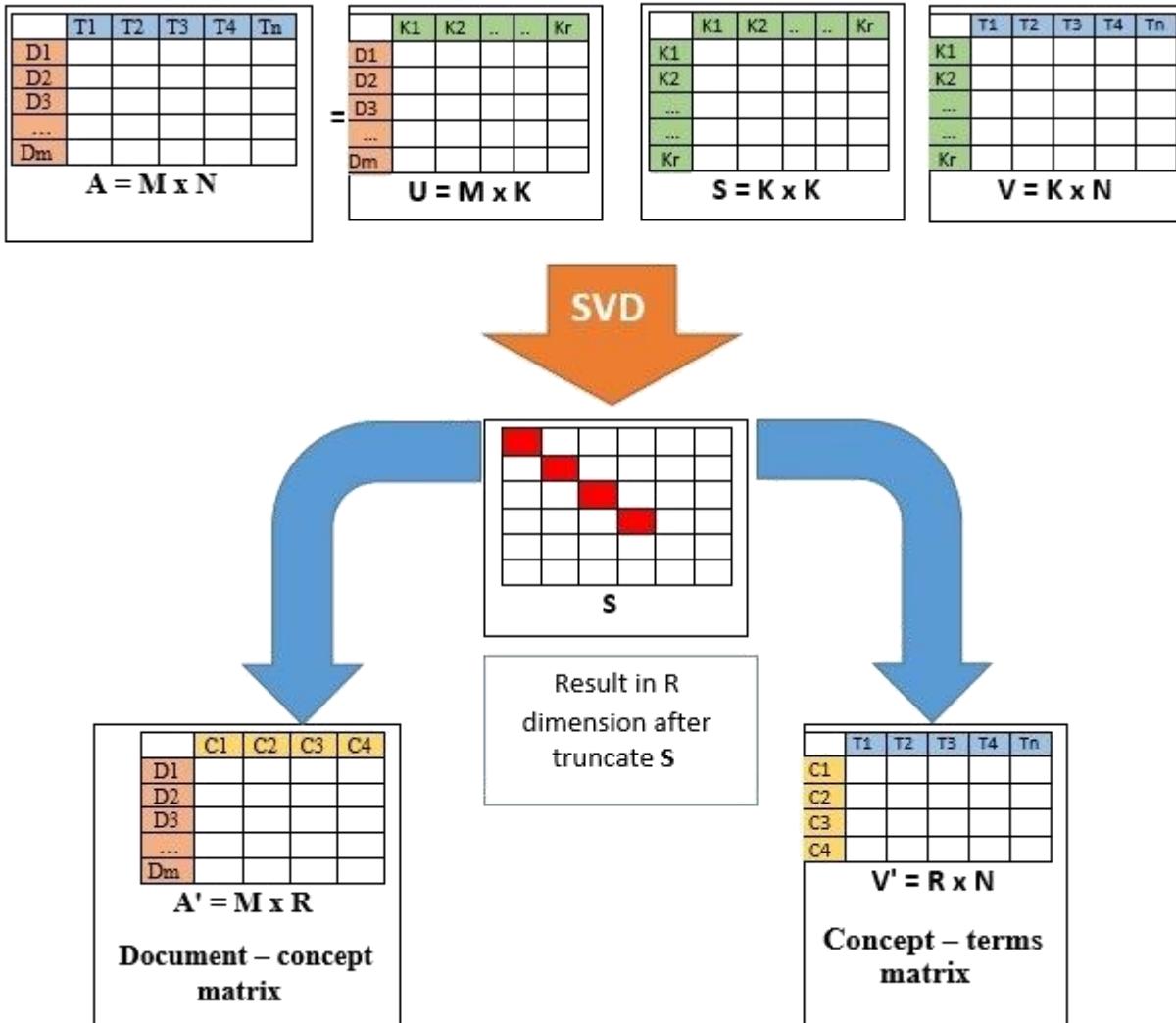
- ✓ Step 2) Multiply the transpose of the matrix U_k

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

- ✓ 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

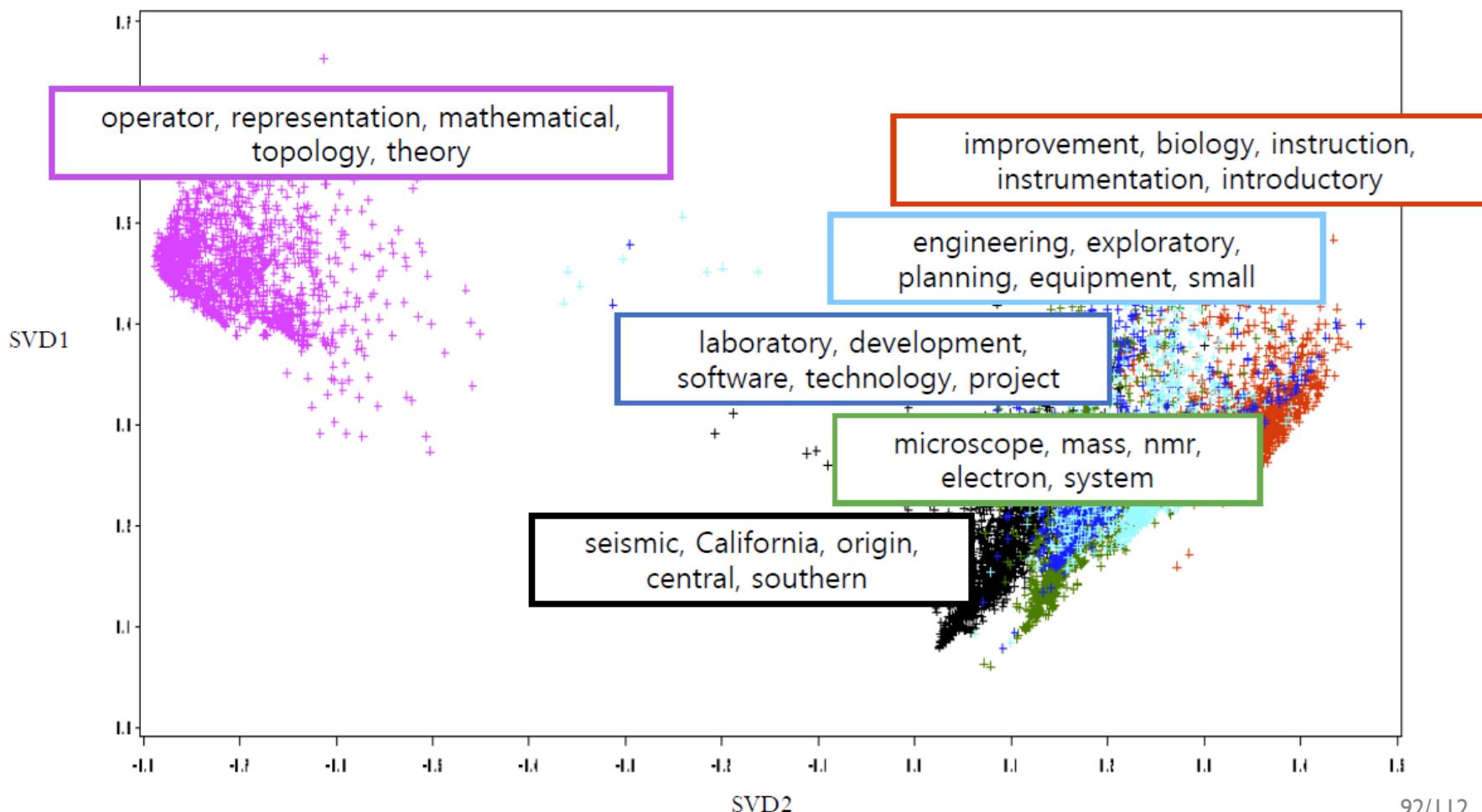
- LSA example

- ✓ 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

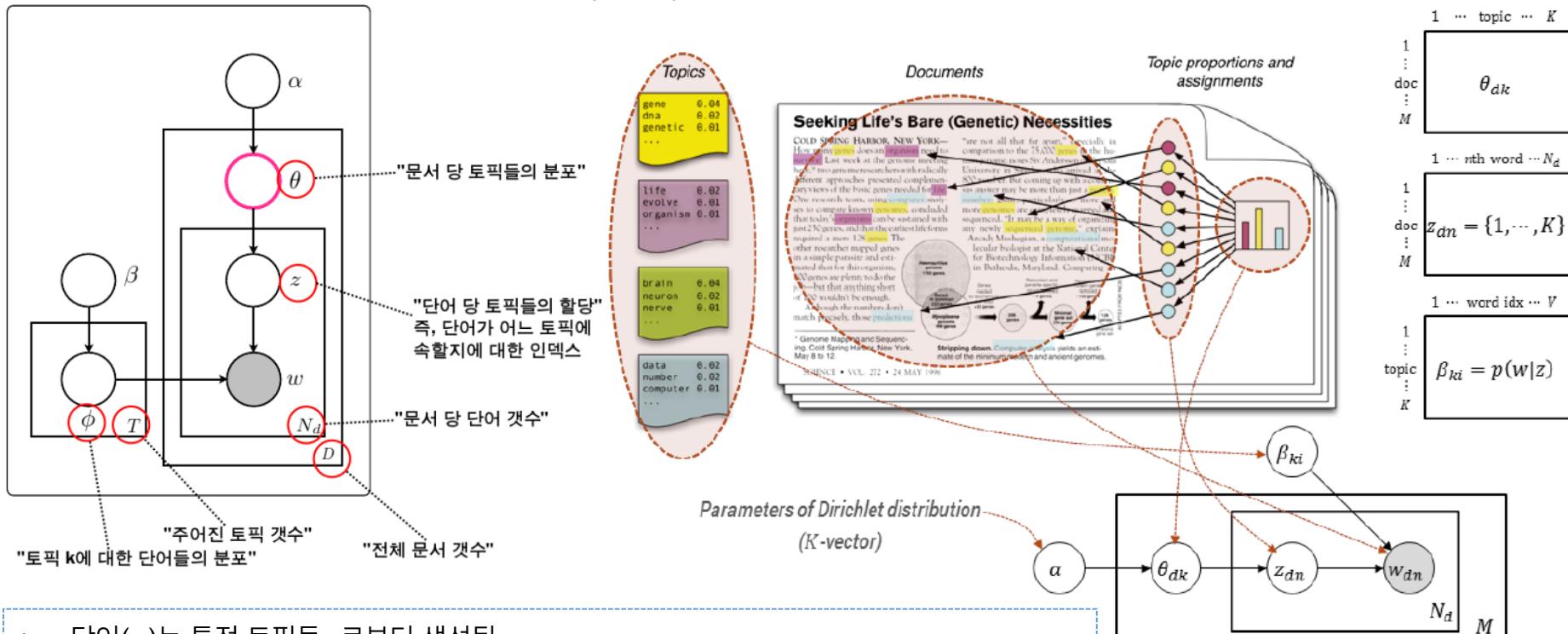
- LSA example
 - ✓ Visualize the project in the reduced 2-D space



Feature Selection/Extraction

- Topic Modeling as a Feature Extractor

- ✓ Latent Dirichlet Allocation (LDA)



- 단어(w)는 특정 토픽들 z 로부터 생성됨
- 해당 문서가 어떤 토픽 비율(topic proportion, θ)을 가질 것인지는 파라미터가 α 인 Dirichlet distribution에 의해 결정됨
- w 만 실제로 관찰할 수 있는 값이고 θ, z, Φ 는 숨겨진 값임
- 문서 생성 프로세스는 먼저 다항분포 θ 로부터 토픽이 나타날 확률 θ 를 추출하고 이를 바탕으로 단어가 나타날 확률인 w 를 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)

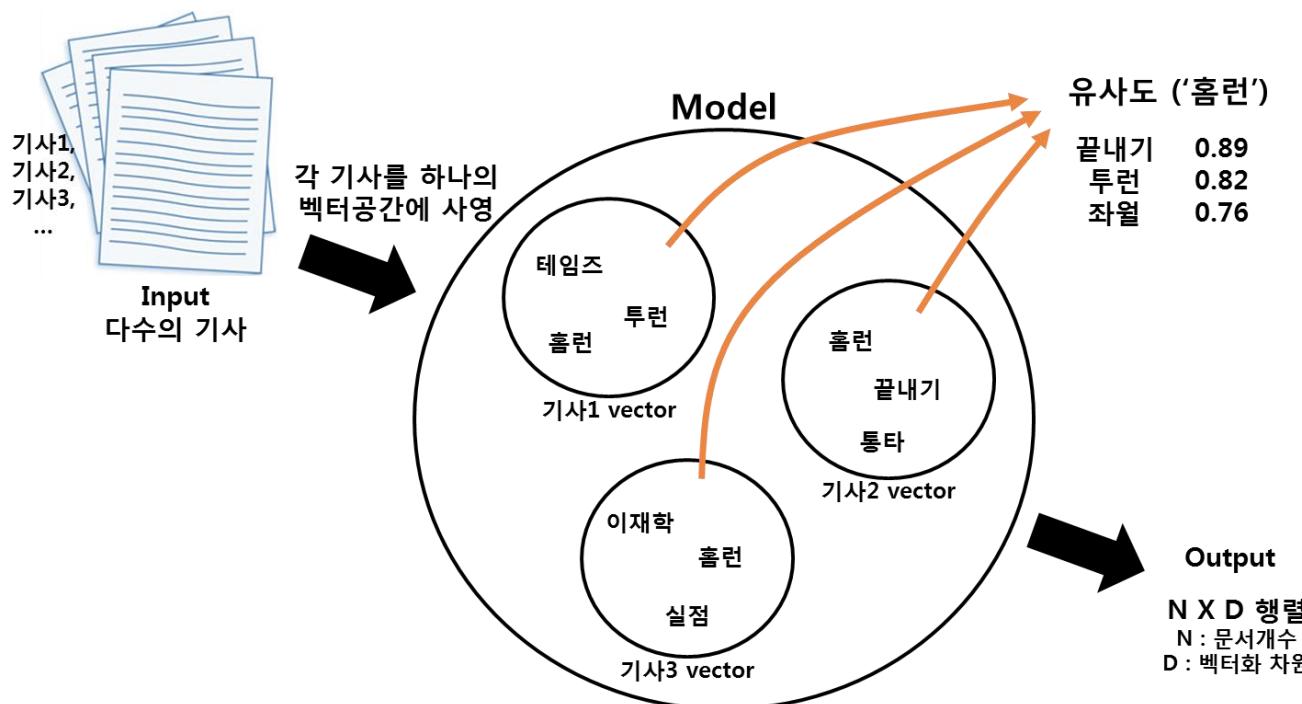
	Topic 1	Topic 2	Topic 3	...	Topic K	Sum
Doc 1	0.20	0.50	0.10	...	0.10	1
Doc 2	0.50	0.02	0.01	...	0.40	1
Doc 3	0.05	0.12	0.48	...	0.15	1
...	1
Doc N	0.14	0.25	0.33	...	0.14	1

(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

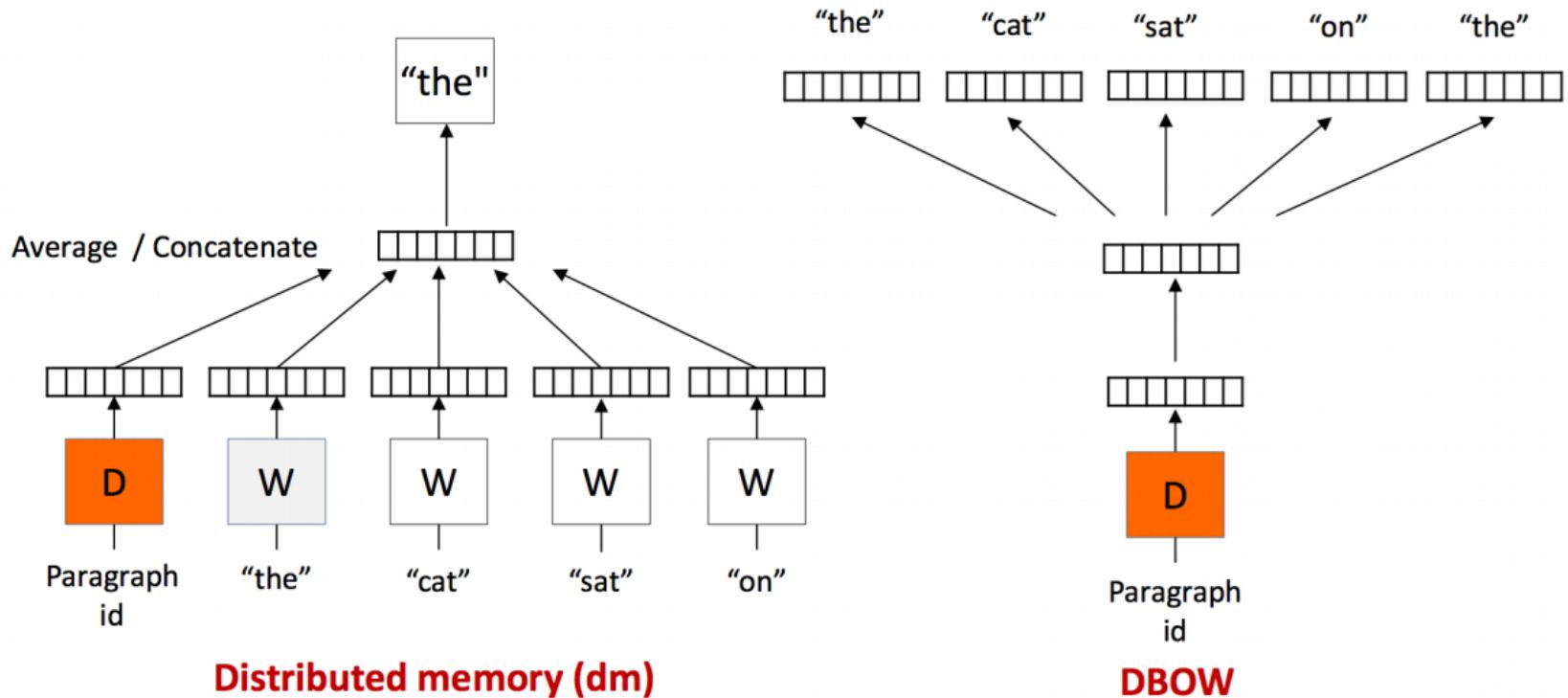
Feature Selection/Extraction

- 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



AGENDA

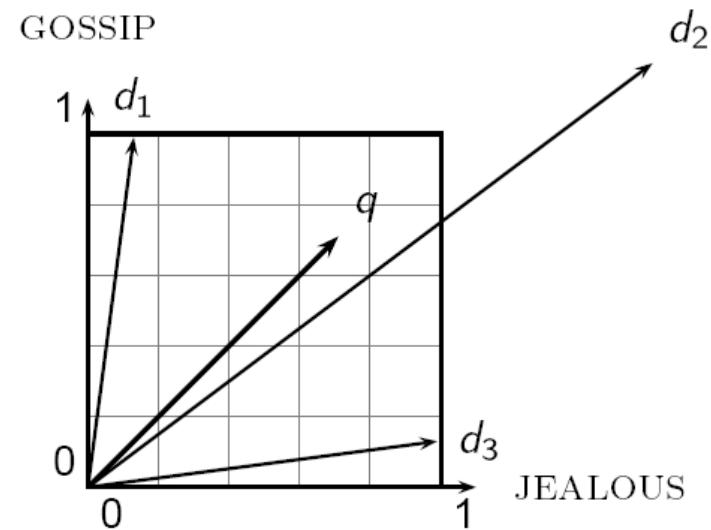
- 01 Text Mining Overview
- 02 TM Process I: Collection & Preprocessing
- 03 TM Process 2: Transformation
- 04 TM Process 3: Dimensionality Reduction
- 05 TM Process 4: Learning & Evaluation

Similarity Between Documents

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

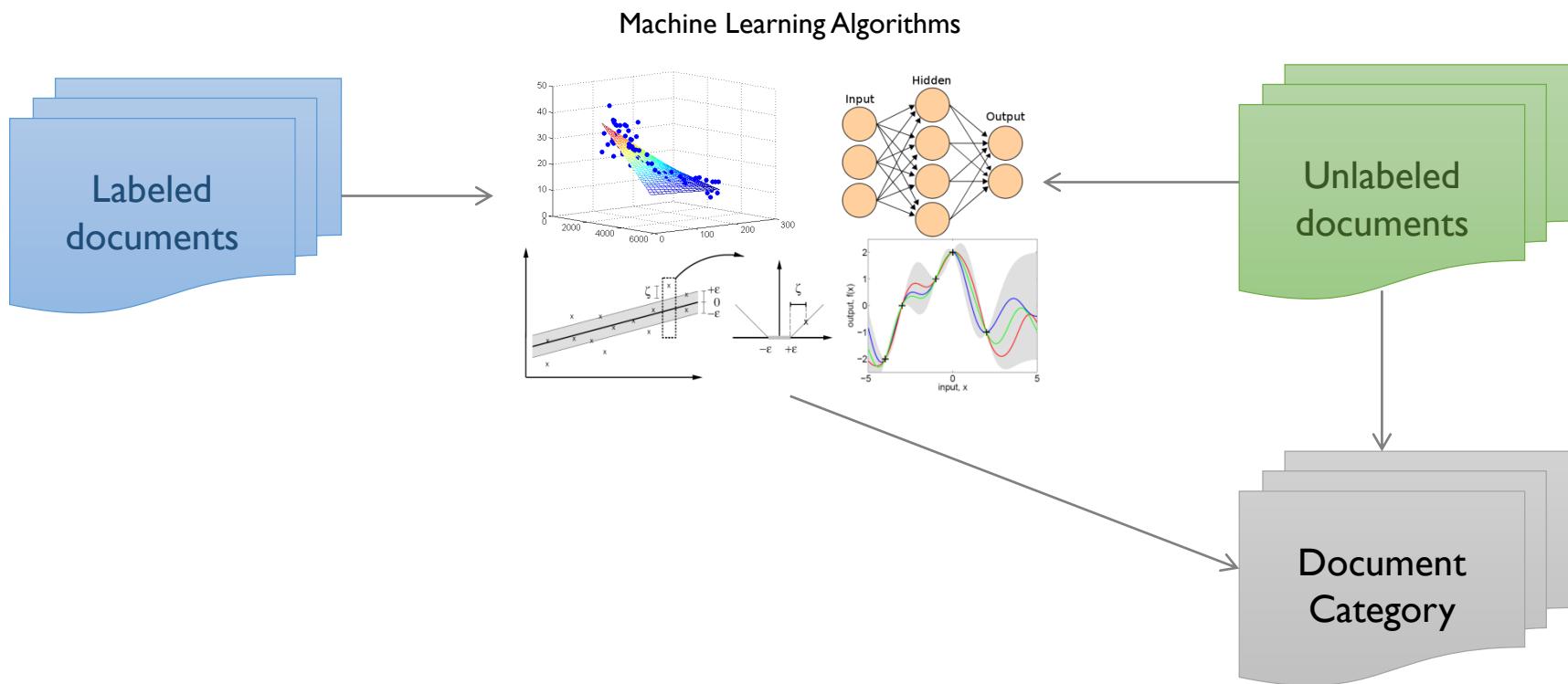
Doc.	Word 1	Word 2	Word 3
Document 1	1	1	1
Document 2	3	3	3
Document 3	0	2	0

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



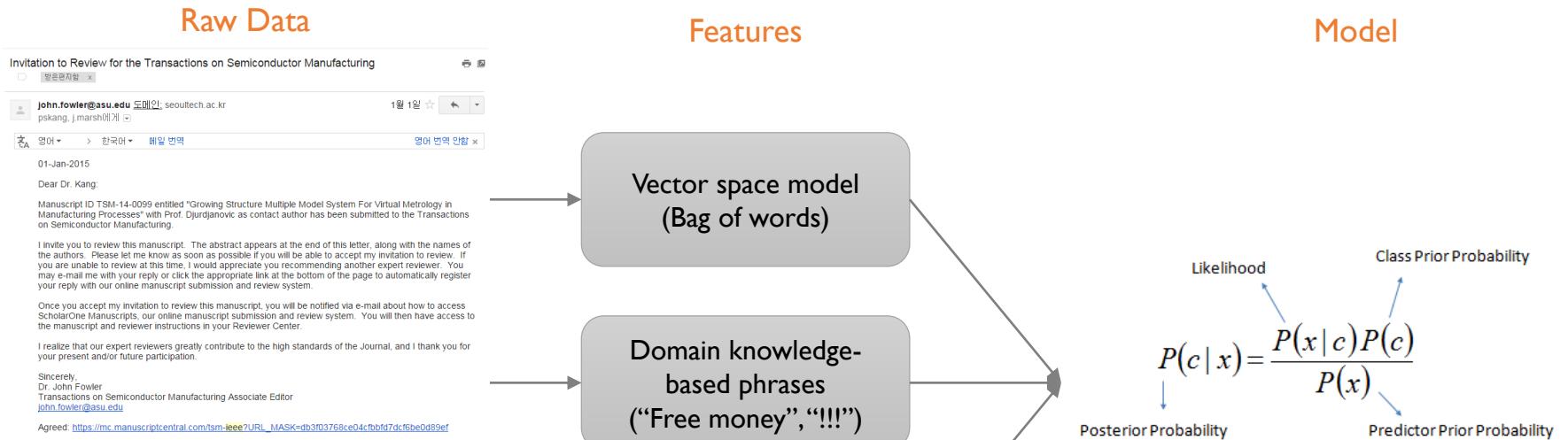
Learning Task I: Classification

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



Learning Task I: Classification

- Spam filtering

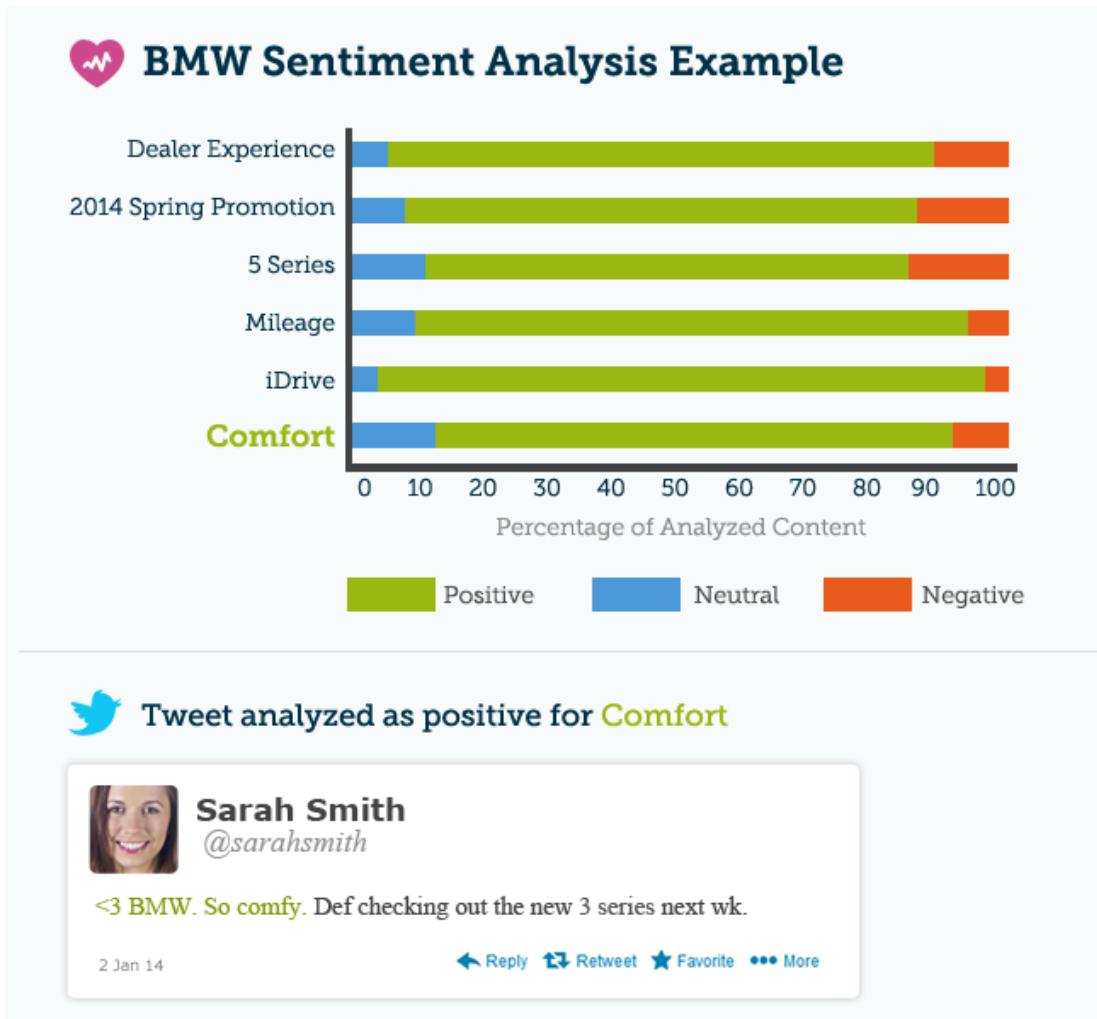


$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

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

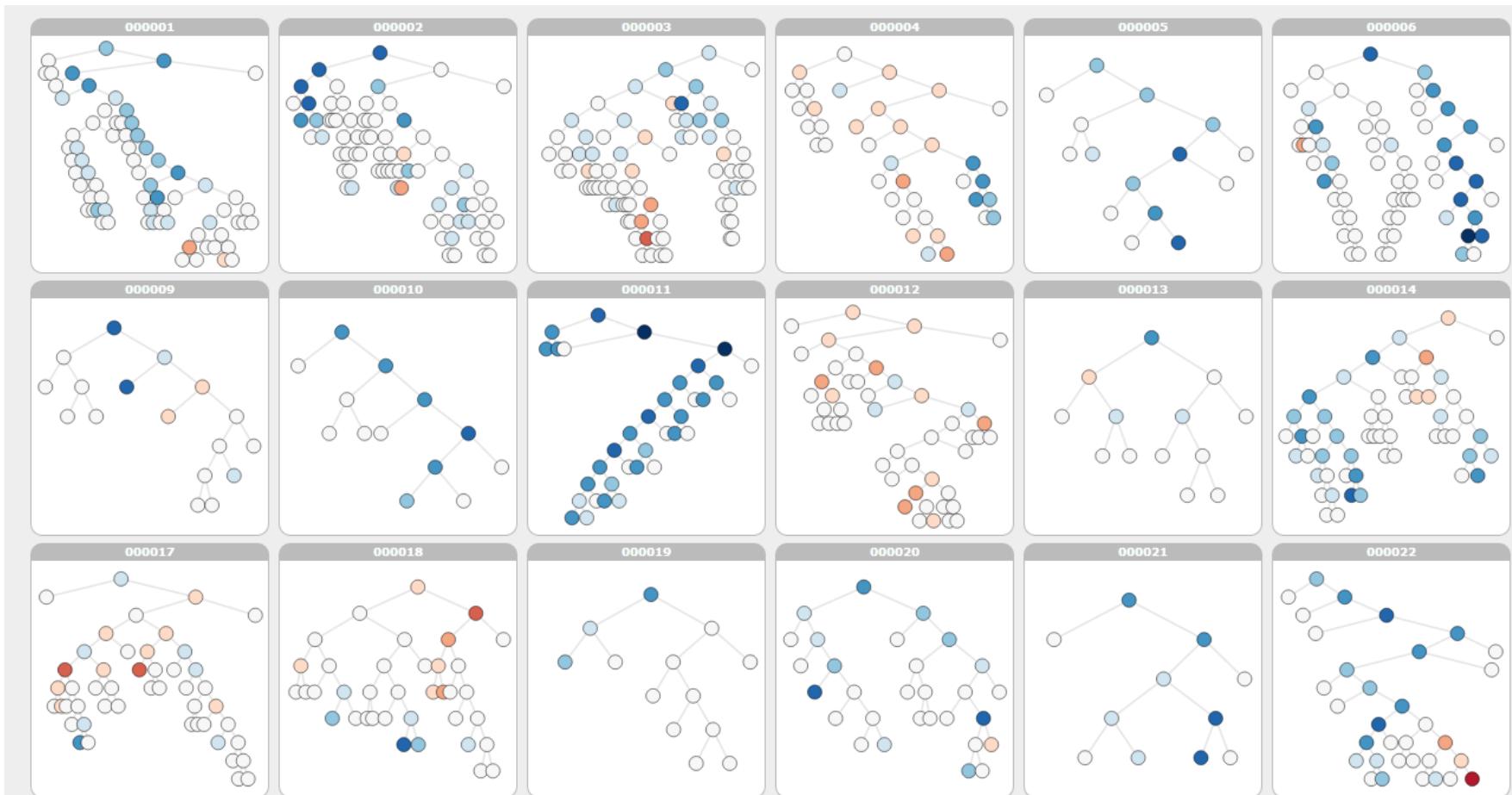


Learning Task I: Classification

Socher et al. (2013)

- Sentiment Analysis

✓ Sentiment tree bank @Stanford NLP Lab

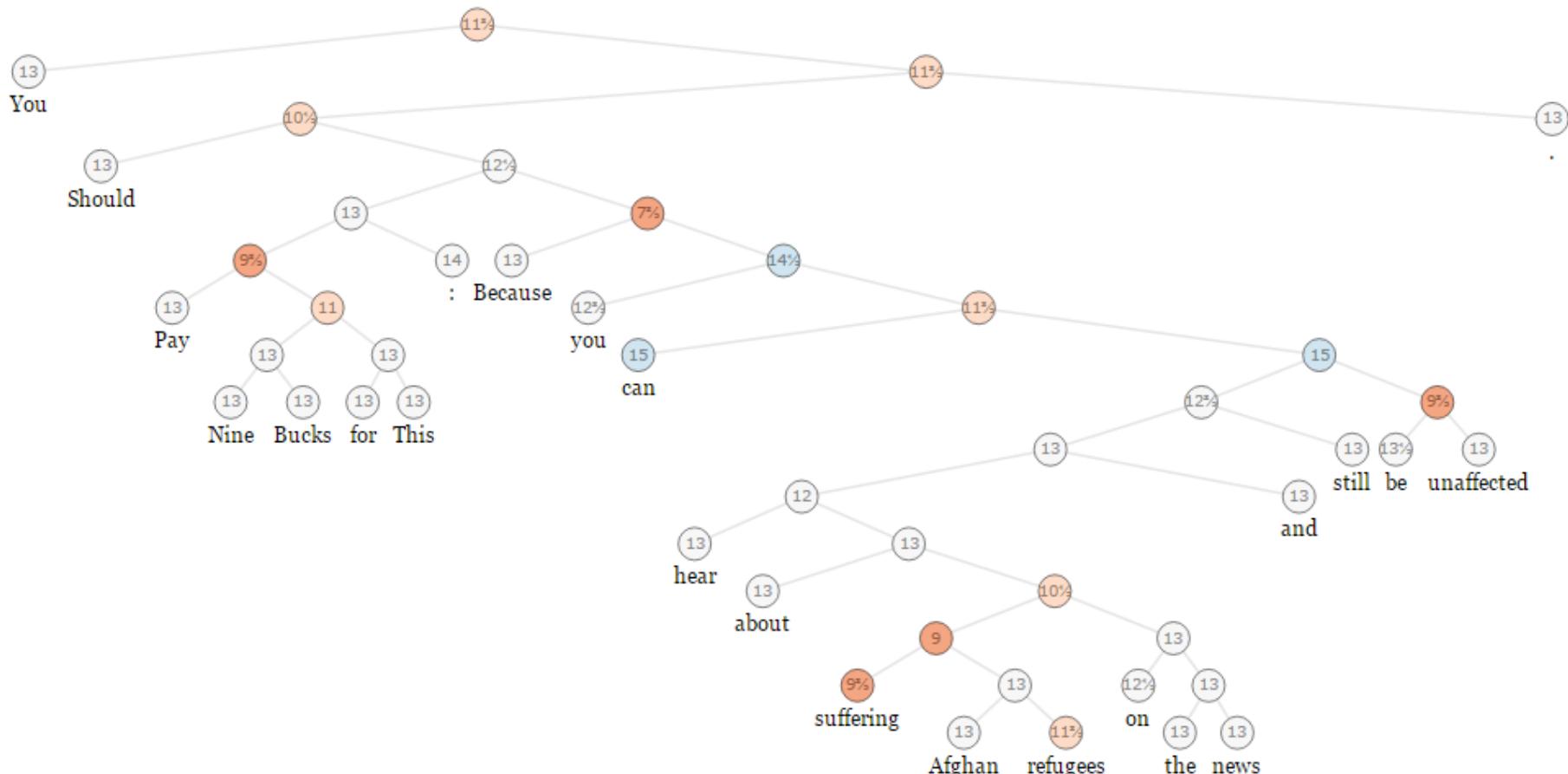


Learning Task I: Classification

Socher et al. (2013)

- Sentiment Analysis

- ✓ Sentiment tree bank @Stanford NLP Lab



Learning Task 2: Clustering

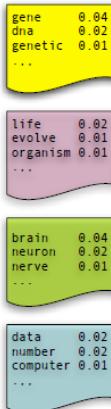
- Document Clustering & Visualization

- ✓ 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 quality management	2014	Interorganizational relationships are an integral part of an organization's decision-making process. This study explores the boundary conditions of Christensen's theory of disruption. Using a two-stage procurement process, we explore the boundary conditions of Christensen's theory of disruption. Using a two-stage procurement process, we explore the boundary conditions of Christensen's theory of disruption.
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 explore the boundary conditions of Christensen's theory of disruption.
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 activities. We propose a two-stage procurement process to address the problem.
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.
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 and consequences of supply chain quality integration are identified.
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 explores the relationship between IT-enabled enterprise agility and organizational culture.
IEEE Transactions on Engineering Management	Accessing external technological knowledge	2014	This study elucidates how a firm accesses external technological knowledge (TK) and how it is used to support innovation. Two research approaches are used to complement each other.
IEEE Transactions on Engineering Management	Investigation on centralized and decentralized design for supply chain	2014	This study presents an investigation of the relationship between internalization of knowledge and design structure. The results show that specialization within groups yields production gains.
IEEE Transactions on Engineering Management	ISO 9000 internalization and organizational structure	2014	This study presents an investigation of the relationship between internalization of knowledge and design structure. The results show that specialization within groups yields production gains.
IEEE Transactions on Engineering Management	Knowledge specialization in ph.D.	2014	Studie
IEEE Transactions on Engineering Management	Overlapping-based design structure	2014	To gain competitive leverage, firms that design and develop complex products see overlapping-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 sustain	2014	In the last decade, industrial research for sustainable development and environmental protection has become an essential consideration while designing a product.
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.
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 a 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 have led to a need for more efficient and effective management of innovation.
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.
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 the openness of the innovation ecosystem.
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	Buyout options allow a bidder to acquire an auctioned product immediately at a low price.
IEEE Transactions on Engineering Management	Contracting a development supplier in t	2014	The development of important components for a new product is frequently outsourced to a development supplier.
IEEE Transactions on Engineering Management	Exploration, exploitation, and growth i	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.

Topics



Features

Documents

Seeking Life's Bare (Genetic) Necessities

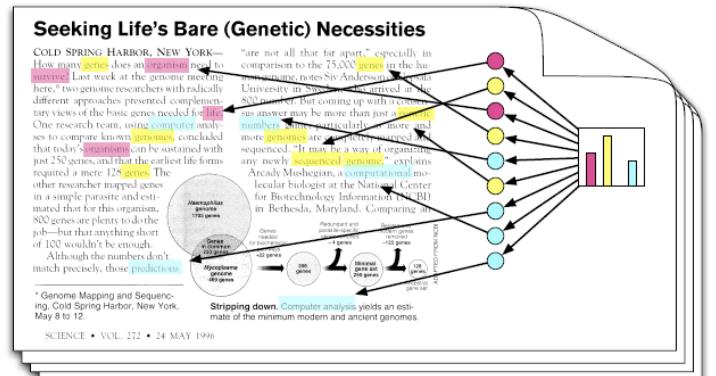
COLD SPRING HARBOR, NEW YORK—“Are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a molecular biologist at Cornell University in Ithaca, New York. But coming up with a list of the bare minimum genes that can support life is more than just a numbers game. “It’s particularly interesting to know what genes are really needed,” says Andersson. “More and more genomes are being sequenced.” It may be a way of organizing any newly sequenced genome,” explains Andrei Mischenko, a molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing 800 genomes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



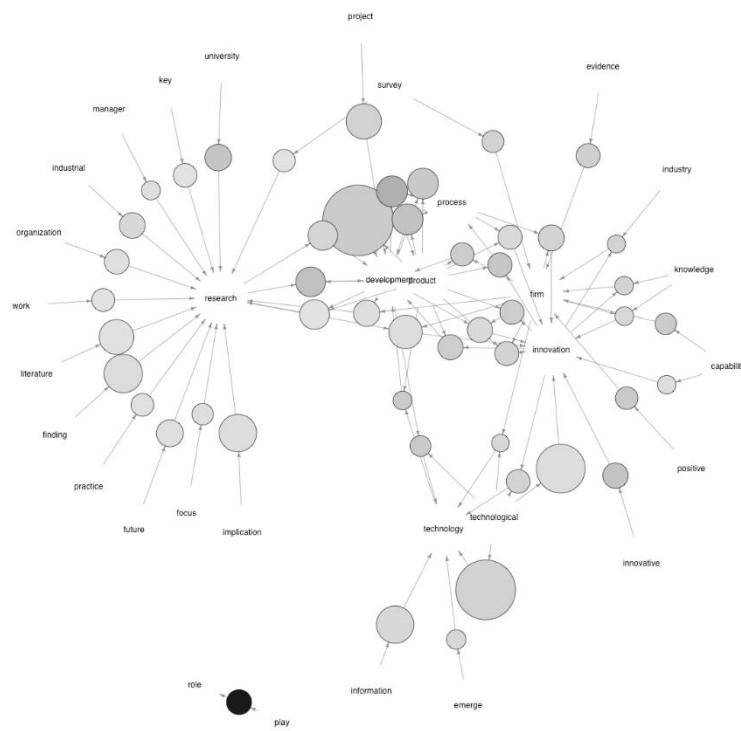
- 8,850 articles from 11 Journals for the recent 10 years
- 21,434 terms after preprocessing

- 50 topics from Latent Dirichlet Allocation (LDA)

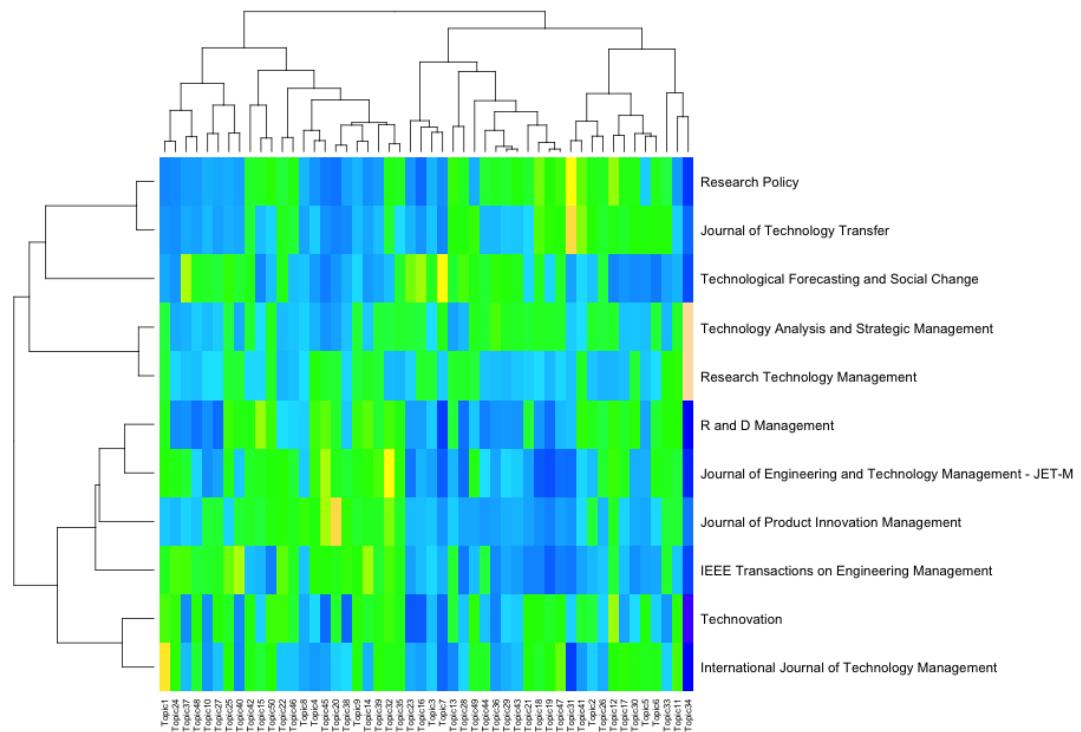
Learning Task 2: Clustering

- Document Clustering & Visualization

Keywords association

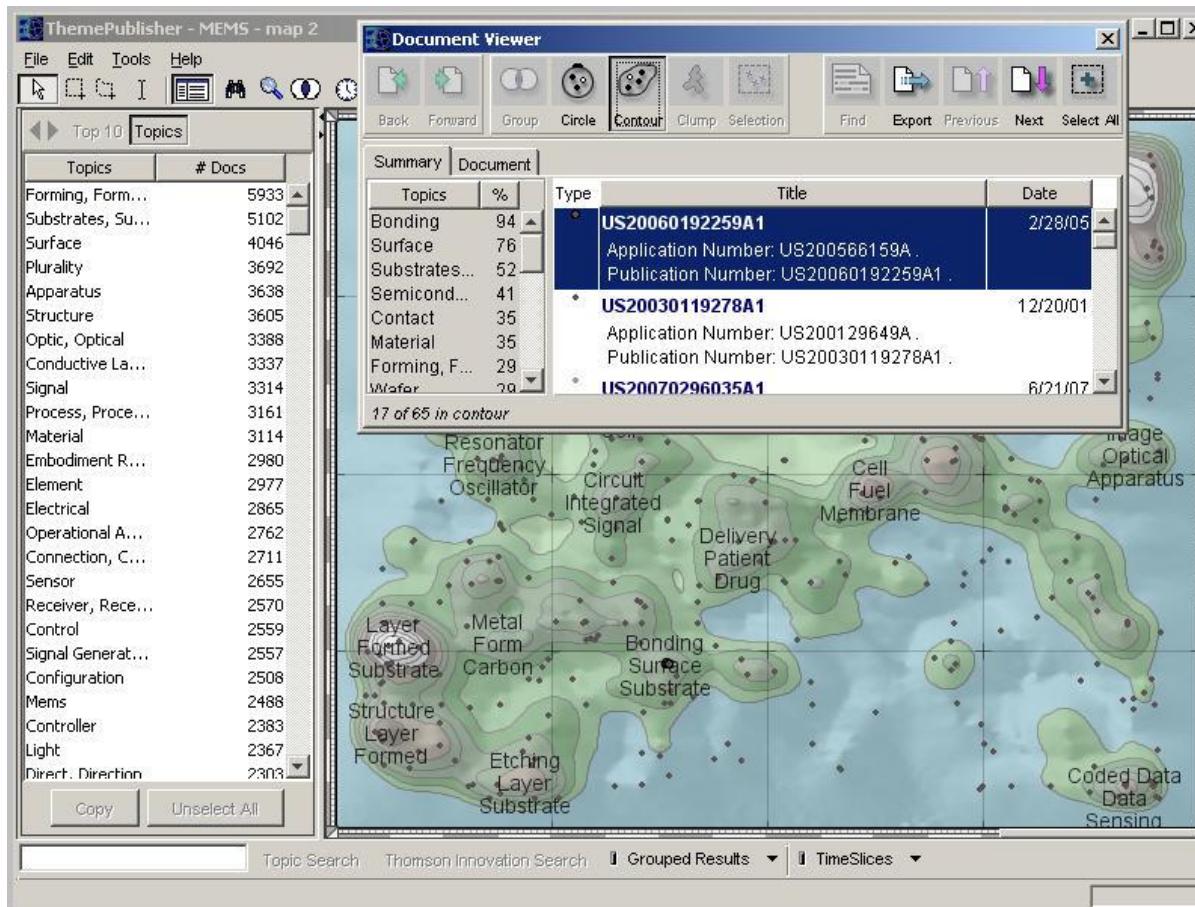


Journal/Topic Clustering



Learning Task 2: Clustering

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



Learning Task 3: Information Extraction/Retrieval

Yang et al. (2013)

- Information extraction/retrieval

- ✓ Find useful information from text databases
- ✓ Examples: Question Answering

The screenshot shows the SQuAD 2.0 website interface. At the top, there's a purple header with the text "SQuAD 2.0" and "The Stanford Question Answering Dataset". Below the header, a message says "Predictions by r-net+ (ensemble) (Microsoft Research Asia)" with an F1 score of 85.4. The main content area is titled "Super_Bowl_50" and "The Stanford Question Answering Dataset". It displays several questions and their answers:

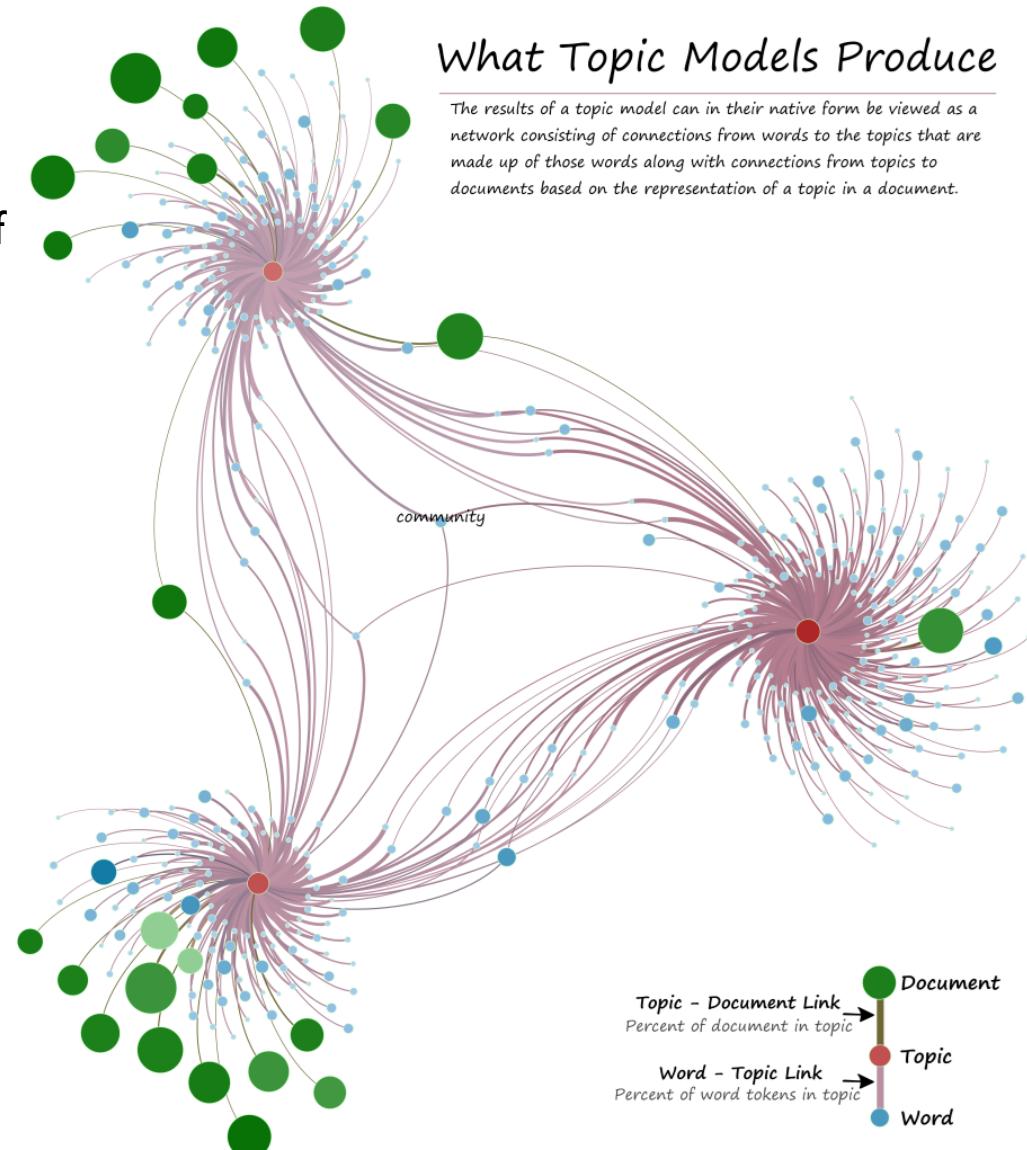
- Which NFL team represented the AFC at Super Bowl 50?**
Ground Truth Answers: Denver Broncos | Denver Broncos | Denver Broncos |
Prediction: Denver Broncos
- Which NFL team represented the NFC at Super Bowl 50?**
Ground Truth Answers: Carolina Panthers | Carolina Panthers | Carolina Panthers |
Prediction: Carolina Panthers
- Where did Super Bowl 50 take place?**
Ground Truth Answers: Santa Clara, California | Levi's Stadium | Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.
Prediction: Levi's Stadium in the San Francisco Bay Area at Santa Clara, California
- Which NFL team won Super Bowl 50?**
Ground Truth Answers: Denver Broncos | Denver Broncos | Denver Broncos |
Prediction: Denver Broncos

At the bottom left, there's a link: [https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Super_Bowl_50.html?model=r-net%20\(ensemble\)%20\(Microsoft%20Research%20Asia\)&version=1.1](https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Super_Bowl_50.html?model=r-net%20(ensemble)%20(Microsoft%20Research%20Asia)&version=1.1)

Learning Task 3: Information Extraction/Retrieval

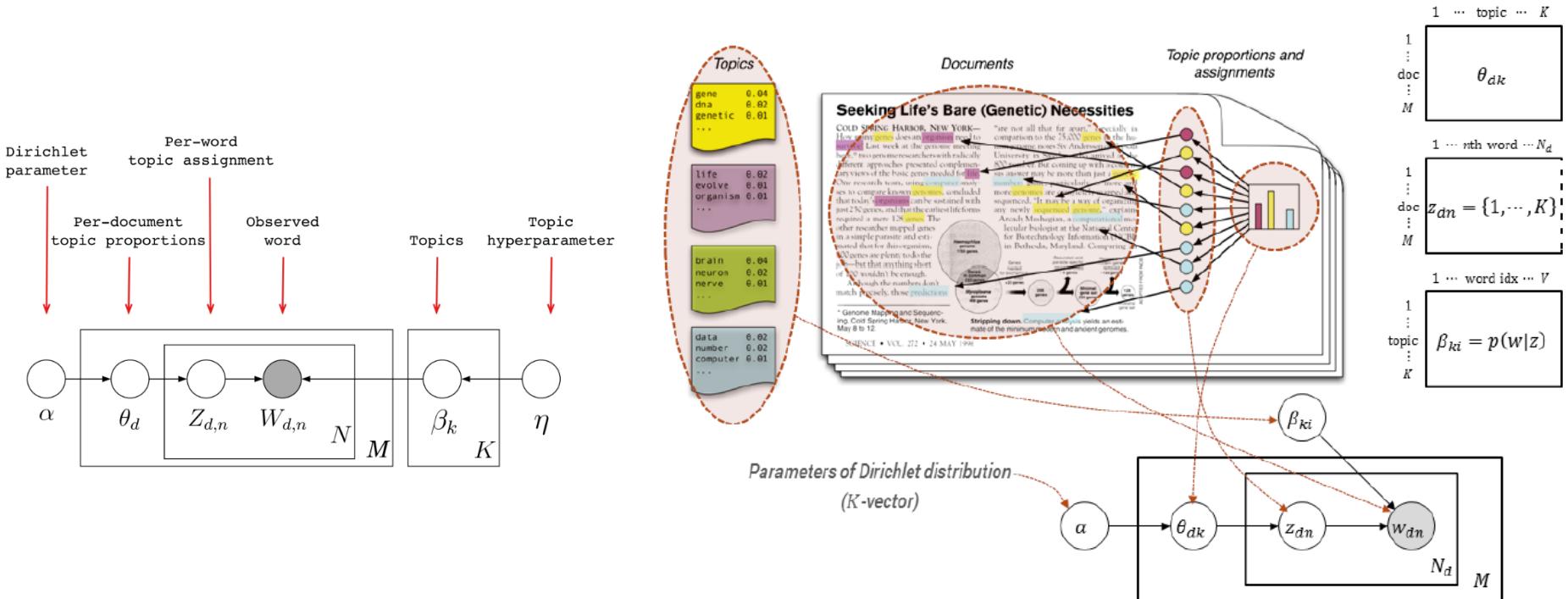
Topic Modeling

- ✓ 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



ANY
questions?

References

Research Papers

- 송서하, 김준홍, 김형석, 박재선, 강필성 (2019). 비정형 텍스트 데이터를 활용한 금융 기업 조기 경보 모형 개발: 부실은행 예측을 중심으로. 대한산업공학회지 45(5), 430-438.
- 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. (2018). Sentiment classification with word attention based on weakly supervised learning with a convolutional neural network. Knowledge-based Systems 152, 70-82.
- Lee, H. & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. Journal of Technology Transfer 43 (5), 1291-1317.
- Mo, K. Park, J., Jang, M., & Kang, P. (2018). 단어와 자소 기반 합성공 신경망을 이용한 문서 분류. 대한산업공학회지 44 (6), 442-451.
- 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.
- Seo, S., Seo, D., Jang, M., Jeong, J., & Kang, P. (2020). Unusual customer response identification and visualization based on text mining and anomaly detection, Expert Systems with Applications 144, 113111.
- 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.

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., Heckerman, 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>
- <https://www.topbots.com/top-ai-nlp-research-papers-2019/>