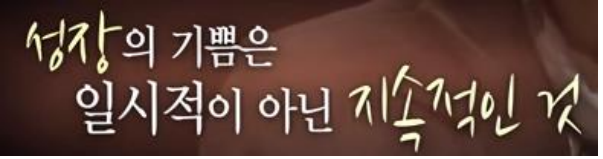
“AI LAWYER” 인공지능 변호사 – I

Artificial Intelligence Lawyer 1
(법령해석 문서 분류)

2017.10.18
민상식, Jason Min



얼마나 깊이 살았는가

무언가를 하고 그 결과에 대해 생각한다면/ 이는 자신만을 위해 그 일을 했다는 의미이다.
우리가 하는 행동 대부분의 결과는/ 눈에 드러나지 않는다./ 한정된 세상을 사는데/
행동의 결과에는 한계가 없기 때문이다.
행동의 결과를 모두 볼 수 있다면/ 그 행동 자체는 거의 의미를 갖지 못할 것이다.
지금 하는 일을 묵묵히 수행하라./ 그리고 이 순간이 앞으로 올 시간을 위해/ 기여할 것임을 믿어라.
중요한 것은 얼마나 오래 살았느냐가 아니라/ 얼마나 깊이 살았느냐이다.
우리는 세상과 밀접한 관련을 맺고 살아간다./ 하지만 그러면서도 세상은 한갓 환영이고/
어딘가에는 또 다른 차원 높은 세상이/ 있으리라는 생각이 든다.

- 톨스토이 -

인공지능 변호사 개발

[법의 이해]
- 법학 스터디

[인공지능 이해]
- 인공지능 스터디

[프로그래밍 이해]
- 하둡, 파이썬, 클라우드컴퓨팅 등

[환경 구성]
- 언어 : Python (Anaconda, Gensim, NLTK 등)

[정보 수집]
- 사례 기반 법적용 해석 - 법 데이터 4대 법 규정 수집 후 유사도 분석
- **비조치 의견(금융위),** 판례 등
- KISA 의견 등

[법령 분석 - 통계학적]

(정보분석)

1957 TF-IDF (한스 피터 룬, http://www.bloter.net/archives/264262, A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317)

(Topic Modeling)

2003 LDA(잠재 디리클레 할당)

(Word Embedding)

2013 Word2Vec (구글)

[법령 분석 – 머신러닝]

[인공지능변호사 구현]

1. 법 카테고리 자동분류(20171018)
- 법 해석 의견 조정된 내용 정리

2. 검색기
- 검색기

3. 법 해석기 (QnA) 변호사 만들기

Why?

영국 옥스퍼드대가 2013년 발표한 '고용의 미래(The Future of Employment)'라는 연구에 따르면, 인공지능 발달에 따른 컴퓨터 자동화에 의해 사라질 직업군의 확률은 법률사무직이 94%, 법원 속기사가 50%로 매우 높은 것으로 나타났다

THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?*

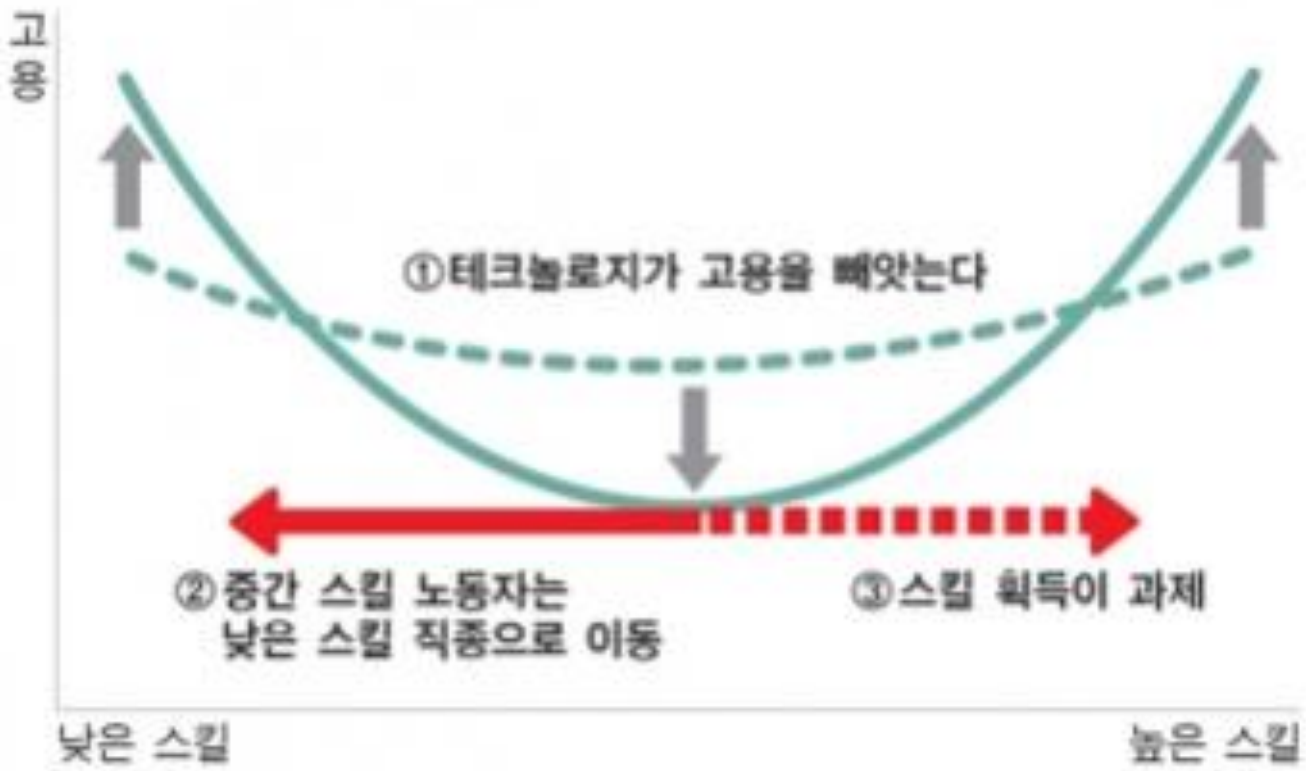
Carl Benedikt Frey† and Michael A. Osborne‡

September 17, 2013

Computerisable				
Rank	Probability	Label	SOC code	Occupation
115.	0.035	0	23-1011	Lawyers
276.	0.41	1	23-1012	Judicial Law Clerks
353.	0.64		23-1021	Administrative Law Judges, Adjudicators, and Hearing Officers
609.	0.94	1	23-2011	Paralegals and Legal Assistants

'인공지능의 법률 분야에서의 응용사례'를 발표하며 법률가들의 일반적인 사고 패턴이 ▲문제가 되는 법적 쟁점을 확정 ▲판례·문헌 검색 ▲주어진 사례를 기존 판례에 적용할지 판단 등의 순서를 거친다

"인공지능은 판례·문헌 검색 부분에서 장점을 가진다"면서도 "법적 쟁점을 확정하는 일은 현재 기술로는 어렵고 향후 인공지능의 발전 단계를 지켜봐야 한다"는 의견을 냈다. 특히 "판례 적용 여부는 인간의 고유한 통찰력이 필요한 지적 작업"이라며 "인공지능이 아무리 발전하더라도 인간을 대체할 수 없다"고 강조


고용
①테크놀로지가 고용을 빼앗는다
②중간 스킬 노동자는
낮은 스킬 직종으로 이동
③스킬 획득이 과제
낮은 스킬
높은 스킬

자료: 동양경제

COIN : COntract INtelligence

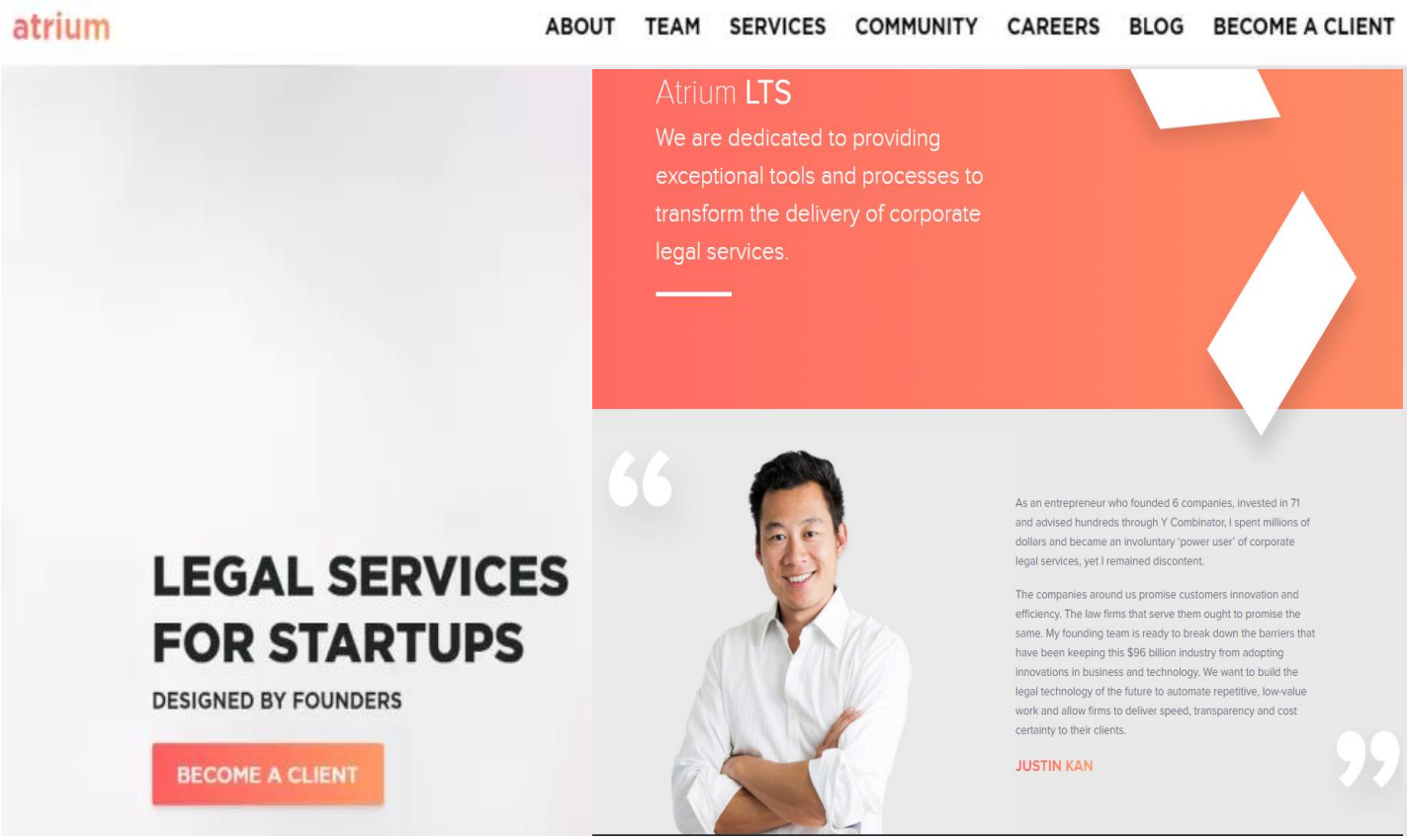
Bloomberg Markets Markets Tech Pursuits Politics Opinion Businessweek

JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours

By Hugh Son
February 27, 2012, 7:31 PM EST Updated on February 28, 2012, 7:24 AM EST

- New software does in seconds what took staff 360,000 hours
- Bank seeking to streamline systems, avoid redundancies

At JPMorgan Chase & Co., a learning machine is parsing financial deals that once kept



https://www.atriumlts.com/, http://www.venturesquare.net/753779

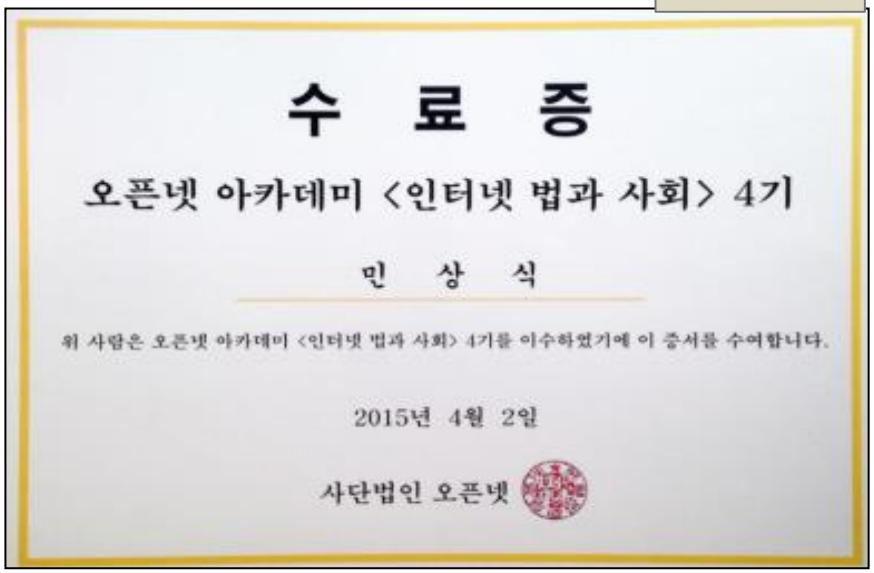
ARE YOU
PREPARED?

TIME TO PLAN

[법의 이해]
- 법학 스터디

2015.3.6

2015.4.2



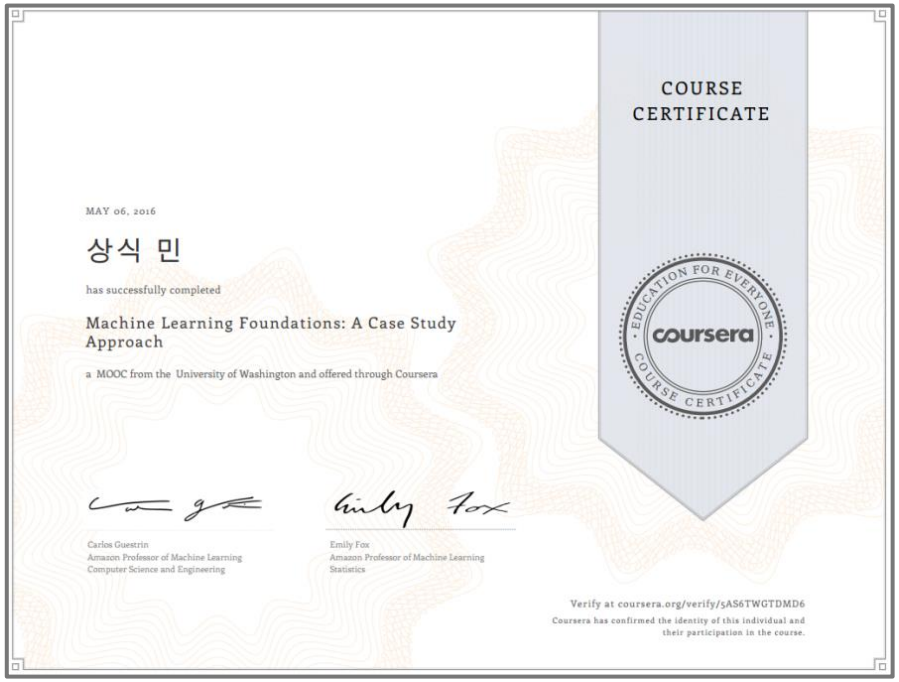
2015.3.1 ~ 2017.2.09

2017.8.22

[인공지능 이해]

- 인공지능 스터디

2016.3.06



COURSE CERTIFICATE

MAY 06, 2016

상식 민

has successfully completed

Machine Learning Foundations: A Case Study Approach

a MOOC from the University of Washington and offered through Coursera

EDUCATION FOR EVERYONE · coursera · COURSE CERTIFICATE

Carlos Guestrin
Amazon Professor of Machine Learning
Computer Science and Engineering

Emily Fox
Amazon Professor of Machine Learning
Statistics

Verify at coursera.org/verify/sAS6TWGTDMD6
Coursera has confirmed the identity of this individual and their participation in the course.

2017.4.11

인증번호 CERTIFICATE ID
33565e6a971e4c6192b6e92216a7a918

K-MOOC

CERTIFICATE

이 수 증

Min,Sangshik

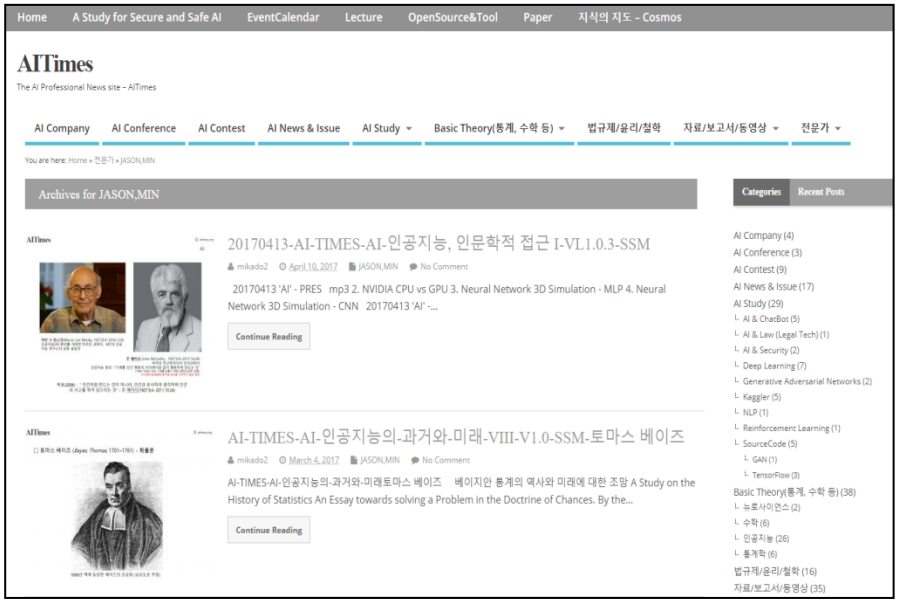
has successfully completed.

인공지능 및 기계학습 **Artificial Intelligence and Machine Learning**
(KAISTk.KCS470)

위 과정을 성실히 이수하여 이 증서를 수여합니다.
an online learning Initiative of KAISTk through K-MOOC.

이수일자 ISSUE DATE
2017년 4월 11일 April 11, 2017

오혜연 Alice Oh
교수 Professor
한국과학기술원 KAIST



Home | A Study for Secure and Safe AI | EventCalendar | Lecture | OpenSource&Tool | Paper | 지식의 지도 – Cosmos

AITimes

The AI Professional News site – AITimes

AI Company | AI Conference | AI Contest | AI News & Issue | AI Study | Basic Theory(통계, 수학 등) | 법규제/윤리/철학 | 자료/보고서/동영상 | 전문가

You are here: Home » 전문가 » JASON,MIN

Archives for JASON,MIN

20170413-AI-TIMES-AI-인공지능, 인문학적 접근 I-VL1.0.3-SSM

mikado2 | April 10, 2017 | JASON,MIN | No Comment

20170413 'AI' - PRES mp3 2. NVIDIA CPU vs GPU 3. Neural Network 3D Simulation - MLP 4. Neural Network 3D Simulation - CNN 20170413 'AI' -...

Continue Reading

AI-TIMES-AI-인공지능의-과거와-미래-VIII-V1.0-SSM-토마스 베이즈

mikado2 | March 4, 2017 | JASON,MIN | No Comment

AI-TIMES-AI-인공지능의-과거와-미래토마스 베이즈 베이지안 통계의 역사와 미래에 대한 조망 A Study on the History of Statistics An Essay towards solving a Problem in the Doctrine of Chances. By the...

Continue Reading

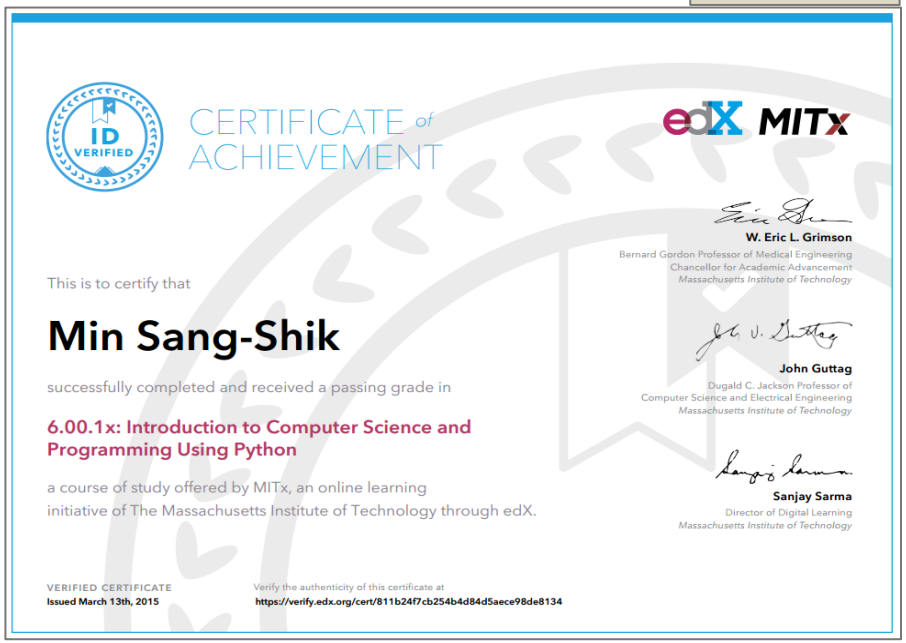
Categories | Recent Posts

AI Company (4)
AI Conference (3)
AI Contest (9)
AI News & Issue (17)
AI Study (29)
AI & ChatBot (5)
AI & Law (Legal Tech) (1)
AI & Security (2)
Deep Learning (7)
Generative Adversarial Networks (2)
Kaggler (5)
NLP (1)
Reinforcement Learning (1)
SourceCode (5)
GAN (1)
TensorFlow (3)
Basic Theory(통계, 수학 등) (38)
뉴로사이언스 (2)
수학 (6)
인공지능 (26)
통계학 (6)
법규제/윤리/철학 (16)
자료/보고서/동영상 (35)

[프로그래밍 이해]

- Computer Science, 하둡, 파이썬, 클라우드컴퓨팅 등

2015.3.13



ID VERIFIED

CERTIFICATE of ACHIEVEMENT

edX MITx

W. Eric L. Grimson
Bernard Gordon Professor of Medical Engineering
Chancellor for Academic Advancement
Massachusetts Institute of Technology

John Guttag
Dugald C. Jackson Professor of
Computer Science and Electrical Engineering
Massachusetts Institute of Technology

Sanjay Sarma
Director of Digital Learning
Massachusetts Institute of Technology

This is to certify that

Min Sang-Shik

successfully completed and received a passing grade in

6.00.1x: Introduction to Computer Science and Programming Using Python

a course of study offered by MITx, an online learning initiative of The Massachusetts Institute of Technology through edX.

VERIFIED CERTIFICATE
Issued March 13th, 2015

Verify the authenticity of this certificate at
https://verify.edx.org/cert/811b24f7cb254b4d84d5aece98de8134

2017.3.18

AWSome Day

amazon web services

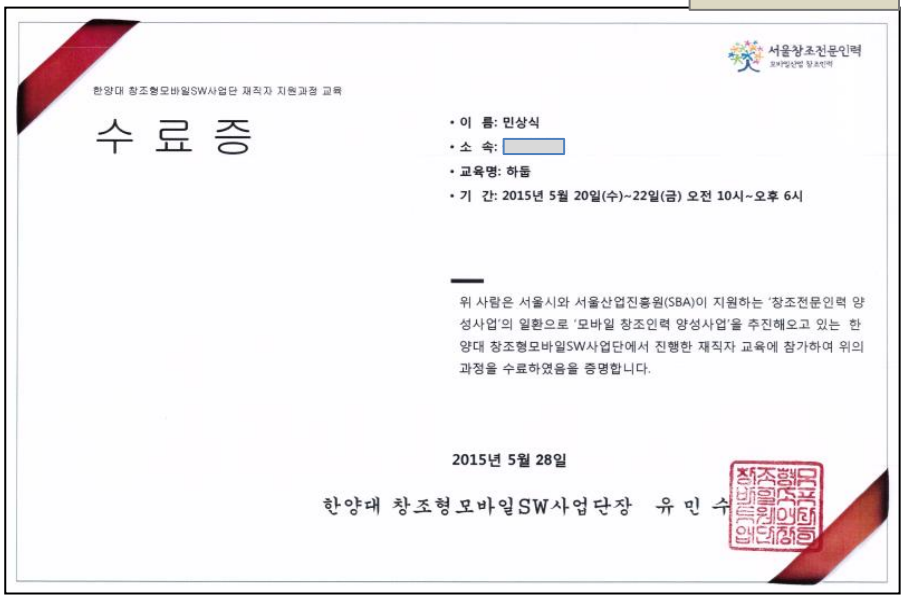
Certificate of Attendance

This is to certify that 민상식
has attended AWSome Day in Seoul on
18 March, 2017.

Jinuk Lee, Technical Trainer,
Amazon Web Services

Junpil Ahn, Technical Trainer,
Amazon Web Services

2015.5.22



서울창조전문인력

한양대 창조형모바일SW사업단 재직자 지원과정 교육

수 료 증

- 이 름: 민상식
- 소 속:
- 교육명: 하둡
- 기 간: 2015년 5월 20일(수)~22일(금) 오전 10시~오후 6시

위 사람은 서울시와 서울산업진흥원(SBA)이 지원하는 '창조전문인력 양성사업'의 일환으로 '모바일 창조인력 양성사업'을 추진해오고 있는 한양대 창조형모바일SW사업단에서 진행한 재직자 교육에 참가하여 위의 과정을 수료하였음을 증명합니다.

2015년 5월 28일

한양대 창조형모바일SW사업단장 유 민 수

READY TO GO
READY
READY TO GO

Competition on Legal Information Extraction and Entailment 검토
- 법률정보 추출 및 분석 대회 (JAPAN) - * 2017년, 4회

Overview of COLIEE 2017

8 pages • Published: June 3, 2017

Yoshinobu Kano, Mi-Young Kim, Randy Goebel and Ken Satoh

Abstract

We present the evaluation of the legal question answering Competition on Legal Information Extraction/Entailment (COLIEE) 2017. The COLIEE 2017 Task consists of two sub-Tasks: legal information retrieval (Task 1), and recognizing entailment between articles and queries (Task 2). Participation was open to any group based on any approach, and the tasks attracted 10 teams. We received 9 submissions to Task 1 (for a total of 17 runs), and 8 submissions to Task 2 (for a total of 20 runs).

Keyphrases: Information Retrieval, juris-informatics, Legal Document Processing, Question Answering, Textual Entailment

In: Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel and Tiago Oliveira (editors). COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment, vol 47, pages 1--8

http://www.nms.kcl.ac.uk/icail2017/cfcoliee.php
http://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2017/
https://easychair.org/publications/paper/350845

https://easychair.org/publications/volume/COLIEE_2017

Papers

Overview of COLIEE 2017
Yoshinobu Kano, Mi-Young Kim, Randy Goebel and Ken Satoh
Pages 1-8

A Civil Code Article Information Retrieval System based on Phrase Alignment with Article Structure Analysis and Ensemble Approach
Masaharu Yoshioka and Daiki Onodera
Pages 9-22

Multiple Agent Based Entailment System(MABES) for RTE
Byungtaek Jung, Chiseung Soh, Kihyun Hong, Seungtak Lim and Young-Yik Rhim
Pages 23-30

Recognizing entailments in legal texts using sentence encoding-based and decomposable attention models
Nguyen Truong Son, Phan Viet Anh and Nguyen Le Minh
Pages 31-42

Improving Legal Information Retrieval by Distributional Composition with Term Order Probabilities
Danilo S. Carvalho, Vu Tran, Khanh Van Tran and Nguyen Le Minh
Pages 43-56

Analyzable Legal Yes/No Question Answering System using Linguistic Structures
Yoshinobu Kano, Reina Hoshino and Ryosuke Taniguchi
Pages 57-67

Legal Information Retrieval Using Topic Clustering and Neural Networks
Rohan Nanda, Adebayo Kolawole John, Luigi Di Caro, Guido Boella and Livio Robaldo
Pages 68-78

Legal Question Answering System using Neural Attention
Ayaka Morimoto, Daiki Kubo, Motoki Sato, Hiroyuki Shindo and Yuji Matsumoto
Pages 79-89

EPiC Series in Computing
Volume 47, 2017, Pages 1–8
COLIEE 2017. Fourth Competition on Legal Information Extraction/Entailment

Overview of COLIEE 2017

Yoshinobu Kano[1], Mi-Young Kim[2], Randy Goebel[2], and Ken Satoh[3]
[1] Faculty of Informatics, Shizuoka University, Japan
[2] University of Alberta, Canada
[3] National Institute of Informatics, Japan
kano@inf.shizuoka.ac.jp, {miyoung2, rgoebel}@ualberta.ca, ksatoh@nii.ac.jp

Abstract

We present the evaluation of the legal question answering Competition on Legal Information Extraction/Entailment (COLIEE) 2017. The COLIEE 2017 Task consists of two sub-Tasks: legal information retrieval (Task 1), and recognizing entailment between articles and queries (Task 2). Participation was open to any group based on any approach, and the tasks attracted 10 teams. We received 9 submissions to Task 1 (for a total of 17 runs), and 8 submissions to Task 2 (for a total of 20 runs).

Legal Question Answering Data Corpus The corpus of legal questions is drawn from **Japanese Legal Bar exams**, and the **relevant Japanese Civil Law articles** have been also provided.

Question	*A person who made a manifestation of intention which was induced by duress emanated from a third party may rescind such manifestation of intention on the basis of duress, only if the other party knew or was negligent of such fact.*
Related Article	*(Fraud or Duress) Article 96 (1)Manifestation of intention which is induced by any fraud or duress may be rescinded.(2)In cases any third party commits any fraud inducing any person to make a manifestation of intention to the other party, such manifestation of intention may be rescinded only if the other party knew such fact.(3)The rescission of the manifestation of intention induced by the fraud pursuant to the provision of the preceding two paragraphs may not be asserted against a third party without knowledge.*

Question	A special provision that releases warranty can be made, but in that situation, when there are rights that the seller establishes on his/her own for a third party, the seller is not released of warranty.
Related Article	(Special Agreement Disclaiming Warranty)Article 572 Even if the seller makes a special agreement to the effect that the seller will not provide the warranties set forth from Article 560 through to the preceding Article, the seller may not be released from that responsibility with respect to any fact that the seller knew but did not disclose, and with respect to any right that the seller himself/herself created for or assigned to a third party.
Label	Yes

2.1 Task 1

Our goal is to explore and evaluate legal document retrieval technologies that are both effective and reliable. The Task investigates the performance of systems that search a static set of civil law articles using previously unseen queries, and return relevant articles. We say an article is "Relevant" to a query if and only if the query sentence can be entailed from the meaning of the article. If combining the meanings of more than one article (e.g., "A," "B," and "C") can answer a query sentence, then all the articles ("A," "B,", and "C") are considered "Relevant." If a query can be answered by an article "D," and it can be also answered by another article "E" independently, we also say that both "D" and "E" are "Relevant." This Task requires the retrieval of *all* the articles that are relevant to answering a query.

Japanese civil law articles (and English translation) have been provided, and training data consists of query and relevant article pairs. The process of executing the queries over the articles and generating the experimental runs should be entirely automatic. Test data includes only queries but no relevant articles.

ID	Approaches
HUKB	article structure analysis, phrase matching, rank SVM with 15 similarity scores and alignment scores, ensemble
iLis7-1	TF-IDF, LSM, LDA, Word2Vec, LSA
JAISTNLP	TF-IDF
JNLP	ranking related n-gram collections, term order probabilities, relevance disambiguation.
NOR	LDA
UA	TF-IDF, language model

Latent Dirichlet allocation (LDA)

2.2 Task 2

The goal of Phase 2 is to construct Yes/No question answering systems for legal queries, by entailment from the relevant articles. The Task investigates the performance of systems that answer "Y" or "N" to previously unseen queries by somehow comparing the intersection of meanings between queries and relevant articles. Training data consists of triples: a query, relevant articles and a correct answer "Y" or "N." For the competition evaluation, the process of finding relevant articles, executing the queries over the relevant articles and generating the experimental runs should be entirely automatic. Test data includes only queries and corresponding "Y/N" labels for each query.

ID	Approaches
iLis9	TF-IDF, negation detection
JAISTNLP	ranking, encoding-based and attention neural network
KIS	case-role based linguistic analysis, predicate-argument structures
NAIST	Word2vec, attention neural network
NOR	CNN, LSTM
UA	Korean dependency parser, Excite Japanese/Korean machine translation, semantic dictionary, k-means clustering

Run	Prec.	Recall	F	L.	Run	Prec.	Recall	F	L.
HUKB-1	0.658	0.472	0.550	J	JNLP1-T	0.500	0.354	0.414	E
HUKB-2	0.586	0.490	0.534	J	KID17	0.703	0.518	0.596	E
HUKB-3	0.551	0.536	0.543	J	KIS-IE-M	0.263	0.272	0.267	J
iLis7-1	0.734	0.554	0.632	E	KIS-IE-NM	0.346	0.245	0.287	J
iLis7-2	0.654	0.500	0.567	E	NOR17	0.462	0.500	0.480	E
JAISTNLP2-1a-norerank	0.628	0.445	0.521	E	UA-LM	0.602	0.427	0.500	E
JAISTNLP2-1b-rerank	0.615	0.436	0.510	E	UA-TFIDF	0.666	0.472	0.553	E
JNLP1-R	0.686	0.536	0.602	E	VNPT	0.430	0.281	0.340	E
JNLP1-RT	0.689	0.545	0.609	E					

IR results (Task 1) on the formal run data.

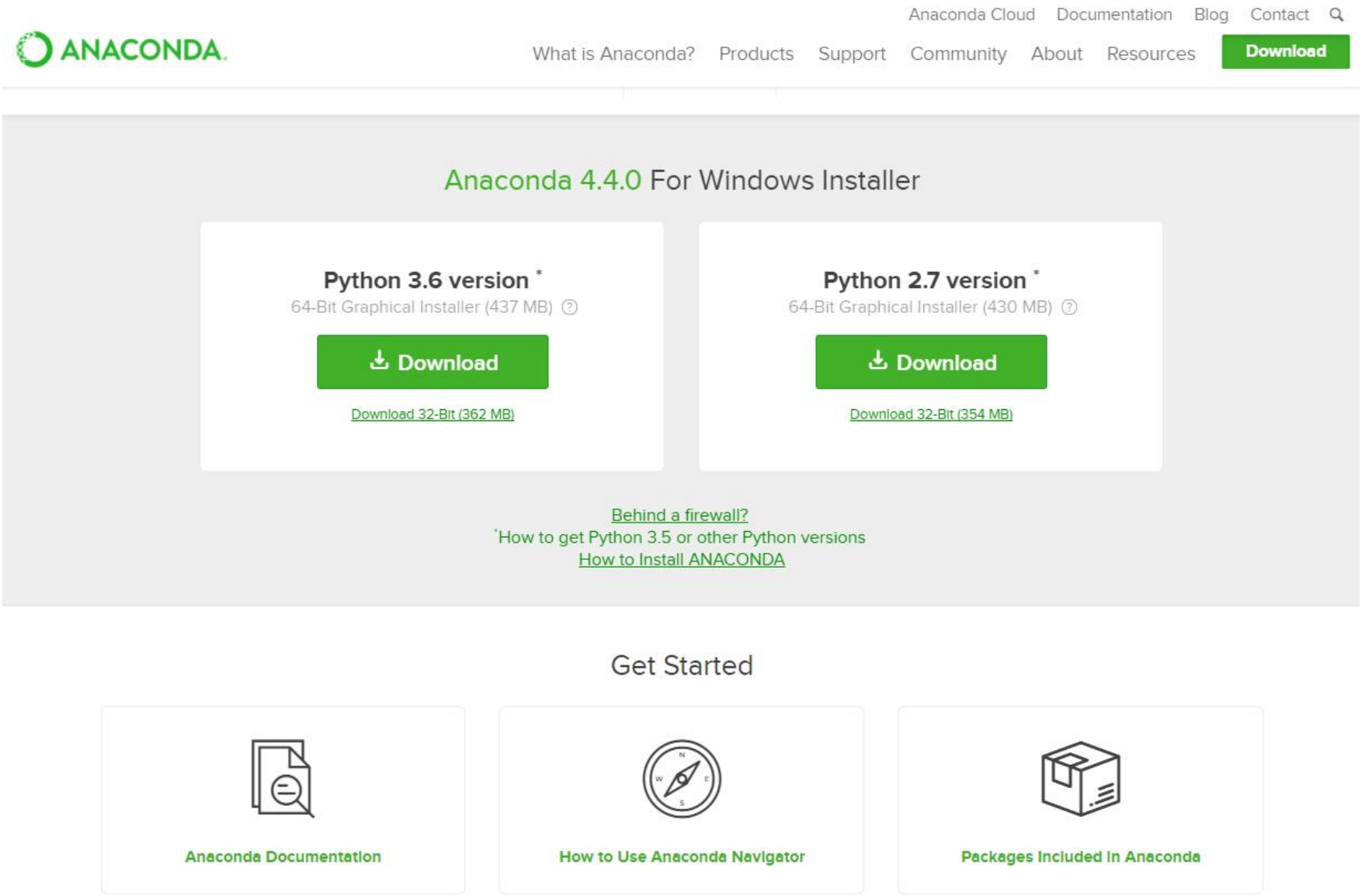
Run	Accuracy	Language	Run	Accuracy	Language
iLis7	0.564	English	KIS-YN-CM	0.538	Japanese
iLis9-1	0.576	English	KIS-YN-CS	0.589	Japanese
iLis9-2	0.538	Japanese	KIS-YN-M	0.576	Japanese
JAISTNLP2-2a-1a-norerank	0.512	English	KIS-YN-S	0.653	Japanese
JAISTNLP2-2a-1b-rerank	0.474	English	NAIST1	0.615	Japanese
JAISTNLP2-2b-1a-norerank	0.487	English	NAIST2	0.653	Japanese
JAISTNLP2-2b-1b-rerank	0.500	English	NAIST3	0.474	Japanese
JNLP1-R	0.435	English	NOR17	0.538	English
JNLP1-RT	0.487	English	UA-LM	0.717	Japanese
KIS-YN-A	0.538	Japanese	UA-TFIDF	0.692	Japanese

Entailment results (Task 2) on the formal run data

LET'S GO

[환경 구성]

- 언어 : Python (Anaconda, Gensim, NLTK 등)




```
conda search python
conda create -n py35 python=3.5.3 anaconda
activate py35
deactivate py35
```

http://thrillfighter.tistory.com/466

https://www.lucypark.kr/slides/2015-pyconkr/#2

PyCon Korea 2015

한국어와 NLTK, Gensim의 만남

(부제: 영화 리뷰를 컴퓨터가 이해할 수 있는 형식으로 표현해서 센티멘트 분석하기)

2015-08-29

Lucy Park (박은정)

http://radimrehurek.com/gensim/install.html

gensim

topic modelling for humans

Download
latest version from the Python Package Index

Direct install with:
easy_install -U gensim

Home | Tutorials | Install | Support | API | About

```
>>> from gensim import corpora, models, similarities
>>>
>>> # Load corpus iterator from a Matrix Market file on disk.
>>> corpus = corpora.MmCorpus('/path/to/corpus.mm')
>>>
>>> # Initialize Latent Semantic Indexing with 200 dimensions.
>>> lsi = models.LsiModel(corpus, num_topics=200)
>>>
>>> # Convert another corpus to the latent space and index it.
>>> index = similarities.MatrixSimilarity(lsi[another_corpus])
>>>
>>> # Compute similarity of a query vs. indexed documents
>>> sims = index[query]
```

Gensim is a FREE Python library

- Scalable statistical semantics
- Analyze plain-text documents for semantic structure
- Retrieve semantically similar documents

Features

Hover your mouse over each feature for more info.

- Scalability
- Platform independent
- Robust
- Open source
- Efficient implementations
- Converters & I/O formats
- Similarity queries
- Support

주요단어, 아이디어 정의

코퍼스(Corpus or Corpora 복수형, 말뭉치) : 특정한 목적을 가지고 언어의 표본을 추출한 집합
문서의 집합체

연어(collocation) : 다른 어떤 단어와 함께 나타나려는 경향
특정한 뜻을 나타낼 때 흔히 함께 쓰이는 단어

Bag of words : bag {a, a, b, c, c, c} = bag {c, a, c, b, a, c}, 순서는 무관

You shall know a word by the company it keeps (Firth, J. R. 1957:11)

존 루퍼트 퍼스 John Rupert Firth

(June 17, 1890 in Keighley, Yorkshire – December 14, 1960 in Lindfield, West Sussex)

단일정보분석(법률개정안 1건)

법률 개정안 - konlpy.corpus import kobill, 1809890.txt

지방공무원법 일부개정법률안
(정의화의원 대표발의)
 의 안
 번 호
9890
발의연월일 : 2010. 11. 12.
발 의 자 : 정의화. 이명수. 김을동
이사철. 여상규. 안규백
황영철. 박영아. 김정훈
김학송 의원(10인)
제안이유 및 주요내용
 초등학교 저학년의 경우에도 부모의 따뜻한 사랑과 보살핌이 필요한 나이이나, 현재 공무원이 자녀를 양육하기 위하여 육아휴직을 할 수 있는 자녀의 나이는 만 6세 이하로 되어 있어 초등학교 저학년인 자녀를 돌보기 위해서는 해당 부모님은 일자리를 그만 두어야 하고 이는 곧 출산의욕을 저하시키는 문제로 이어질 수 있을 것임. 따라서 육아휴직이 가능한 자녀의 연령을 만 8세 이하로 개정하려는 것임(안 제63조제2항제4호).
...
...
신 ·구조문대비표

현 행	개 정 안
제63조(휴직) ① (생 략)	제63조(휴직) ① (현행과 같음)
② 공무원이 다음 각 호의 어느 하나에 해당하는 사유로 휴직을 원하면 임용권자는 휴직을 명할 수 있다. 다만, 제4호의 경우에는 대통령령으로 정하는 특별한 사정이 없으면 휴직을 명하여야 한다.	② ----------.----------
1. ~ 3. (생 략)	1. ~ 3. (현행과 같음)

https://m.blog.naver.com/2feelus/220384206922

단일정보분석(법률개정안 1건)

법률 개정안 - from konlpy.tag import Twitter ,
morphs 함수(형태소분석 – 형태소별로 단순분리)

http://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/
http://konlpy-ko.readthedocs.io/ko/v0.4.3/api/konlpy.tag/

tokens_ko

['지방공무원법', '일부', '개정', '법률', '안', '(', '정의화', '의원', '대표', '발의', ')', '의', '안', '번', '호', '9890', '발의', '연월일', ':', '2010', '.', '11', '.', '12', '.', '발', '의', '자', ':', '정의화', '. ', '이명수', '. ', '김을동', '이사철', '. ', '여상규', '. ', '안규백', '황영철',
...
...
...
'분석관', '김', '태', '완', '(', '02', '-', '788', '-', '4649', ',', 'tanzania@assembly.go.kr', ')', '-', '11', '-']

https://m.blog.naver.com/2feelus/220384206922

단일정보분석(법률개정안 1건)

법률 개정안 - import nltk
ko = nltk.Text ()

http://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/
http://konlpy-ko.readthedocs.io/ko/v0.4.3/api/konlpy.tag/

tokens_ko

['지방공무원법', '일부', '개정', '법률', '안', '(', '정의화', '의원', '대표', '발의', ')', '의', '안', '번', '호', '9890', '발의', '연월일', ':', '2010', '.', '11', '.', '12', '.', '발', '의', '자', ':', '정의화', '. ', '이명수', '. ', '김을동', '이사철', '. ', '여상규', '. ', '안규백', '황영철',
...
...
...
'분석관', '김', '태', '완', '(', '02', '-', '788', '-', '4649', ',', 'tanzania@assembly.go.kr', ')', '-', '11', '-']

```
#4. 토큰 정보및 단일 토큰 정보 알아내기
print(len(ko.tokens))       # returns number of tokens (document length)
print(len(set(ko.tokens)))  # returns number of unique tokens
```

전체 토큰 수 : 1707
고유 토큰 수 : 476

https://m.blog.naver.com/2feelus/220384206922

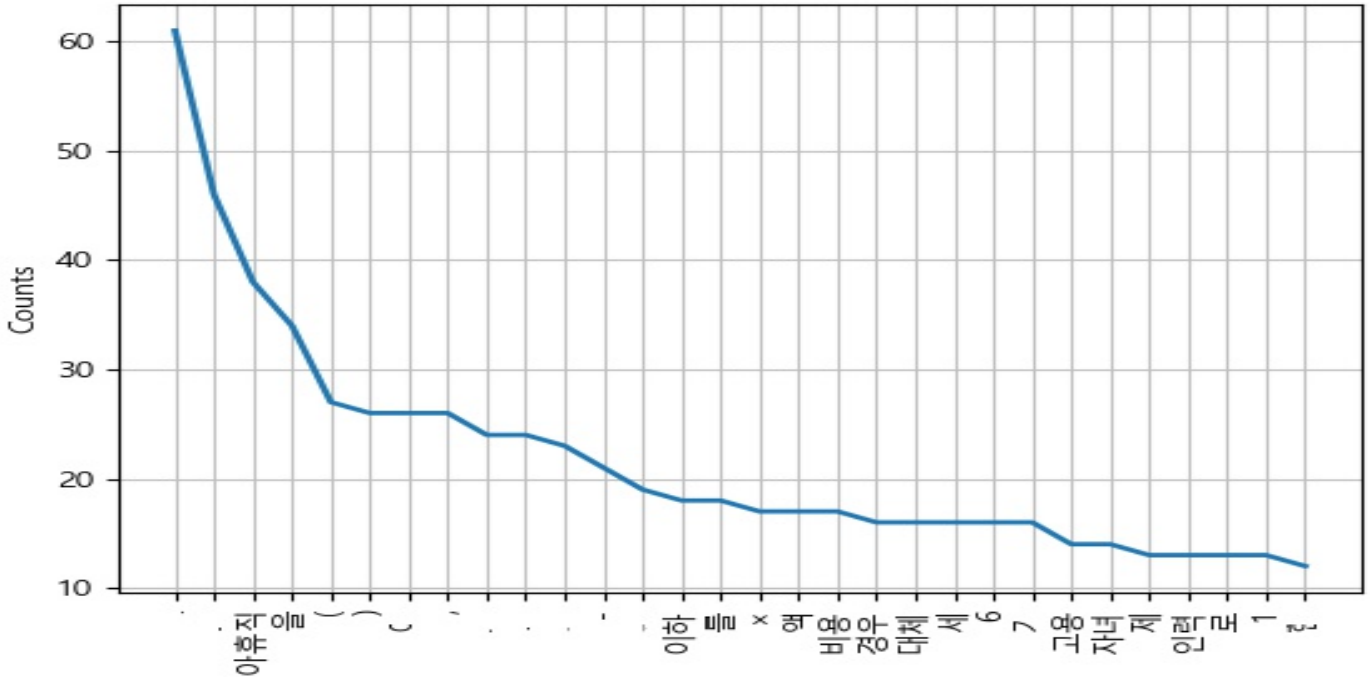
단일정보분석(법률개정안 1건)

법률 개정안 - import nltk
ko = nltk.Text ()

http://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/
http://konlpy-ko.readthedocs.io/ko/v0.4.3/api/konlpy.tag/

```
특정 단어에 대해서 빈도수 확인하기
print("육아휴직 빈도수 : " + str(ko.count(str('육아휴직'))))
print("법률 빈도수 : " + str(ko.count(str('법률'))))
```

육아휴직 빈도수 : 38
법률 빈도수 : 7



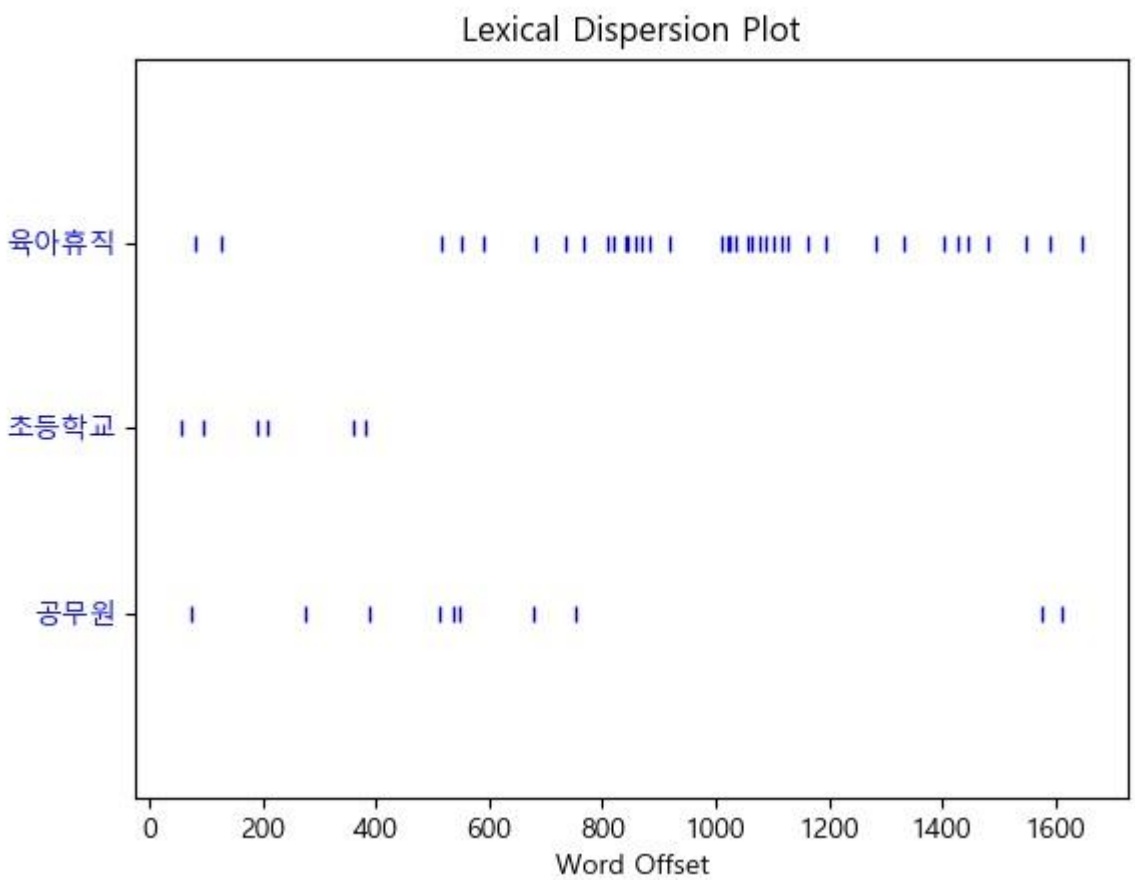
https://m.blog.naver.com/2feelus/220384206922

단일정보분석(법률개정안 1건)

법률 개정안 - import nltk
ko = nltk.Text ()

http://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/
http://konlpy-ko.readthedocs.io/ko/v0.4.3/api/konlpy.tag/

```
print ("분산차트보기")
pylab.show = lambda: pylab.savefig('1809890_dispersion_plot.jpg')
ko.dispersion_plot(['육아휴직', '초등학교', '공무원'])
```



https://m.blog.naver.com/zteelus/220384206922

단일정보분석(법률개정안 1건)

법률 개정안 - import nltk
ko = nltk.Text ()

http://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/
http://konlpy-ko.readthedocs.io/ko/v0.4.3/api/konlpy.tag/

```
print ("색인하기 (본문속에서 해당 검색어가 포함된 문장을 찾아주는것)")
#ko.concordance(unicode('법'))
ko.concordance('법')
```

Displaying 4 of 4 matches:
학년 이하 를 말한 다) 의 자녀 를 " 로 한 다 . 부 칙 이 **법** 은 공포 한 날 부터 시행 한 다 . - 3 - 신 · 구조 문대비
미 첨부 사유 서 1 . 재정 수반 요인 개정안 에서 「 국가공무원 **법** 」 제 71 조제 2 항제 4 호 중 국가공무원 의 육아 휴직 가능
만 6 세 이하 에서 만 8 세 이하 로 하고 , 「 지방 공 무 원 **법** 」 제 63 조제 2 항제 4 호 중 지방 공무원 의 육아휴직 가능
을 만 6 세 이하 에서 만 8 세 이하 로 하고 , 「 교육 공무원 **법** 」 제 44 조제 1 항 제 7 조 중 교육 공무원 의 육아휴직 가능

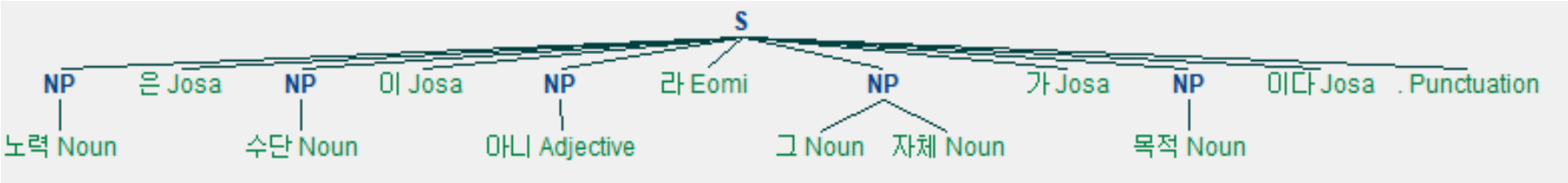
https://m.blog.naver.com/2feelus/220384206922

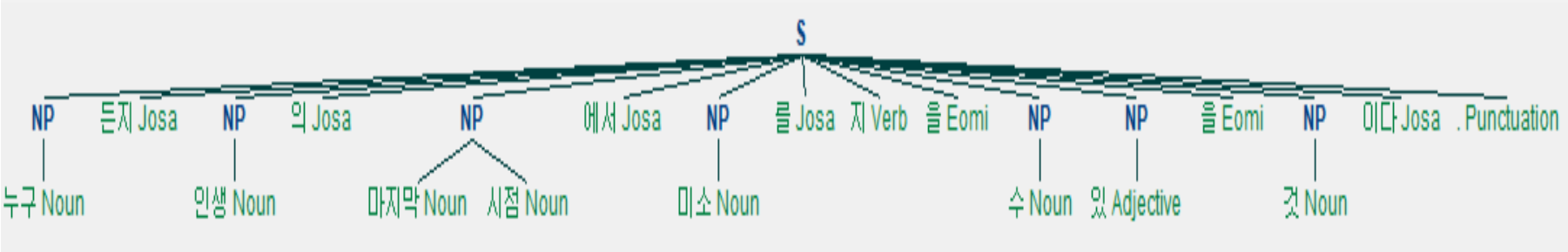
https://m.blog.naver.com/2feelus/220384206922

형태소 분석

```
의미단위로나누기(Tagging and chunking) 10.1 형태소 분석기(POS tagging)
from konlpy.tag import Twitter;
t = Twitter()
tags_ko = t.pos('작고 노란 강아지가 고양이에게 짖었다')
print(tags_ko)
```

```
노력은 수단이 아니라 그 자체가 목적이다.
노력하는 것 자체에 보람을 느낀다면,
누구든지 인생의 마지막 시점에서 미소를 지을 수 있을 것이다.
```





단일정보분석(법률개정안 1건)

법률 개정안 - from gensim import corpora
dictionary_ko = corpora.Dictionary(texts_ko)
dictionary_ko.save('ko.dict') # save dictionary to file for future use

```
from gensim import corpora
print('₩n₩nencode tokens to integers')
dictionary_ko = corpora.Dictionary(texts_ko)
dictionary_ko.save('ko.dict')  # save dictionary to file for future use
```

```
...
...
결혼/NounrM►X
신청/NounrK{X
소재/NounrMX지원하기/VerbrM|X신고한/VerbrM*X
가입/NounrMJX
사례/NounrMKX예상되는/VerbrMX곤란하/AdjectiverM►X계약서/NounrMRX나성린/NounrM►X
서류/NounrM►X
긴급/NounrM►X
경과/NounrM►X응/NounrMZX바/NounrM'X방산수/NounrM)X박은수/NounrM►X
172/Numberr M►X
징계/Nounr!M►X는/Eomir"K►X정영희/Nounr#M►X
다만/Nounr$K►X등록한/Verbr%MbX이/Determinerr&K►X정/Nounr'M►X사/Nounr(KjX
주권/Nounr)M►X안홍준/Nounr*MfX
공정/Nounr+M►X심/Nounr,M►X項第/Foreignr-M)X
격상/Nounr.M6X
소송/Nounr/M►X참여하게/Verbr0M►X
...
...
...
```

https://m.blog.naver.com/2feelus/220384206922
https://datascienceschool.net/view-notebook/6927b0906f884a67b0da9310d3a581ee/
http://dalpo0814.tistory.com/13

SO WHAT?

자연언어처리(Natural Language Processing)

분포가설(Distributional Hypothesis), 벡터공간모델(Vector Space Models)

* 언어학적 측면에서 Harris (1954), Firth (1957)의 분포 가설(Distributional Hypothesis)을 활용한 의미 유사도를 측정

아이디어 : '빈도'를 '의미'로 변환할 수 있다

문서 집합(말뭉치)에 속하는 각각의 **문서***들을 벡터공간의 **벡터로 표현**할 수 있다.

* 단어 또는 형태소로 분리 한 후 분석

벡터공간에 벡터로 표현된 **문서**들 사이의 **거리가 가깝다면 의미가 유사**하다.

distributional hypothesis : **비슷한 맥락에 등장하는 단어**들은 **유사한 의미**를 지니는 경향이 있다.
statistical semantics hypothesis : **언어 사용의 통계적 패턴**은 사람들이 **의미하는 바를 이해**하는 데 쓰일 수 있다.
bag of words hypothesis : 어떤 문서에 출현한 단어들의 빈도는 문서와 쿼리의 관련성을 나타내는 경향이 있다.
어떤 문서가 쿼리 문서와 유사한 벡터라면 그 의미도 비슷하다.
Latent relation hypothesis : 비슷한 패턴으로 **동시에 등장하는 단어쌍**은 유사한 **의미적 관계**를 지니는 경향이 있다.

방법론 : 단어-문서행렬(Term-Document Matrix),
단어-문맥행렬(Word-Context Matrix),
페어-패턴행렬(Pair-Pattern Matrix),
Word2Vec, Glove, Fasttext 등

Word Embedding : 단어를 벡터로 변환

'one-hot encoding'방식 : 단어 하나에 인덱스 정수를 할당(Bag of Words)

S1. "I am a boy"
S2. "I am a girl"

["I": 0, "am": 1, "a": 2, "boy": 3, "girl": 4]

S1 OHE : [11110]
S2 OHE : [11101]

SVD(특이값분해, Singular Value Decomposition), PCA(주성분분석, Principal Component Analysis)
-> LSA(잠재의미분석, Latent Sematic Analysis)-> NNLM, Word2Vec, Glove, Fasttext 등

- Word Embedding
문장속의 단어간의 관계를 **Unsupervised Learning 방식으로 분석**하여 만들어지는 수십~수백차원의 벡터로서 특징(Feature)화 되는 단어들을 만들어내는것. 단어가 가지고있는 벡터간의 연산을 통해 다른단어와의 관계를 만들어내게 된다

embedding EN

뜻 포함 위 매장 (수학)

- 미분기하학에서, 매장(埋藏, 영어: embedding) 또는 묻기는 그 상이 정의역과 위상동형인 단사 몰입이다.

Definition of embedding in English Dictionary

- 명사 (Noun) PL **embeddings** SUF **-ing**
  1. (mathematics) A map which maps a subspace (smaller structure) to the whole space (larger structure).
- 동사 (Verb)
  1. present participle of embed.

Deerwester et ak.(1990)과 Landauer and Dumais(1997)은 이 기법을 적용하면 단어와 문맥 간의 내재적인 의미(latent/hidden meaning)을 효과적으로 보존할 수 있게 돼 결과적으로 문서 간 유사도 측정 등 모델의 성능 향상에 도움
Rapp(2003)은 입력 데이터의 노이즈 제거, Vozalis and Margaritis(2003)은 입력데이터의 **sparsity**를 줄이는 효과

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/04/06/pcasvdlsa/

Word Embedding 방법론

2003 **Neural Network Language Model(NNLM,** Bengio**)** → 2013 **Word2Vec** (Google Mikolov) → 2014 **GloVe (스탠포드 대학)** → 2016 **Fasttext (페이스북)**

CBOW **Skip-Gram**

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

https://research.fb.com/projects/fasttext/

A. Neural Network Language Model(NNLM), Bengio(2003)

P(간다P(간다|발,없는,말이,천리)발,없는,말이,천리)

$$P(w_t|w_{t-1},\ldots,w_{t-n+1}) = \frac{exp(y_{w_t})}{\sum_i exp(y_i)}$$

조건부확률을 최대화하는 방향으로 학습

직전까지 등장한 n−1개 단어들로 n번째 단어를 맞추는 **N-gram** 모델이 그 본질

ex : '발', '없는', '말이', '천리' 네 개 단어로 '간다'를 맞추자는 것

단어 임베딩은 신경망을 이용하여 언어 모형을 만들려는 시도에서 나옴

"A Neural Probabilistic Language Model", Bengio, et al. 2003
http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

https://datascienceschool.net/view-notebook/6927b0906f884a67b0da9310d3a581ee/

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/03/30/word2vec/

Word Embedding

B. Word2Vec 란?
Distributional Hypothesis 에 근거한 방법론

"Efficient Estimation of Word Representations in Vector Space", Mikolov, et al. 2013
https://arxiv.org/pdf/1301.3781v3.pdf
"word2vec Parameter Learning Explained", Xin Rong,
http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf

2003

Neural Network Language Model(NNLM, Bengio**)**

2013

Word2Vec
(Google Mikolov**)**

CBOW | **Skip-Gram**

Neural Network Language Model(NNLM)을 계승하면서도 학습 속도와 성능을 비약적으로 끌어올림

CBOW(Continuous Bag of Words)와 **Skip-Gram** 두 가지 방식이 있음
- 전자는 주변에 있는 단어들을 가지고 중심에 있는 단어를 맞추는 방식 예시 : 나는 ___ 를 먹는다.
- 후자는 중심에 있는 단어로 주변 단어를 예측하는 방법 예시 : ____ 과자__ _____

말뭉치(코퍼스, Corpus or Corpora) : 나는 밥을 먹는다. 나는 과자를 먹었다. 나는 라면을 먹는다.

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/03/30/word2vec/

Word Embedding

B. Word2Vec 란?

비슷한 위치에 등장하는 단어들은 그 의미도 유사할 것이라는 전제
Word2Vec(Skip-Gram)은 다음 식을 최대화하는 걸 목표로 함

$$p(o|c) = \frac{exp(u_o^T v_c)}{\sum_{w=1}^{W} exp(u_w^T v_c)}$$

O : 주변단어(surrounding word)
C : 중심단어(context word)

p(o|c) : 중심단어(c)가 주어졌을 때 주변단어(o)가 등장할 **조건부확률**
이 식을 최대화하는 것은 중심단어로 주변단어를 잘 맞춘다는 의미

'외나무다리'가 등장했을 때 '원수'라는 표현이 나올 것이라는 사실을 예측

u와 v는 단어벡터들
예시 : '외나무다리'라는 중심단어 벡터가 v_c, '원수'라는 주변단어 벡터가 u_o

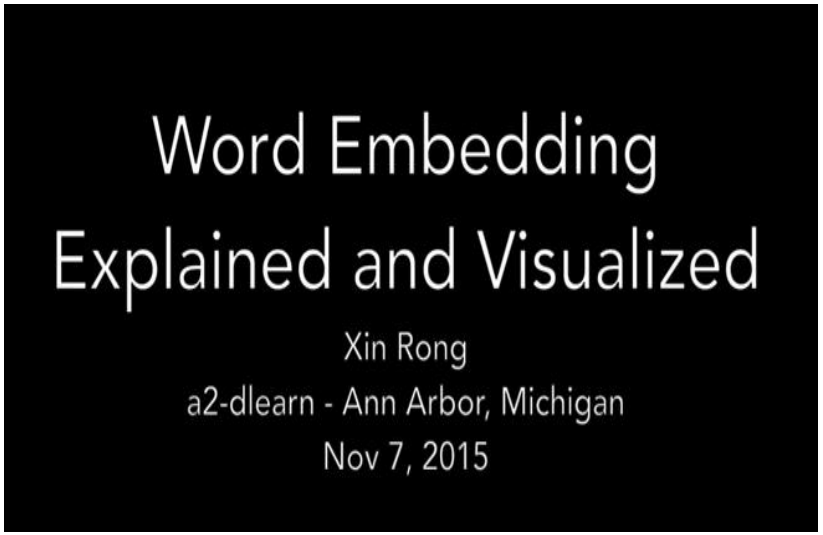
[학습 과정]
사용자가 주변단어 몇 개를 볼 지(**window**)를 정해주면, Word2Vec은 말뭉치를 window 크기로 슬라이딩, 중심단어별로 주변단어들을 보고 각 단어에 해당하는 벡터들의 요소값들을 조금씩 업데이트함으로써 단어를 벡터로 임베딩
* Word2Vec은 window 내에 등장하지 않는 단어에 해당하는 벡터는 중심단어 벡터와 벡터공간상에서 멀어지게끔 (**내적값 줄이기**), 등장하는 주변단어 벡터는 중심단어 벡터와 가까워지게끔(**내적값 키우기**) 한다는 것

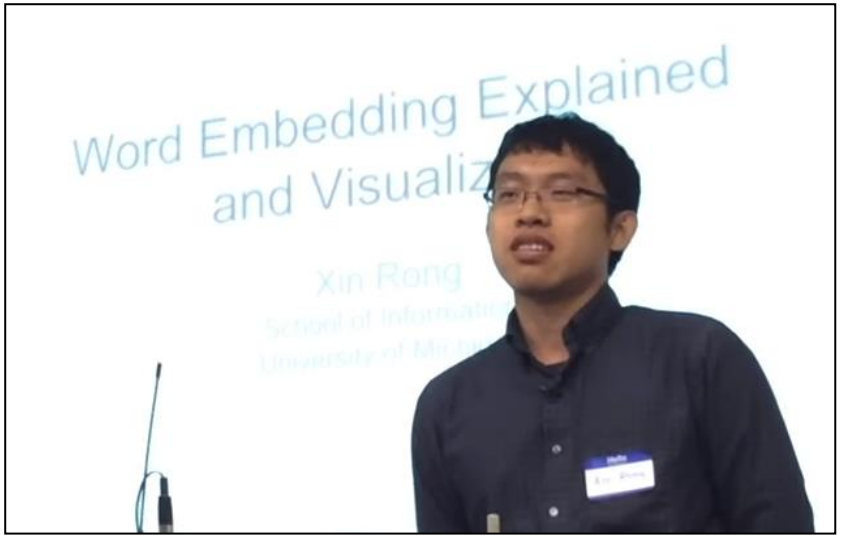
주변 단어를 몇 개 볼지를 정하고 동시에 등장하는 단어의 빈도수를 세어서 행렬로 변환한 '단어-문맥행렬'과 매우 유사함, Word2Vec은 기존 count 기반 방법론처럼 자주 같이 등장하는 단어들의 정보(Co-occurrence)를 보존
* Word2Vec은 본질적으로 기존 count 기반의 방법론과 다르지 않다는 점을 논증
Neural Word Embedding as Implicit Matrix Factorization, Omer and Yoav(2014)

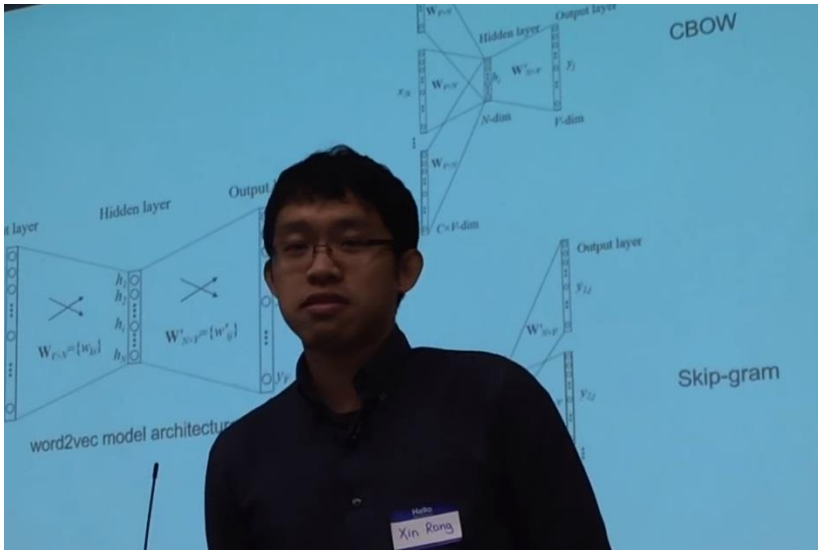
ng/

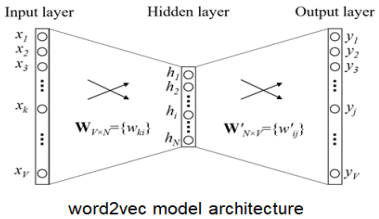
Word Embedding

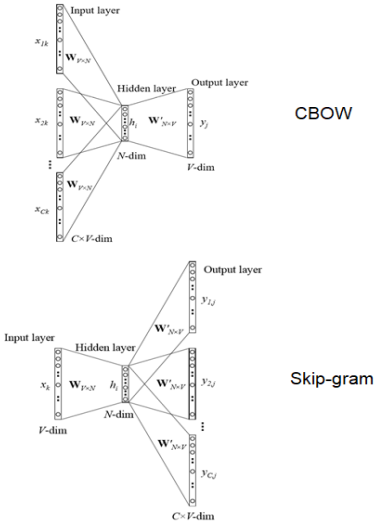
B. Word2Vec - word2vec Parameter Learning Explained, Xin Rong







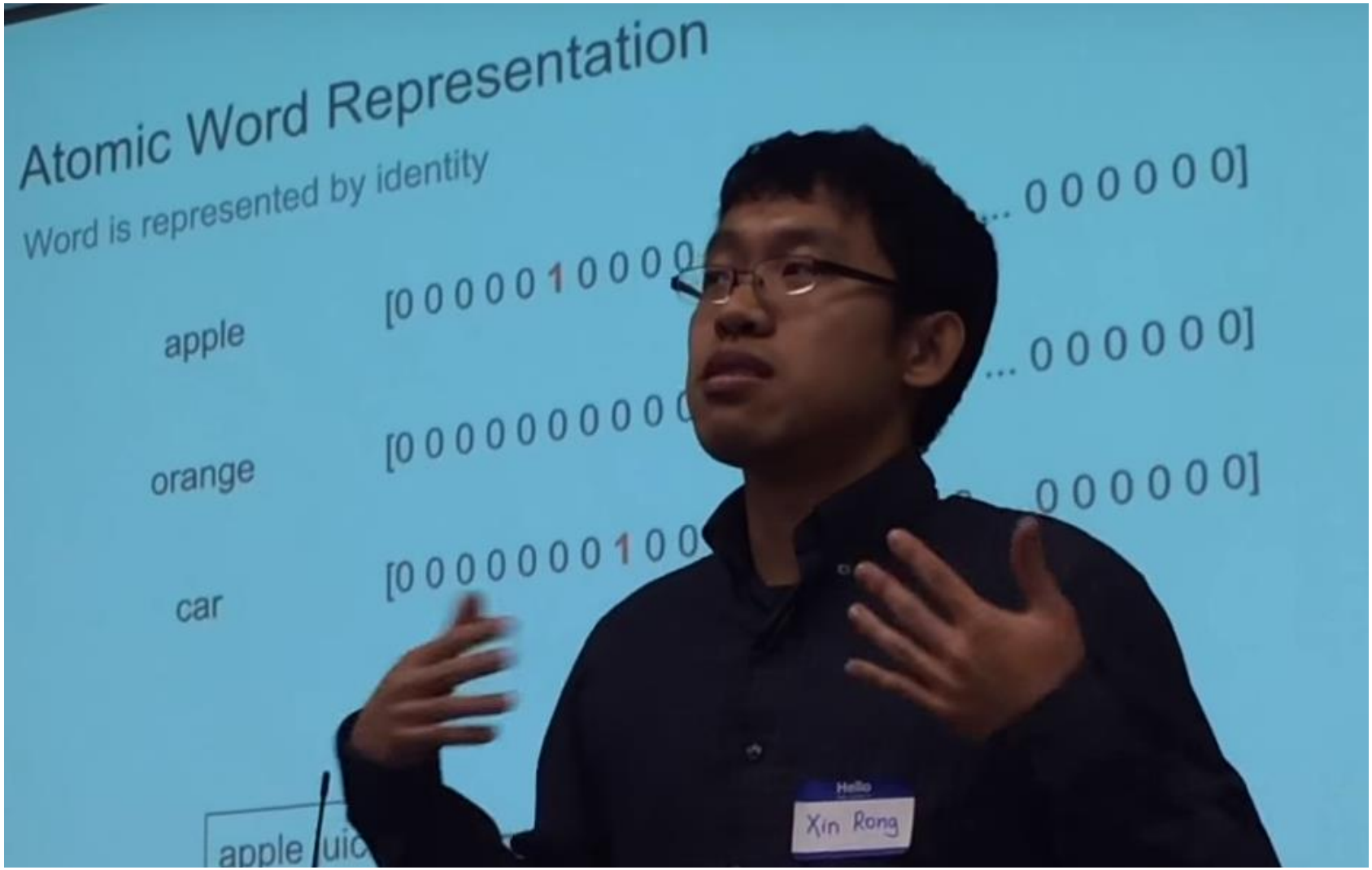




https://ronxin.github.io/wevi/

Word Embedding

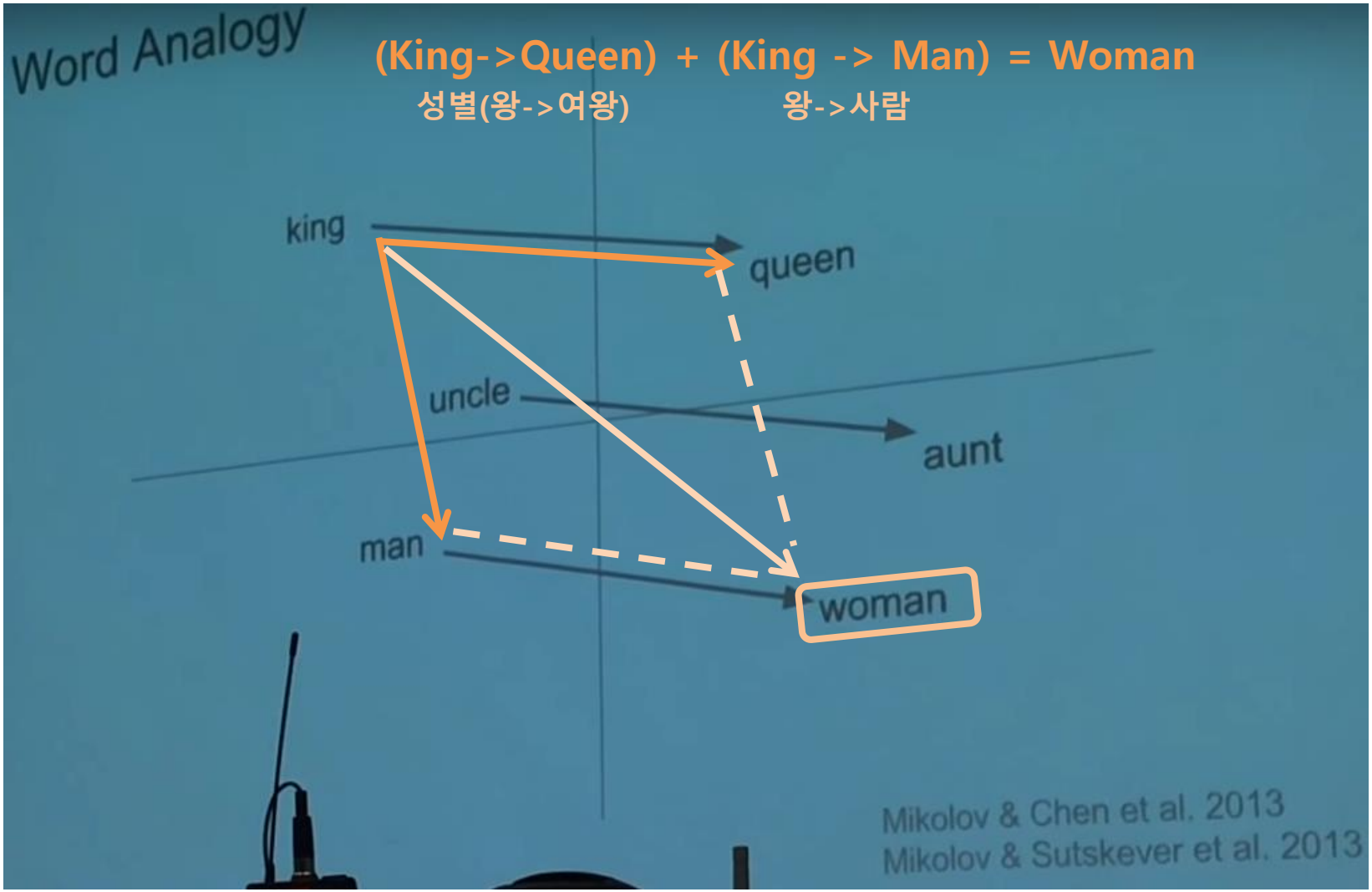
B. Word2Vec - word2vec Parameter Learning Explained, Xin Rong



https://ronxin.github.io/wevi/

Word Embedding

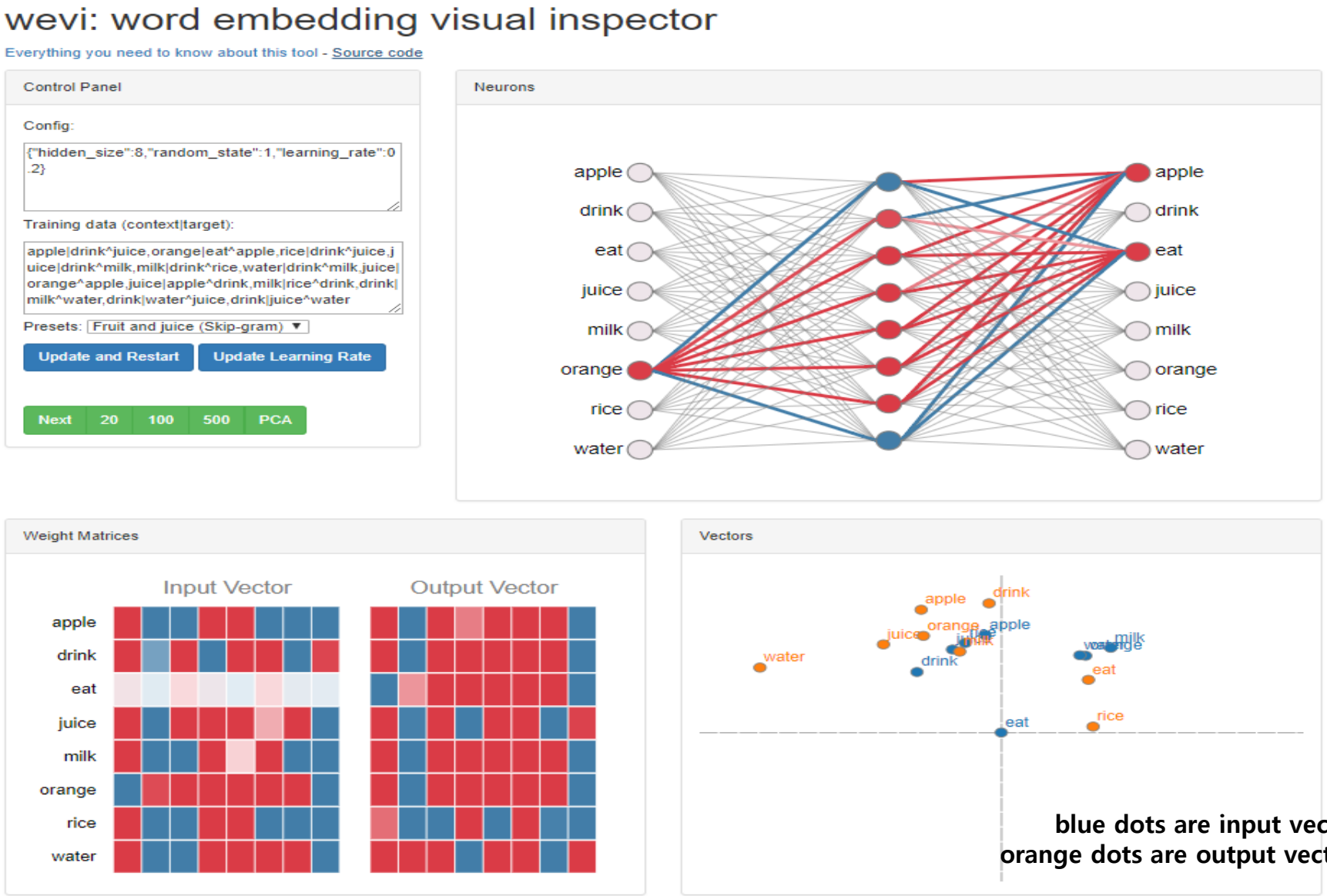
B. Word2Vec - word2vec Parameter Learning Explained, Xin Rong



https://ronxin.github.io/wevi/

Word Embedding

B. Word2Vec 란?



https://ronxin.github.io/wevi/

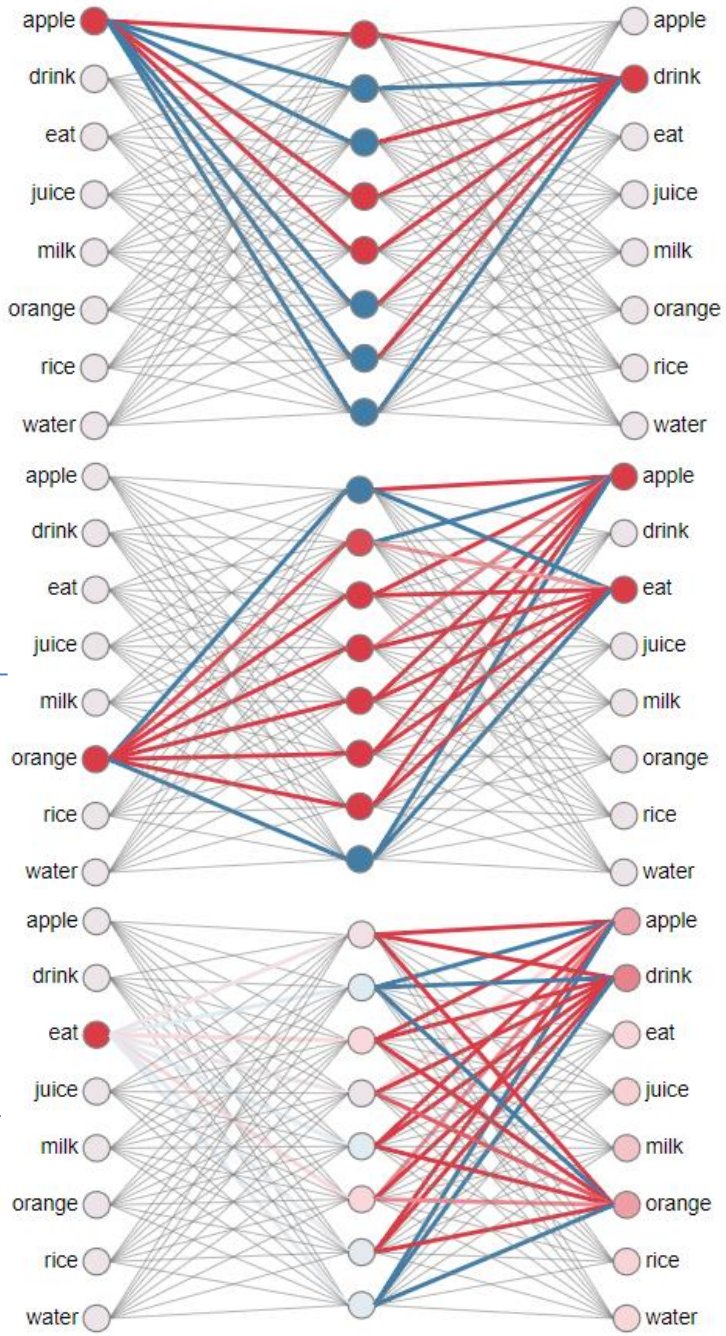
Word Embedding

B. Word2Vec 란?

{"hidden_size":8,"random_state":1,
"learning_rate":0.2}

Training data (context|target):

apple|drink^juice,
orange|eat^apple,
rice|drink^juice,
juice|drink^milk,
milk|drink^rice,
water|drink^milk,
juice|orange^apple,
juice|apple^drink,
milk|rice^drink,
drink|milk^water,
drink|water^juice,
drink|juice^water



https://ronxin.github.io/wevi/, https://github.com/ronxin/wevi

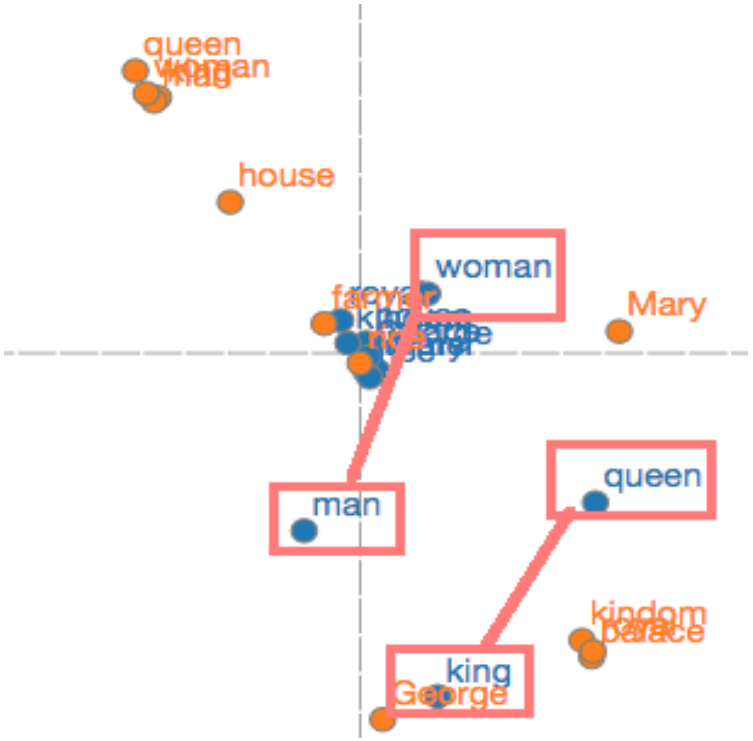
Word Embedding

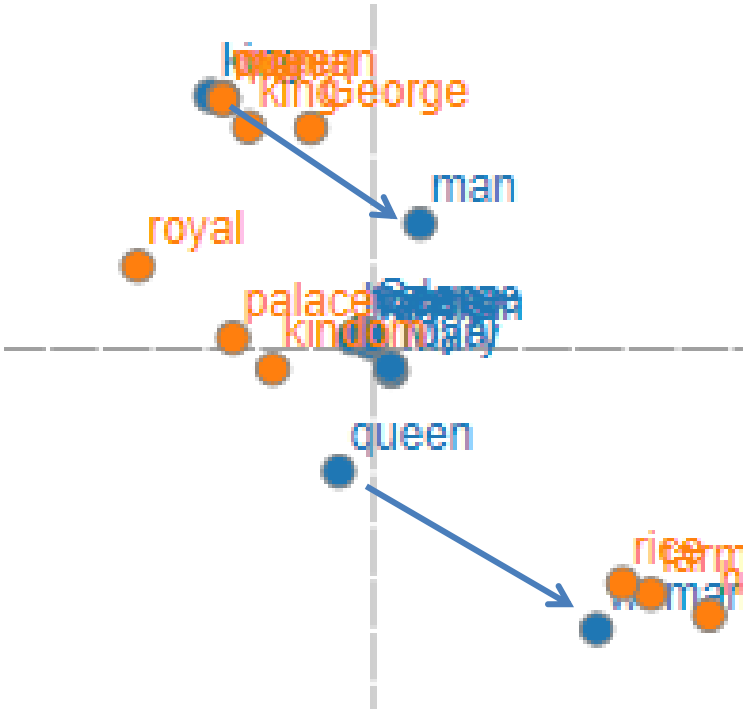
B. Word2Vec 란?

king|kindom,queen|kindom,king|palace,queen|palace,king|royal,queen|royal,king|George,queen|Mary,man|rice,woman|rice,man|farmer,woman|farmer,man|house,woman|house,man|George,woman|Mary

https://ronxin.github.io/wevi/

you can see the infamous analogy: "king - queen = man - woman





https://www.youtube.com/watch?v=D-ekE-Wlcds&feature=youtu.be

https://code.google.com/archive/p/word2vec/

Google Code Archive

Search this site

Projects Search About

Project

Source

Issues

Wikis

Downloads

word2vec

Tool for computing continuous distributed representations of words.

Introduction

This tool provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These representations can be subsequently used in many natural language processing applications and for further research.

Quick start

- Download the code: svn checkout http://word2vec.googlecode.com/svn/trunk/
- Run 'make' to compile word2vec tool
- Run the demo scripts: **./demo-word.sh** and **./demo-phrases.sh**
- For questions about the toolkit, see http://groups.google.com/group/word2vec-toolkit

Project Information

The project was created on Jul 30, 2013.

- License: Apache License 2.0
- 945 stars
- svn-based source control

Labels:

NeuralNetwork MachineLearning NaturalLanguageProcessing WordVectors Google

I know I can do this.

Word Embedding - Exercise 1 – Kaggle Movie Review Data

kaggle | Search kaggle | Competitions | Datasets | Kernels | Discussion | Jobs | ...

Dataset

Word2Vec tutorial - Suite

Training the model -

KarimBELAYATI • last updated 3 months ago

0

Overview | Data | Kernels | Discussion | Activity | Download (54 MB) | New Kernel

Data Set

The labeled data set consists of 50,000 IMDB movie reviews specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 review labeled training set does not include any of the same movies as the 25,000 review test set. In addition, there are another 50,000 IMDB reviews provided without any rating labels.

File descriptions

labeledTrainData - The labeled training set. The file is tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review.
testData - The test set. The tab-delimited file has a header row followed by 25,000 rows containing an id and text for each review. Your task is to predict the sentiment for each one. unlabeledTrainData - An extra training set with no labels. The tab-delimited file has a header row followed by 50,000 rows containing an id and text for each review. sampleSubmission - A comma-delimited sample submission file in the correct format. Data fields

id - Unique ID of each review sentiment - Sentiment of the review; 1 for positive reviews and 0 for negative reviews review - Text of the review

https://www.kaggle.com/belayati/word2vec-tutorial-suite
https://ratsgo.github.io/natural%20language%20processing/2017/03/08/word2vec/

Word Embedding - Exercise 1

이름	수정한 날짜	유형	크기
labeledTrainData.tsv	2017-06-23 오후...	TSV 파일	32,770KB

	A	B	C
1	id	sentiment	review
2	5814_8	1	With all this stuff going down at the moment with MJ i've started listening to his music, watching the odd documentary here and ther
3	2381_9	1	₩The Classic War of the Worlds₩" by Timothy Hines is a very entertaining film that obviously goes to great effort and lengths to faithf
4	7759_3	0	The film starts with a manager (Nicholas Bell) giving welcome investors (Robert Carradine) to Primal Park . A secret project mutating a
5	3630_4	0	It must be assumed that those who praised this film (₩the greatest filmed opera ever,₩" didn't I read somewhere?) either don't care fo
6	9495_8	1	Superbly trashy and wondrously unpretentious 80's exploitation, hooray! The pre-credits opening sequences somewhat give the false ir
7	8196_8	1	I dont know why people think this is such a bad movie. Its got a pretty good plot, some good action, and the change of location for H
8	7166_2	0	This movie could have been very good, but comes up way short. Cheesy special effects and so-so acting. I could have looked past that
9	10633_1	0	I watched this video at a friend's house. I'm glad I did not waste money buying this one. The video cover has a scene from the 1975 m
10	319_1	0	A friend of mine bought this film for 짙1, and even then it was grossly overpriced. Despite featuring big names such as Adam Sandler,
11	8713_10	1	 This movie is full of references. Like ₩Mad Max II₩", ₩"The wild one₩" and many others. The ladybug쨈s face it쨈s a cle
12	2486_3	0	What happens when an army of wetbacks, towelheads, and Godless Eastern European commies gather their forces south of the borde
13	6811_10	1	Although I generally do not like remakes believing that remakes are waste of time; this film is an exception. I didn't actually know so f
14	11744_9	1	₩Mr. Harvey Lights a Candle₩" is anchored by a brilliant performance by Timothy Spall. While we can predict that his titu
15	7369_1	0	I had a feeling that after ₩Submerged₩", this one wouldn't be any better... I was right. He must be looking for champagne money, anc
16	12081_1	0	note to George Litman, and others: the Mystery Science Theater 3000 riff is ₩I don't think so, *breeder*₩". my favorite riff
17	3561_4	0	Stephen King adaptation (scripted by King himself) in which a young family, newcomers to rural Maine, find out about the pet cemete
18	4489_1	0	`The Matrix' was an exciting summer blockbuster that was visually fantastic but also curiously thought provoking in its `Twilight Zone'-
19	3951_2	0	Ulli Lommel's 1980 film 'The Boogey Man' is no classic, but it's an above average low budget chiller that's worth a look. The sequel, 19
20	3304_10	1	This movie is one among the very few Indian movies, that would never fade away with the passage of time, nor would its spell bindinc
21	9352_10	1	Most people, especially young people, may not understand this film. It looks like a story of loss, when it is actually a story about being
22	3374_7	1	₩Soylent Green₩" is one of the best and most disturbing science fiction movies of the 70's and still very persuasive even by today's sta
23	10782_7	1	Michael Stearns plays Mike, a sexually frustrated individual with an interesting moral attitude towards sexuality. He has no problem og
24	5414_10	1	This happy-go-luck 1939 military swashbuckler, based rather loosely on Rudyard Kipling's memorable poem as well as his novel ₩Soldi
25	10492_1	0	I would love to have that two hours of my life back. It seemed to be several clips from Steve's Animal Planet series that was spliced int
26	3350_3	0	The script for this movie was probably found in a hair-ball recently coughed up by a really old dog. Mostly an amateur film with lame
27	6581_7	1	Looking for Quo Vadis at my local video store, I found this 1985 version that looked interesting. Wow! It was amazing! Very much a Ke
28	2203_3	0	Note to all mad scientists everywhere: if you're going to turn your son into a genetically mutated monster, you need to give him a scar
29	680_1	0	What the ... is this ? This must, without a doubt, be the biggest waste of film, settings and camera ever. I know you can't set your a

labeledTrainData

Word Embedding - Exercise 1

이름	수정한 날짜	유형	크기
labeledTrainData.tsv	2017-06-23 오후...	TSV 파일	32,770KB
sampleSubmission.csv	2017-10-02 오후...	Microsoft Excel ...	277KB
testData.tsv	2017-06-23 오후...	TSV 파일	31,958KB
unlabeledTrainData.tsv	2017-06-23 오후...	TSV 파일	65,705KB

	A	B
1	id	review
2	9999_0	Watching Time Chasers, it obvious that it was made by a bunch of friends. Maybe they were sitting around one day in film school and said, ₩He
3	45057_0	I saw this film about 20 years ago and remember it as being particularly nasty. I believe it is based on a true incident: a young man breaks into a
4	15561_0	Minor Spoilers In New York, Joan Barnard (Elvire Audrey) is informed that her husband, the archeologist Arthur Barnard (John Saxon
5	7161_0	I went to see this film with a great deal of excitement as I was at school with the director, he was even a good friend of mine for a while. But sor
6	43971_0	Yes, I agree with everyone on this site this movie is VERY VERY bad. To even call this a movie is an insult to all movies ever made. It's 40 minutes
7	36495_0	Jennifer Ehle was sparkling in ₩Pride and Prejudice.₩" Jeremy Northam was simply wonderful in ₩"The Winslow Boy.₩" With actors of this calibe
8	49472_0	Amy Poehler is a terrific comedian on Saturday Night Live, but her role in this movie doesn't give her anything to work with. Her character, a pub
9	36693_0	A plane carrying employees of a large biotech firm--including the CEO's daughter--goes down in thick forest in the Pacific Northwest. When the
10	316_0	A well made, gritty science fiction movie, it could be lost among hundreds of other similar movies, but it has several strong points to keep it nea
11	32454_0	Incredibly dumb and utterly predictable story of a rich teen girl who, not given love by her parents, starts a girl gang. They rob gas stations, rape
12	37128_0	After reading the comments for this movie, I am not sure whether I should be angry, sad or sickened. Seeing comments typical of people who a)
13	19439_0	It's hard to describe Elfen Lied to someone who hasn't heard of it; the closest I can think of is a strange mix of extreme violence, romance and dr
14	10760_0	Of all the bile-inducing vomitoriums to squeeze out of Tinseltown's bloody sphincter over the past 50 years, this is indisputably the worst of the
15	15073_0	This is quite an underrated Hitchcock movie. While its not one of his best movies, this is actually quite good. The plot is very World War 2-inspire
16	33119_0	Being a huge Gary Oldman fan, I had high hopes for this film. And the ₩Deleverence/Straw Dogs₩"-esquire story about an English couple vacati
17	38735_0	For the most part the acting was poorly done (especially on part of the Park Ranger), and the stories were anything but scary. Predictable and rat
18	12041_0	Ram Gopal Varma does it again yet another movie which draws you into the reality of Underworld and introduces you to the machinery of Mum
19	41565_0	I gave it 2 for some attempt at historical detail... To echo another reviewer, all the actors behave as in the 20th/21st centuries, not the Victorian 1
20	48612_0	I wanted to watch this one sense I am a fan of films such as ₩Last House on the Left₩" and ₩"Maniac₩". This film follows in this same gener on
21	17525_0	Che is a good film for chronicling the events leading to the Cuban Revolution and certainly has a mystic potency surrounding Che and his politi
22	32062_0	This film is very disturbing. It reminded me of the love and affection and materialistic things that we all want for ourselves. It's a movie about ho
23	48625_0	Let me put it this way. I love bad movies. I pretty much grew up with MST3K since I was about 10. I often rent cheesy horror and action films for
24	48176_0	From the combined ministries of Jack and Rexella Van Impe and John Hagee comes Revelation, the second in a trilogy of films about somebody'
25	16232_0	Ghost World is one of the better teen or comming of age movie i have seen in a while. The story was very sweet, and it gives people who are ho
26	12528_0	Outstanding film dealing with a 16 year old who tries to bring stability to a totally dysfunctional family. There is a truly outstanding
27	39976_0	This movie did play like a dream or maybe even like an autobiography - it was disjointed. It was basically a bunch of mostly random scenes thro
28	25857_0	This film is supposedly based on the true account of a prisoner who escaped from Devil's Island in French Guiana. However, the book by Henri C

unlabeledTrainData

준비 개수: 99995 100%

Word Embedding - Exercise 1

이름	수정한 날짜	유형	크기
labeledTrainData.tsv	2017-06-23 오후...	TSV 파일	32,770KB
sampleSubmission.csv	2017-10-02 오후...	Microsoft Excel ...	277KB
testData.tsv	2017-06-23 오후...	TSV 파일	31,958KB
unlabeledTrainData.tsv	2017-06-23 오후...	TSV 파일	65,705KB

	A	B
1	id	review
2	12311_10	Naturally in a film who's main themes are of mortality, nostalgia, and loss of innocence it is perhaps not surprising that it is rated more highly by
3	8348_2	This movie is a disaster within a disaster film. It is full of great action scenes, which are only meaningful if you throw away all sense of reality. Let
4	5828_4	All in all, this is a movie for kids. We saw it tonight and my child loved it. At one point my kid's excitement was so great that sitting was impossib
5	7186_2	Afraid of the Dark left me with the impression that several different screenplays were written, all too short for a feature length film, then spliced
6	12128_7	A very accurate depiction of small time mob life filmed in New Jersey. The story, characters and script are believable but the acting drops the ball
7	2913_8	...as valuable as King Tut's tomb! (OK, maybe not THAT valuable, but worth hunting down if you can). I notice no one has commented on this mo
8	4396_1	This has to be one of the biggest misfires ever...the script was nice and could have ended a lot better.the actors should have played better and m
9	395_2	This is one of those movies I watched, and wondered, why did I watch it? What did I find so interesting about it? Being a truck driver myself, I di
10	10616_1	The worst movie i've seen in years (and i've seen a lot of movies). Acting is terrible, there is no plot whatsoever, there is no point whatsoever, i fel
11	9074_9	Five medical students (Kevin Bacon, David Labraccio; William Baldwin, Dr. Joe Hurley; Oliver Platt, Randy Steckle; Julia Roberts, Dr. Rachel Mannus
12	9252_3	'The Mill on the Floss' was one of the lesser novels by Mary Ann Evans, who wrote under the male pseudonym George Eliot. I tried to read this d
13	9896_9	I just saw this film at the phoenix film festival today and loved it. The synopsis was listed in our program as ₩an old Shakespearean actor invites
14	574_4	₩The Love Letter₩" is one of those movies that could have been really clever, but they wasted it. Focusing on a letter wreaking havoc in a small
15	11182_8	Another fantastic offering from the Monkey Island team and though it was a long time coming and had to survive the departure of Ron Gilbert
16	11656_4	This was included on the disk ₩Shorts: Volume 2₩"--a rather dull collection of short films. Shorts are among my favorite style of films but someh
17	2322_4	I'm not really much of an Abbott & Costello fan (although I do enjoy ₩Who's On First₩") and, to be honest, there wasn't much in this movie tha
18	8703_1	This movie was dreadful. Biblically very inaccurate. Moses was 80 years old when he led the people out of Egypt, the movie has him about forty.
19	7483_1	I don't think I've ever gave something a 1/10 rating, but this one easily gets the denomination. I find it hard just to sit through one of his jokes. I
20	6007_10	Excellent story-telling and cinematography. Poignant, biting social commentary. Superb effects. Well-filmed and acted.
21	12424_4	I completely forgot that I'd seen this within a couple of days, which is pretty revealing in itself. The umpteenth version of Gaston 'Phantom of the
22	4672_1	I like action movies. I have a softspot for ₩B₩" flicks with bad dialogue and wooden acting. So, I've been wracking my brain to come up with on
23	10841_3	This is one of the worst Sandra Bullock movie since Speed 2 But not quite that bad. I really lost it with those out of the blue not so ₩special effe
24	8954_7	Watched this flick on Saturday afternoon cable. Man, did it drag. I got the metaphors, symbolism, and all that stuff. No, I didn't care one way or a
25	7392_1	I went to see ₩TKIA₩" with high expectations, which might have influence on my opinion on it. I have seen all of the Dogme films, and this TKIA
26	10288_8	All credit to writer/director Gilles Mimouni who fashioned this winding, twisting tale of deceit and betrayal. While keeping the utmost control, he
27	5343_4	As a writing teacher, there are two ending I never allow my students to use: ₩Then I woke up₩" and ₩"Then I Got Run Over by a Truck.₩" I am
28	4950_1	I don't know why this has gotten any decent reviews as it could be the weakest horror comedy I've ever seen. Englund is just in it for a cameo an
29	9257_4	This film was released in the UK under the name Blood Bites. It was banned outright and never submitted again for release. As The

testData

Word Embedding - Exercise 1

Read 25000 labeled train reviews, 25000 labeled test reviews, and 50000 unlabeled reviews
2017-10-02 22:24:42,461 : INFO : collecting all words and their counts
2017-10-02 22:24:45,961 : INFO : collected 123504 word types from a corpus of 17798082 raw words and 795538 sentences
2017-10-02 22:24:46,110 : INFO : estimated required memory for 16490 words and 300 dimensions: 47821000 bytes
2017-10-02 22:24:46,390 : INFO : training model with 6 workers on **16490 vocabulary** and **300 features**, using sg=0 hs=0 sample=1e-05 negative=5 **window=10**
2017-10-02 22:25:43,623 : INFO : saved 300features_40minwords_10context

이름	수정한 날짜	유형	크기
data	2000-11-22 오전...	파일 폴더	
300features_40minwords_10context	2000-11-22 오전...	파일	59,280KB
run_log.txt	2000-11-22 오전...	텍스트 문서	25KB
sentences.pickle	2000-11-22 오전...	PICKLE 파일	246,172KB
w2v_sample.py	2000-11-23 오전...	PY 파일	5KB

```
# For the second time
model = word2vec.Word2Vec.load("300features_40minwords_10context")

print("Q1. doesnt_match [man woman child kitchen]")
print(model.doesnt_match("man woman child kitchen".split()))
print("Q2. doesnt_match [france england germany berlin]")
print(model.doesnt_match("france england germany berlin".split()))
print("Q3. doesnt_match [paris berlin london china]")
print(model.doesnt_match("paris berlin london china".split()))

print("Q4. most_similar [man]")
print(model.most_similar("man"))
print("Q5. most_similar [queen]")
print(model.most_similar("queen"))
print("Q6. most_similar [terrible]")
print(model.most_similar("terrible"))
```

Word Embedding - Exercise 1

[결과]

Q1. doesnt_match [man woman child kitchen]
Kitchen

Q2. doesnt_match [france england germany berlin]
Berlin

Q3. doesnt_match [paris berlin london china]
China

Q4. most_similar [man]
[('murderer', 0.9646390676498413), ('seeks', 0.9643397927284241), ('priest', 0.9583220481872559), ('obsessed', 0.9529876708984375), ('patient', 0.9518172144889832), ('accused', 0.9511740207672119), ('prostitute', 0.9504649043083191), ('determined', 0.9503258466720581), ('lonely', 0.94843989610672), ('learns', 0.9481915235519409)]

Q5. most_similar [queen]
[('preston', 0.9884580373764038), ('duke', 0.9870703220367432), ('belle', 0.9855383634567261), ('princess', 0.9838896989822388), ('sally', 0.9837985038757324), ('karl', 0.9834704399108887), ('marshall', 0.9831289649009705), ('cole', 0.9830288887023926), ('virginia', 0.9829562306404114), ('veronica', 0.9828841686248779)]

Q6. most_similar [terrible]
[('horrible', 0.9897130131721497), ('awful', 0.976392924785614), ('lame', 0.9656811952590942), ('horrid', 0.9629285335540771), ('alright', 0.9597908854484558), ('boring', 0.9586129188537598), ('mess', 0.9550553560256958), ('cringe', 0.9532005786895752), ('badly', 0.9410943388938904), ('horrendous', 0.9403845071792603)]

https://ratsgo.github.io/natural%20language%20processing/2017/03/08/word2vec/

Word Embedding - Exercise 1

```
model.wv.save_word2vec_format('300features_40minwords_10context.txt', binary=False)
```

A16491 | disingenuous 0.065291 -0.011626 0.104388 -0.013313 0.065584 -0.053475 -0.047935 0.052099 -0.039

16469	inhumanity 0.062615 0.000141 0.118493 -0.016572 0.060124 -0.042580 -0.061475 0.043932 -0.047785 0.041030 -0.033785 -0.033839 0.0
16470	objection 0.082133 0.006591 0.126224 -0.001112 0.056210 -0.071137 -0.049406 0.057289 -0.064228 0.040781 -0.034007 -0.028995 0.03
16471	uncensored 0.059816 -0.007009 0.111052 -0.010294 0.034936 -0.037102 -0.048609 0.084554 -0.046209 0.021842 -0.034450 -0.016625 0
16472	lookout 0.065192 -0.001321 0.104255 -0.009723 0.036137 -0.041666 -0.039423 0.075729 -0.068599 0.023492 -0.029275 -0.014758 0.038
16473	uncovering 0.082729 0.033476 0.117745 0.009669 0.031754 -0.056303 -0.046046 0.051891 -0.098918 0.014284 -0.023918 -0.022070 0.07
16474	colorless 0.051449 -0.008189 0.075467 -0.012868 0.060212 -0.028053 -0.040537 0.037834 -0.050642 0.025742 -0.016234 -0.036306 0.04
16475	dominatrix 0.079583 0.027954 0.090587 0.034038 0.032119 -0.046981 -0.035562 0.058485 -0.102870 0.001484 -0.016309 -0.026510 0.07
16476	dong 0.072526 0.003858 0.099715 -0.005476 0.041275 -0.037135 -0.042155 0.064802 -0.076078 0.027900 -0.024414 -0.027716 0.042206
16477	grable 0.045424 -0.010232 0.068509 -0.010923 0.035688 -0.019364 -0.024808 0.054730 -0.059797 0.012360 -0.011559 -0.022958 0.0402
16478	pandora 0.054560 -0.001376 0.095983 -0.012359 0.042158 -0.019393 -0.050921 0.068880 -0.058302 0.016643 -0.028182 -0.026255 0.03
16479	insatiable 0.076519 0.014951 0.099842 0.013946 0.046610 -0.047857 -0.047781 0.052447 -0.081810 0.018840 -0.022099 -0.033795 0.057
16480	farts 0.098136 0.020354 0.109245 0.036706 0.046962 -0.073489 -0.045725 0.075917 -0.082278 0.016028 -0.032291 -0.024917 0.052259 0
16481	captions 0.060328 -0.015657 0.112117 -0.017126 0.055934 -0.031616 -0.063998 0.069225 -0.030888 0.032269 -0.037303 -0.027586 0.018
16482	snore 0.075110 -0.009759 0.098363 0.011266 0.058267 -0.045813 -0.052961 0.075294 -0.041588 0.025364 -0.030925 -0.027624 0.017363
16483	emilio 0.042342 -0.008230 0.063953 -0.013089 0.037293 0.000343 -0.033365 0.061751 -0.061577 0.012001 -0.013618 -0.029022 0.03888
16484	rewritten 0.050912 -0.029217 0.104865 -0.036348 0.063454 -0.024686 -0.059097 0.062118 -0.005092 0.037802 -0.033321 -0.030523 -0.0
16485	barman 0.062423 -0.014140 0.087392 -0.008390 0.063597 -0.030009 -0.044421 0.064284 -0.035645 0.030124 -0.027477 -0.030743 0.024
16486	briggs 0.051813 -0.008025 0.064855 -0.000556 0.023652 -0.019824 -0.022766 0.076636 -0.079579 0.012051 -0.014689 -0.016837 0.0378
16487	faris 0.068148 -0.010047 0.075907 -0.000659 0.049653 -0.044018 -0.021908 0.065353 -0.064937 0.023574 -0.016363 -0.028781 0.028573
16488	patchy 0.036621 -0.036722 0.068435 -0.026754 0.077924 -0.005753 -0.059377 0.043017 0.006042 0.038097 -0.025821 -0.042218 0.00302
16489	wasps 0.069448 0.029793 0.093991 0.015921 0.028818 -0.026912 -0.054661 0.067657 -0.088521 0.000179 -0.027535 -0.018127 0.077593
16490	backlash 0.055496 -0.007321 0.111126 -0.025181 0.047458 -0.033871 -0.059793 0.066283 -0.049800 0.041193 -0.038003 -0.023593 0.03
16491	disingenuous 0.065291 -0.011626 0.104388 -0.013313 0.065584 -0.053475 -0.047935 0.052099 -0.039893 0.045941 -0.031212 -0.035681

16490 x 300

Word Embedding - Exercise 2

Papers.csv

This file contains one row for each of the 403 NIPS papers from this year's conference. It includes the following fields

- **Id** - unique identifier for the paper (equivalent to the one in NIPS's system)
- **Title** - title of the paper
- **EventType** - whether it's a poster, oral, or spotlight presentation
- **PdfName** - filename for the PDF document
- **Abstract** - text for the abstract (scraped from the NIPS website)
- **PaperText** - raw text from the PDF document (created using the tool pdftotext)

Authors.csv

This file contains id's and names for each of the authors on this year's NIPS papers.

- **Id** - unique identifier for the author (equivalent to the one in NIPS's system)
- **Name** - author's name

PaperAuthors.csv

This file links papers to their corresponding authors.

- **Id** - unique identifier
- **PaperId** - id for the paper
- **AuthorId** - id for the author

https://www.kaggle.com/cherishzhang/clustering-on-papers
https://github.com/benhamner/nips-2015-papers/blob/master/src/download_papers.py

Word Embedding - Exercise 2

```
step one: extract keywords from Title, Abstract and PaperText based on tf-idf
step two: keywords are used to build the word2vec model
step three: from keywords to paper document, average the top-n keywords vector to represent the whole paper
```

```
papers_data['Title_clean'] = papers_data['Title'].apply(lambda x:clean_text(x))
papers_data['Abstract_clean'] = papers_data['Abstract'].apply(lambda x:clean_text(x))
papers_data['PaperText_clean'] = papers_data['PaperText'].apply(lambda x: clean_text(x))



#title2kw = extract_tfidf_keywords(papers_data['Title_clean'],3)
abstract2kw = extract_tfidf_keywords(papers_data['Abstract_clean'], 20)
text2kw = extract_tfidf_keywords(papers_data['PaperText_clean'],100)

print ("[abstract2kw]", abstract2kw)
print ("[text2kw]", text2kw)
```

(TFIDF) 403개의 문서-> abstract, papertext-> 각 20개, 100개 씩 키워드 추출

[
['possibl', 'onli', 'data', 'qualiti', 'involv', 'reduct', 'multipl', 'label', 'popular', 'address', 'lower', 'fast', 'challeng', 'form', 'machin learn', 'low', 'obtain', 'rate', 'natur', 'make'],
['loss', 'convex', 'robust', 'classif', 'strong', 'solut', 'prove', 'ani', 'paper propos', 'result', 'label', 'nois', 'limit', 'standard', 'make', 'random', 'howev', 'class', 'experi', 'linear'],
...
]

https://www.kaggle.com/cherishzhang/clustering-on-papers
https://github.com/benhamner/nips-2015-papers/blob/master/src/download_papers.py

Word Embedding - Exercise 2

```
"""
    step three:
    average top-n keywords vectors and compute similarities
"""
doc2vecs = []
for i in range(0, len(abstract2kw)):
    vec = [0 for k in range(100)]
    for j in range(0, len(abstract2kw[i])):
        if abstract2kw[i][j] in word2vec_model:
            vec += word2vec_model[abstract2kw[i][j]]

    for j in range(0, len(text2kw[i])):
        if text2kw[i][j] in word2vec_model:
            vec += word2vec_model[text2kw[i][j]]
    doc2vecs.append(vec)

similarities = squareform(pdist(doc2vecs, 'cosine'))
```

https://www.kaggle.com/cherishzhang/clustering-on-papers
https://github.com/benhamner/nips-2015-papers/blob/master/src/download_papers.py

Word Embedding - Exercise 2

```
"""
    k-means clustering and wordcloud(it can combine topic-models
    to give somewhat more interesting visualizations)
"""
num_clusters = 10
km = KMeans(n_clusters=num_clusters)
km.fit(doc2vecs)
clusters = km.labels_.tolist()
```

	A	B	C	D
1	Id	Title	EventType	Cluster
2	5802	Planar Ultrametrics for Image Segmentation	Poster	7
3	5677	Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing	Poster	0
4	6019	Algorithmic Stability and Uniform Generalization	Poster	2
5	5814	Parallel Correlation Clustering on Big Graphs	Poster	9
6	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8
7	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8
8	5714	Robust Portfolio Optimization	Poster	5
9	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8
10	6019	Algorithmic Stability and Uniform Generalization	Poster	2
11	5802	Planar Ultrametrics for Image Segmentation	Poster	7
12	6019	Algorithmic Stability and Uniform Generalization	Poster	2
13	5937	Logarithmic Time Online Multiclass prediction	Spotlight	6
14	5941	Learning with Symmetric Label Noise: The Importance of Being Unhinged	Spotlight	1
15	5802	Planar Ultrametrics for Image Segmentation	Poster	7
16	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8
17	5937	Logarithmic Time Online Multiclass prediction	Spotlight	6
18	5941	Learning with Symmetric Label Noise: The Importance of Being Unhinged	Spotlight	1
19	5941	Learning with Symmetric Label Noise: The Importance of Being Unhinged	Spotlight	1
20	5802	Planar Ultrametrics for Image Segmentation	Poster	7
21	5714	Robust Portfolio Optimization	Poster	5
22	5941	Learning with Symmetric Label Noise: The Importance of Being Unhinged	Spotlight	1
23	6019	Algorithmic Stability and Uniform Generalization	Poster	2
24	5937	Logarithmic Time Online Multiclass prediction	Spotlight	6
25	5677	Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing	Poster	0
26	5714	Robust Portfolio Optimization	Poster	5
27	5714	Robust Portfolio Optimization	Poster	5
28	5937	Logarithmic Time Online Multiclass prediction	Spotlight	6
29	5937	Logarithmic Time Online Multiclass prediction	Spotlight	6
30	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8
31	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8
32	5776	Expressing an Image Stream with a Sequence of Natural Sentences	Poster	8

https://www.kaggle.com/cherishzhang/clustering-on-papers
https://github.com/benhamner/nips-2015-papers/blob/master/src/download_papers.py

[CASE1 : 사례 기반 법적용 해석] - 법 데이터, 비조치 의견(금융위), 판례 등

법령해석

회신일 ~ 진행단계 전체

분야 전체 법령명 법령명검색

검색조건 항목선택

검색

번호	기관	분야	제목	일련번호	진행단계
985	금융위원회	공통	VAN결제 방식 간소화(생략, 축소)와 이를 가맹점에 적용…	170417	완료
984	금융위원회	공통	실질주주명세의 금융거래정보 해당 여부 질의	170414	완료
983	금융위원회	공통	지급보증 업무 영위 가능 여부	170410	완료
982	금융위원회	공통	더치페이 서비스 관련 법령해석 요청	170408	완료
981	금융위원회	공통	테스트베드 통과 로보어드바이저의 수익률 광고 가능 여부	170376	완료
980	금융위원회	공통	전자금융업자가 다른 전자금융업자 또는 전자금융보조업…	170372	완료
979	금융위원회	공통	사외이사의 결격사유 중 '주된 자문계약'의 범위 해석	170345	완료
978	금융위원회	공통	여신전문금융회사의 합병에 따른 절차	170341	완료
977	금융위원회	공통	여신전문금융업법상 대주주의 "특수관계인" 정의	170342	완료
976	금융위원회	공통	여신전문금융업법 제54조의3제1항제3호의 사후보고 요…	170333	완료

1 2 3 4 5 6 7 8 9 10

금융규제민원포털

통합검색 검색

법령해석 및 비조치의견 | 금융개혁 현장점검 | 옴부즈만

법령해석 및 비조치의견서
- 제도 개요
- 요청하기
- 회신사례
- 회신현황
- 법령해석 비조치의견서
- 법령해석 관련 정보

홈 > 법령해석 및 비조치의견서 > 회신현황 > 법령해석

법령해석

- 회신

제목	매수평균단가 미만 매도시 매도수수료
처리구분	완료
소관부서	
등록자	관리자
회신일	2017-07-25
첨부파일	회신문(170220).hwp (0.01Mbyte)
질의요지	○ 투자자가 개별 주식 매도시 손실*을 본 경우, 해당 거래에 대한 거래수수료를 면제하는 것이 「자본시장과 금융투자업에 관한 법률」 상 손실보전 금지 규제, 정당한 사유 없는 투자자간 차별 금지 규제에 위반되는지 여부 * 손실 여부는 투자자의 평균매수가격을 산정하고 평균매수가격 이하의 가격으로 매도하는 것으로 판단
회답	□ 「자본시장과 금융투자업에 관한 법률」 (이하 '법'이라 합니다) 제55조의 '손실의 보전'이란 투자자에 대하여 원본의 전부 또는 일부를 보장하는 일체의 행위를 의미하는바, 금융투자업자가 투자자에 대하여 수수료를 할인하거나 면제하는 행위는 수수료 할인 · 면제 약정의 시기, 수수료 할인 · 면제의 동기 또는 목적, 수수료 · 할인 · 면제행위와 손실보전의 결과 사이의 인과관계, 금융투자업자와 해당 투자자 간의 거래기간 등을 종합적으로 고려하여 원본의 전부 또는 일부의 보장과 동일한 결과가 발생하는 예외적인 상황에 해당하지 않는 한, 원칙적으로 법 제55조의 손실의 보전에 해당하지 아니함을 알려드립니다.
이유	위 회답과 같음

금융위원회 공통 VAN결제 방식 간소화(생략, 축소)와 이를 가맹점에 적용… 170417 완료

http://better.fsc.go.kr/user/extra/fsc/123/fsc_lawreq/view/jsp/LayOutPage.do?lawreqIdx=1612

(수집) WEB Page -> HTML FILES

Project › 20170918_AI_LAW › Sourcecode › fsc_Crawling › output_crawling › fsc_lawreq_html_crawling

fsc_lawreq_html

이름	수정한 날짜	유형	크기
280.html	2000-11-08 오전...	HTML 파일	4KB
281.html	2000-11-08 오전...	HTML 파일	3KB
282.html	2000-11-08 오전...	HTML 파일	3KB
283.html	2000-11-08 오전...	HTML 파일	2KB
284.html	2000-11-08 오전...	HTML 파일	5KB
285.html	2000-11-08 오전...	HTML 파일	3KB
286.html	2000-11-08 오전...	HTML 파일	4KB
292.html	2000-11-08 오전...	HTML 파일	3KB
294.html	2000-11-08 오전...	HTML 파일	70KB
296.html	2000-11-08 오전...	HTML 파일	3KB
297.html	2000-11-08 오전...	HTML 파일	3KB
298.html	2000-11-08 오전...	HTML 파일	3KB
299.html	2000-11-08 오전...	HTML 파일	3KB
300.html	2000-11-08 오전...	HTML 파일	5KB
301.html	2000-11-08 오전...	HTML 파일	4KB
302.html	2000-11-08 오전...	HTML 파일	4KB
303.html	2000-11-08 오전...	HTML 파일	73KB
304.html	2000-11-08 오전...	HTML 파일	10KB
305.html	2000-11-08 오전...	HTML 파일	3KB
312.html	2000-11-08 오전...	HTML 파일	4KB
313.html	2000-11-08 오전...	HTML 파일	4KB
314.html	2000-11-08 오전...	HTML 파일	4KB
315.html	2000-11-08 오전...	HTML 파일	4KB
316.html	2000-11-08 오전...	HTML 파일	4KB
317.html	2000-11-08 오전...	HTML 파일	3KB

(구조분석) #subContent > div > div

```
▼<div class="subContentInner">
  <!--
                    ##########################################################################################
                    #    Content Start                                                                       #
                    ##########################################################################################
                    -->
  <h4 class="brownMinus2">회신</h4>
 ▼<div class="tblType_top6Graypatten view replay mb50">
   <div class="tblType_top7BrownpattenBar"></div>
  ▼<table>
    <caption>회신의 제목, 처리구분, 소관부서, 등록자, 회신일, 답변내용 및 첨부된 파일을 알수 있는 표입니다.</caption>
   ▼<colgroup>
     <col class="col20">
     <col class="colAuto">
    </colgroup>
   ▼<tbody>
    ▼<tr>
      <th scope="row">제목</th>
      <td>매수평균단가 미만 매도시 매도수수료 </td>
    </tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▼<tr>
      <th scope="row">등록자</th>
      <td>관리자</td>
    </tr>
    ▼<tr>
      <th scope="row">회신일</th>
      <td class="engNum">2017-07-25</td>
    </tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▶<tr>…</tr>
    ▼<tr>
      <th scope="row">이유</th>
     ▼<td>
      ▼<p class="minH100">
       ▼<span style="color: black; font-family: 한컴바탕; font-size: 11pt; mso-bidi-font-family: 한컴바탕; mso-ansi-language: KO; mso-hansi-
          <font face="굴림">위 회답과 같음</font>
        </span>
       </p>
      </td>
    </tr>
   </tbody>
  </table>
 </div>
 <!--
                    ##########################################################################################
                    #    Content END                                                                         #
                    ##########################################################################################
                    -->
```

(DATA) CSV FILE 생성 (전제 2238개 질의/답변 데이터)

	A	B
1	NO	DATA
2	280	정보처리시스템 개발업무 사용장소(1) 외부주문등에 의한 정보처리시스템의 개발업무에 사용되는 업무장소와 내부업무장소를 분리하는 경우, 그 분리의 정도 부서를 구분하여 표
3	281	비밀번호 입력방식□ ACS(Auto Call System)를 통한 비밀번호 입력방식이 전자금융감독규정 제33조(이용자 비밀번호 관리) 제2항제4호 내지 5호에서 규정하고 있는 핀패드 등 보안장치 또는 전자적 장치'에 해당하는지 여부□ 귀 사의 ACS는 전자금융감독규정 제33조제2항제4호 또는 제5호에서 규정하고 있는 핀패드(Pin-Pad)와 동등 또는 유사한 방식으로 판단됨□ 핀패드는 통상 고객의 통장개설 신청서, 전표 등에 기재된 비밀번호가 금융회사 직원의 업무처리 또는 전표 폐기과정에서 유출되는 것을 방지하기 위하여, ㅇ 고객이 거래용지 등에 비밀번호를 쓰는 대신 손으로 직접 입력할 수 있게 하는 장치(금융용어사전, 금융감독원)를 의미함 □ 귀 사의 ACS는 고객이 비밀번호를 신청서, 전표 등에 기재하지 않고 고객의 단말기를 이용하여 직접 입력한다는 점에서 본질적으로 핀패드와 동등 또는 유사한 방식으로 판단됨 ㅇ 다만, 귀 사의 ACS는 통상 금융회사의 비밀번호 입력기를 사용하는 통상적 핀패드와 달리 고객의 단말기를 사용한다는 점에서 더욱 엄격한 관리가 필요할 것으로 판단됨
4	282	금융실명법상 본인확인 방법㈜ㅇㅇㅇ은 금융실명거래 및 비밀보장에 관한 법률 제3조에 의한 실지명의 확인 방법으로 모바일 기기를 통한 실시간 영상 통화' 방식을 인정할 수 있는
5	283	단기매매차익 반환 기산일□ ㅇㅇ금융지주(상장법인)와 ㅇㅇ은행(비상장법인)이 합병(14.11.1.)한 후 존속법인인 ㅇㅇ은행이 신규 상장(14.11.18.)된 경우, 합병 전 (舊)ㅇㅇ금융지주
6	284	청약철회 제한 규정 배제 가능여부1. 보험업법 제102조의4제1항 및 동법 시행령 제48조의2제1항제1호에도 불구하고, 진단계약에 대해서도 청약철회를 허용할 수 있는지 여부 2. 보험약관에 진단계약은 청약철회를 할 수 없다'고 기재되어 있음에도 진단계약에 대해 청약철회를 허용할 경우 보험업법 제127조의3(기초서류 기재사항 준수의무) 위반에 해당하는지(질의 1) 보험회사가 진단계약에 대해 청약철회를 허용한다 하여도 보험업법 제102조의4제1항을 위반한 것이 아닙니다. (질의 2) 보험약관에도 불구하고 보험회사가 진단계약에 대해 청약철회를 허용한다 하더라도 이를 보험업법 제127조의3 위반이라 할 수 없습니다. 다만, 이 경우 보험회사는 보험업법 제127조에 따라 진단계약에 대해 청약철회를 허용하는 방향으로 기초서류를 변경하여 운용하는 것이 타당하며, 이렇게 약관의 규정과 실제 운용이 다른 경우 해당 사안에 대한 검사 등을 통해 기초서류 변경을 권고할 수 있을 것입니다.[질의 1] □ 보험업법 제102조의4제1항 및 동법 시행령 제48조의2제1항의 규정은 동법 시행령 제48조의2제1항 각 호에 해당하지 않는 보험계약에 대해서는 보험회사가 청약철회를 거부하여서는 안된다는 규정입니다. □ 따라서, 보험회사가 보험업법 시행령 제48조의2제1항 각 호에 해당하는 보험계약에 대한 청약철회를 허용한다 하여도 보험회사가 보험업법 제102조의4제1항을 위반한 것이라 할 수 없습니다. [질의 2] □ 보험약관의 규정은 계약자는 청약일로부터 15일 이내에 청약을 철회할 수 있으나 진단계약 등에 대해서는 철회할 수 없다는 것으로, 진단계약에 대해 계약자의 청약 철회권을 제한하는 규정입니다. ㅇ 따라서 보험업법 제127조의3은 보험회사의 약관 준수 의무를 규정한 것으로 상기 약관의 규정이 보험회사의 의무라 할 수 없으므로 진단계약의 청약철회에 대해 보험회사가 이를 수용하더라도 보험업법 제127조의3을 위반한 것이라 할 수 없습니다. □ 다만, 제재가 곤란하다 하더라도 보험약관과 다르게 진단계약에 대해 보험회사가 재량적으로 청약철회를 수용할 수 있게 됨으로써 보험계약자를 차별하게 되거나 진단계약의 계약자로서 청약철회 의사가 있음에도 보험약관에 따라 청약철회를 포기하는 경우가 있을 수 있으므로, 보험회사가 진단계약에 대해서도 청약철회를 허용하고자 하는 경우에는 보험업법 제127조에 따라 기초서류를 변경하여 운용하는 것이 타당합니다. □ 추후 등 사안에 대하여 현재와 같이 약관에 기재된 내용과 실제 운용이 다른 경우 해당 사항의 검사 등의 결과를 통해 기초서류의 변경을 권고할 수 있을 것입니다.
7	285	역외금융회사등에 대한 해외직접투자에 있어 신고의무자역외금융회사 등에 대한 해외직접투자에 있어 금융기관의 해외진출에 관한 규정 제7조제1항에 따른 신고의무자가 되는 금융기관이 집합투자업자인지 신탁업자인지투자신탁 관계에 있어서 금융기관의 해외진출에 관한 규정 제7조제1항에 따른 신고의무자는 집합투자업자임□ 투자신탁에서의 신탁업자는 집합투자업자의 지시에 따라 투자대상자산의 취득·처분 등을 하는 수동적 지위에 있는 점 투자신탁의 경우 자본시장과 금융투자업에 관한 법률 등 현행 법체계가 투자신탁의 특수성을 인정하고 있는 점 실무적으로 위탁자인 집합투자업자가 신고를 하고 있는 현실 등을 고려할 때 - 집합투자업자에게 신고의무가 있다고 보는 것이 타당함□ 향후 관련 규정을 개정하여 신고의무자를 명확히 규정할 예정
8	286	특정금전신탁을 통한 해외 파생결합증권 취득가능 여부ㅇ 자본시장법 시행령 제7조제4항제5의2호라목에서 '해당 파생결합증권을 국내에서 매매하는 경우 투자매매업자가 그 파생결합증권을 인수하여 전문투자자에게 이를 취득하도록 하거나 투자중개업자를 통하여 전문투자자에게 그 파생결합증권을 매도'해야 하는 것으로 규정 - 여기서 전문투자자에는 '위탁자가 신탁재산인 금전의 운용방법을 지정하는 특정금전신탁을 운용하는 신탁업자'는 제외 - 이 경우 위탁자가 자본시장법상 '일반투자자'인 경우에 한정되는지 아니면 위탁자가 '전문투자자'인 경우도 포함되는지 여부에 대해 질의ㅇ 위탁자가 자본시장법상 전문투자자인 특정금전신탁을 운용하는 신탁업자는 해외 파생결합증권을 취득할 수 있는 전문투자자에 해당함ㅇ 외국 투자매매업자가 국외에서 발행한 파생결합증권을 국내에서 매매하려면 국내 투자매매업자가 그 파생결합증권을 인수하여 전문투자자에게 이를 취득하도록 하거나 투자중개업자를 통하여 전문투자자에게 그 파생결합증권을 매도하여야 합니다. - 특정금전신탁을 운용하는 신탁업자의 경우 위탁자의 지시에 따라 운용되는 특정금전신탁의 성질에 비추어 볼 때 위탁자가 일반투자자일 경우 전문투자자로 보기 어려우므로, 영 제7조제4항제5호의2라목은 특정금전신탁의 신탁업자를 전문투자자의 범주에서 제외하고 있습니다. - 다만, 동일한 이유로, 위탁자가 자본시장법상 전문투자자에 해당하는 경우에는 특정금전신탁의 신탁업자를 전문투자자의 범주에서 제외할 이유가 없을 것으로 보입니다.
9	292	신용카드의 정의에 실물카드 없는 모바일카드가 포함되는지 여부ㅇ 여신전문금융업법 제2조에 의한 신용카드의 정의에 실물카드(플라스틱 카드) 없이 발급되는 모바일카드가 포함되는지 여부ㅇ 실물카드가 없는 경우에도 신용카드가맹점에서 반복하여 결제할 수 있는 경우에는 신용카드에 포함될 수 있습니다.ㅇ 여신전문금융업법상 신용카드의 정의에서 '증표'는 그 형태나 규격 등을 특정하고 있지 않으므로 플라스틱 실물카드만을 가리킨다고 볼 수는 없으며, - 모바일카드가 신용카드가맹점에서 이를 제시함으로써 반복적으로 결제가 가능한 증표로서 기능한다면 여신전문금융업법상 신용카드에 포함될 수 있습니다. * 여신전문금융업법 제2조 제3호 "신용카드"란 이를 제시함으로써 반복하여 신용카드가맹점에서 다음 각 목을 제외한 사항을 결제할 수 있는 증표(證票)로서 신용카드업자(외국에서 신용카드업에 상당하는 영업을 영위 하는 자를 포함한다)가 발행한 것을 말한다. (각 목 생략)
10	294	정보기술부문의 정보보호와 관련 업무를 위탁 받은 전자금융보조업자가 해당 업무를 제3자에게 재위탁할 수 있는 범위 ghltls정보기술부문의 정보보호와 관련 업무를 위탁 받은 전자
		종합금융투자사업자의 신용공여 규모 산정시 PF대출 포함 여부ㅇ 종합금융투자사업자의 신용공여 규모 산정시 기업금융업무 또는 만기 3개월 이내의 프로젝트파이낸싱 관련 대출금액 포함 여부ㅇ 종합금융투자사업자의 기업금융업무 또는 PF 관련 대출 등은 신용공여 규모 산정시 포함됨ㅇ 종합금융투자사업자의 기업 신용공여(PF 대출 포함) 업무는 법 제

FSC_TOTAL_FPRI_20170920_NOnDATA

(DATA ANALYSIS) 문서분류 – 키워드 추출

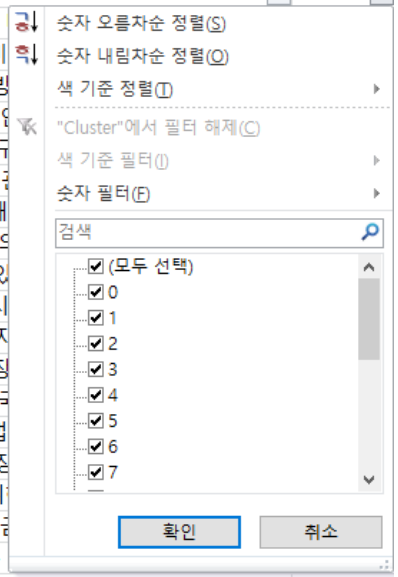
step one: extract keywords from Document
step two: keywords are used to build the word2vec model
step three: Document Clustering (K-Means, 20)

명/Noun', '거래/Noun', ')/Punctuation'], ['국적/Noun', '변경/Noun', '자/Suffix', '의/Josa', '금융/Noun', '거래/Noun', '가능/Noun', '여부/Noun', '관련/Noun', '질의/Noun', 'ㅇ/KoreanParticle', '금융/Noun', '실명/Noun', '거래/Noun', '및/Noun', '비밀/Noun', '보장/Noun', '에/Josa', '관한/Verb', '법률/Noun', '' /Foreign', '이/Noun', '거래/Noun', '의/Josa', '투명/Noun', '성/Suffix', '을/Josa', '확보/Noun', '하고/Josa', '경제/Noun', '정의/Noun', '를/Josa', '실현하는/Verb', '것/Noun', '을/Josa', '목적/Noun', '으로/Josa', '하는/Verb', '것/Noun', '을/Josa', '비추어/Verb', '볼/Noun', '때/Noun', ',/Punctuation', '이/Noun', '와/Josa', '같이/Josa', '동일인/Noun', '에/Josa', '대한/Noun', '형식/Noun', '적/Suffix', '인/Josa', '명의/Noun', '변경/Noun', '사항/Noun', '은/Josa', '제재/Noun', '대상/Noun', '이/Josa', '아닐/Adjective', '것/Noun', '이라고/Josa', '사료/Noun', '되/Verb', '나/Eomi', ',/Punctuation', '혹시/Noun', '라도/Josa', '이러한/Adjective', '해석/Noun', '과는/Josa', '달리/Noun', '금융기관/Noun', '등/Noun', '이/Josa', '어떠한/Adjective', '제재/Noun', '의/Josa', '대상/Noun', '이/Josa', '될/Verb', '수도/Noun', '있는/Adjective', '지/Eomi', 'ㅇ/KoreanParticle', '국적/Noun', '변경/Noun', '등/Noun', '으로/Josa', '거래/Noun', '자의/Noun', '실지/Noun', '명의/Noun', '가/Josa', '변경/Noun', '된/Verb', '경우/Noun', '금융/Noun', '회사/Noun', '는/Josa', '거래/Noun', '자/Suffix', '가/Josa', '제출한/Verb', '각종/Noun', '입증/Noun', '자료/Noun', '를/Josa', '통하여/Verb', '동일인/Noun', '임/Noun', '을/Josa', '확인하고/Verb', '금융/Noun', '거래/Noun', '를/Josa', '하여/Verb', '야/Eomi', '합/Verb', '니다/Eomi', './Punctuation', 'ㅇ/KoreanParticle', '금융/Noun', '실명/Noun', '거래/Noun', '및/Noun', '비밀/Noun', '보장/Noun', '에/Josa', '관한/Verb', '법률/Noun', '제/Noun', '3/Number', '조/Noun', '에/Josa', '따라/Verb', '금융/Noun', '회사/Noun', '는/Josa', '거래/Noun', '자의/Noun', '실지/Noun', '명의/Noun', '에/Josa', '의하여/Verb', '금융/Noun', '거래/Noun', '를/Josa', '하여/Verb', '야/Eomi', '합/Verb', '니다/Eomi', './Punctuation', '거래/Noun', '자의/Noun', '실지/Noun', '명의/Noun', '는/Josa', '주민등록/Noun', '표/Noun', '에/Josa', '기재/Noun', '된/Verb', '성명/Noun', '을/Josa', '의미합/Verb', '니다/Eomi', '(/Punctuation', '동법/Noun', '시행/Noun', '령/Noun', '제/Noun', '3/Number', '조/Noun', ')./Punctuation', 'ㅇ/KoreanParticle', 'A/Alpha', '씨/Noun', '는/Josa', '본인/Noun', '의/Josa', '실지/Noun', '명의/Noun', '로/Josa', '금융/Noun', '자산/Noun', '을/Josa', '보유/Noun', '해왔/Verb', '으며/Eomi', '단/Noun', '한/Verb', '번/Noun', '도/Josa', '타인/Noun', '명의/Noun', '로/Josa', '주식/Noun', '예금/Noun', '등/Noun', '금융/Noun', '자산/Noun', '을/Josa', '보유/Noun', '한/Josa', '바/Noun', '없/Adjective', '습니다/Eomi', './Punctuation', '다만/Noun', ',/Punctuation', '앞서/Noun', '본/Verb', '바/Noun', '와/Josa', '같이/Josa', '1999/Number', '년/Noun', '대한민국/Noun', '국적/Noun', '이/Josa', '자동/Noun', '으로/Josa', '상실/Noun', '됨/Verb', '에/Eomi', '따라/Verb', '신분/Noun', '이/Josa', '내/Determiner', '국/Noun', '인/Suffix', '에서/Josa', '외국인/Noun', '으로/Josa', '변경/Noun', '된/Verb', '바/Noun', '는/Josa', '있으/Adjective', '나/Eomi', ',/Punctuation', '위/Noun', '금융/Noun', '자산/Noun', '은/Josa', '계속/Noun', '해서/Josa', '타인/Noun', '의/Josa', '명의/Noun', '가/Josa', '아닌/Adjective', '모두/Noun', '동일인/Noun', '의/Josa', '명의/Noun', '로/Josa', '보유/Noun', ',/Punctuation', '유지되/Verb', '고/Eomi', '있으/Adjective', '므로/Eomi', '금융/Noun', '실명/Noun', '관련/Noun', '법령/Noun', '에/Josa', '위배/Noun', '되/Verb', '지는/Eomi', '않는/Verb', '것/Noun', '으로/Josa', '판단/Noun', '됩/Verb', '니다/Eomi', './Punctuation', 'ㅇ/KoreanParticle', '금융/Noun', '실명/Noun', '거래/Noun', '및/Noun', '비밀/Noun', '보장/Noun', '에/Josa', '관한/Verb', '법률/Noun', '제/Noun', '3/Number', '조/Noun', '(/Punctuation', '금융/Noun', '실명/Noun', '거래/Noun', ')/Punctuation'], ['보조금/Noun', '환수/Noun', '를/Josa', '위한/Verb', '금융/Noun', '거래/Noun', '정보/Noun', '제공/Noun', '가능/Noun', '여부/Noun', '관련/Noun', '질의/Noun', 'ㅇ/KoreanParticle', '보조금/Noun', '관리/Noun', '에/Josa', '관한/Verb', '법률/Noun', '제/Noun', '33/Number', '조의/Noun', '2/Number', '에/Josa', '의거/Noun', '금융/Noun', '정보/Noun', '를/Josa', '제공/Noun', '받/Verb', '을/Eomi', '수/Noun', '있는/Adjective', '지/Eomi', 'ㅇ/KoreanParticle', '보조금/Noun', '은/Josa', '국세/Noun', '및/Noun', '지방세/Noun', '에/Josa',

(DATA ANALYSIS) 문서분류 (K-Means)

step one: extract keywords from Document
step two: keywords are used to build the word2vec model
step three: Document Clustering (K-Means, 20)

	A	B	C
1	NO	DATA	Cluster
2	280	정보처리시스템 개발업무 사용장소(1) 외부주문등에 의한 정보처리시스템의 개발업무에 사용되는 업무장소와	
3	281	비밀번호 입력방식□ ACS(Auto Call System)를 통한 비밀번호 입력방식이 전자금융감독규정 제33조(이용자 비	
4	282	금융실명법상 본인확인 방법㈜ㅇㅇㅇ은 금융실명거래 및 비밀보장에 관한 법률 제3조에 의한 실지명의 확인 방	
5	283	단기매매차익 반환 기산일□ ㅇㅇ금융지주(상장법인)와 ㅇㅇ은행(비상장법인)이 합병(14.11.1.)한 후 존속법인	
6	284	청약철회 제한 규정 배제 가능여부1. 보험업법 제102조의4제1항 및 동법 시행령 제48조의2제1항제1호에도 불구	
7	285	역외금융회사등에 대한 해외직접투자에 있어 신고의무자역외금융회사 등에 대한 해외직접투자에 있어 금융기	
8	286	특정금전신탁을 통한 해외 파생결합증권 취득가능 여부ㅇ 자본시장법 시행령 제7조제4항제5의2호라목에서 '해	
9	292	신용카드의 정의에 실물카드 없는 모바일카드가 포함되는지 여부ㅇ 여신전문금융업법 제2조에 의한 신용카드의	
10	294	정보기술부문의 정보보호와 관련 업무를 위탁 받은 전자금융보조업자가 해당 업무를 제3자에게 재위탁할 수 있	
11	296	종합금융투자사업자의 신용공여 규모 산정시 PF대출 포함 여부ㅇ 종합금융투자사업자의 신용공여 규모 산정시	
12	297	종합금융투자사업자의 신용공여 규모 산정시 지급보증 포함 여부ㅇ 종합금융투자사업자가 특정 기업에 대한 지	
13	298	투자일임업자가 일임재산에 자신이 발행한 ELS를 편입하는 경우 투자자 동의를 받아야 하는지 여부ㅇ 자본시장	
14	299	일임계좌를 통한 해외시장 상장 ETF 편입 가능 여부ㅇ 자본시장법 제279조에 의해 외국 집합투자업자 등은 외국	
15	300	자신이 인수한 증권에 대한 신탁재산 및 일임재산의 매수 가능 범위ㅇ 투자일임업자 및 신탁업자는 자본시장법	
16	301	투자매매(중개)업을 겸영하는 투자일임(자문)업자가 투자자에 대한 대출업무 등을 영위할 수 있는지ㅇ 자본시장	
17	302	일임재산을 통한 환헤지 투자가 가능한지 여부ㅇ 일임재산으로 해외자산에 투자하는 경우, 환위험을 줄이기 위	
18	303	전자금융거래법 제40조 제6항의 전자금융보조업자의 정보보호업무 재위탁금지 예외 인정 관련 문의 회신전자금	
19	304	표준사업방법서 계약전 알릴의무 관련 유권해석 요구 표준사업방법서상의 계약 전 알릴의무 사항을 확대 또는	
20	305	표준사업방법서 계약전 알릴의무 관련 유권해석 요구에 대한 회신표준사업방법서상의 계약전 알릴의무 사항을 확대 또는 추가없이 그 취지의	7
21	312	소셜커머스의 보험상품 소개 페이지 게시 등□ 소셜커머스에서 보험상품 소개 페이지를 올리는 것이 모집' 행위인지? □ 소셜커머스의 보험	11
22	313	사회복지단체에 대한 무료보험 제공□ 보험회사가 사회공헌 차원에서 사회복지단체 등에 보험상품을 제공하고 보험료를 대납하는 경우 보험	5
23	314	바이럴 마케팅에 대한 수수료 지급□기존 보험계약자, SNS 이용자 등이 지인을 보험회사에 소개하거나 지인에게 보험상품을 추천하여 보험	11
24	315	보험계약 체결 과정중 전화녹취 가능 여부□ 금융위는 '12.11월 "보험업 허가정책"을 발표 ㅇ기존 종합보험사가 전문화.특화된 형태의 보험사	11
25	316	보험회사가 보험수익자의 연락처 수집□보험업법 시행령 §102⑤에 따라 보험회사는 보험수익자 지정·변경 등의 업무를 위해 보험수익자의 이	5
26	317	전자서명관리기준의 모바일 청약에 적용 여부□ 전자서명을 통한 보험계약 체결시 전자문서 작성 및 관리기준('12.1월, 이하 "전자서명관리기	7
27	318	통신판매전문보험회사의 금융기관보험대리점 홈페이지를 통한 모집□ 통신판매전문보험회사로 보험업 허가를 받은 보험회사가 은행 등 금융	11
28	319	보험업 허가정책 하에서 모자 보험회사간 금융기관보험대리점을 통해 대면 모집과 인터넷 모집을 분리하여 각각 위탁할 수 있는지?□ 금융위	5
29	321	자본시장법시행령 제272조1항각호 투자신탁재산 관련 자본시장법 시행령 제273조(보험회사에 대한 특칙)제1항 각 호의 "투자신탁재산"의 범	16
30	323	전자금융감독규정 제15조제1항제3호, 제5호의 단서 적용 회신1. 전자자금융감독규정 제15조 제1항 제3호, 제5호의 각 단서에서, 제3호 '업무상	2
31	324	보험업감독규정 제6-12조 항 4호 라목의 기존보험계약'의 범위□ 보험업감독규정 제6-12조 항 4호 라목의 기존보험계약'의 범위에 대한 해	2



(DATA ANALYSIS) 문서분류 (K-Means)

step one: extract keywords from Document
step two: keywords are used to build the word2vec model
step three: Document Clustering (K-Means, 20)

```
[STEP2]
2017-10-17 00:14:02,038 : INFO : collected 9769 word types from a corpus of 775967 raw words and 2238 sentences
2017-10-17 00:14:02,038 : INFO : Loading a fresh vocabulary
2017-10-17 00:14:02,066 : INFO : min_count=5 retains 4658 unique words (47% of original 9769, drops 5111)
2017-10-17 00:14:02,066 : INFO : min_count=5 leaves 766478 word corpus (98% of original 775967, drops 9489)
2017-10-17 00:14:02,078 : INFO : deleting the raw counts dictionary of 9769 items
2017-10-17 00:14:02,078 : INFO : sample=0.001 downsamples 72 most-common words
2017-10-17 00:14:02,079 : INFO : downsampling leaves estimated 594246 word corpus (77.5% of prior 766478)
2017-10-17 00:14:02,079 : INFO : estimated required memory for 4658 words and 100 dimensions: 6055400 bytes
2017-10-17 00:14:02,090 : INFO : resetting layer weights
2017-10-17 00:14:02,141 : INFO : training model with 4 workers on 4658 vocabulary and 100 features, using sg=0 hs=0 sample=0.001 negative=5
window=5
2017-10-17 00:14:03,145 : INFO : PROGRESS: at 68.11% examples, 2049520 words/s, in_qsize 7, out_qsize 0
2017-10-17 00:14:03,547 : INFO : worker thread finished; awaiting finish of 3 more threads
2017-10-17 00:14:03,549 : INFO : worker thread finished; awaiting finish of 2 more threads
2017-10-17 00:14:03,550 : INFO : worker thread finished; awaiting finish of 1 more threads
[STEP3]
2017-10-17 00:14:03,555 : INFO : worker thread finished; awaiting finish of 0 more threads
2017-10-17 00:14:03,555 : INFO : training on 3879835 raw words (2970389 effective words) took 1.4s, 2105504 effective words/s
```

```
[clusters]
[8, 7, 8, 7, 17, 7, 16, 9, 6, 2, 2, 16, 16, 18, 4, 16, 1, 7, 7, 11, 5, 11, 11, 5, 7, 11, 5, 16, 2, 2, 1, 0, 7, 9, 8, 16, 7, 7, 16, 16, 16, 5, 16,
7, 7, 7, 7, 16, 8, 2, 2, 6, 7, 5, 8, 7, 7, 9, 7, 8, 7, 4, 17, 16, 2, 2, 7, 2, 16, 16, 2, 0, 1, 0, 7, 8, 0, 4, 1, 7, 8, 12, 5, 4, 6, 16, 0, 17, 16,
11, 17, 18, 16, 12, 7, 19, 2, 0, 16, 17, 0, 10, 0, 6, 16, 4, 6, 16, 0, 7, 16, 0, 16, 16, 2, 4, 7, 5, 8, 7, 4, 2, 6, 5, 6, 5, 6, 5, 6, 8, 8, 7, 17,
4, 5, 7, 4, 19, 7, 7, 17, 11, 2, 6, 0, 4, 7, 6, 9, 2, 5, 2, 17, 0, 14, 17, 7, 7, 19, 19, 0, 9, 7, 2, 9, 7, 8, 8, 8, 2, 2, 7, 6, 4, 8, 17, 16, 15,
6, 5, 0, 6, 14, 4, 16, 7, 17, 16, 18, 6, 2, 7, 7, 0, 16, 7, 7, 16, 1, 2, 6, 7, 8, 16, 9, 8, 0, 18, 2, 6, 6, 7, 8, 0, 16, 8, 19, 2, 17, 16, 4, 0, 0,
16, 2, 9, 16, 16, 2, 7, 7, 8, 1, 14, 7, 6, 7, 6, 6, 7, 7, 16, 1, 4, 8, 1, 2, 5, 9, 2, 1, 8, 16, 17, 7, 16, 4, 4, 6, 2, 5, 5, 6, 17, 7, 8, 13, 18,
6, 17, 6, 7, 8, 17, 18, 9, 7, 9, 6, 5, 16, 5, 17, 7, 6, 18, 7, 16, 9, 8, 6, 0, 7, 7, 5, 9, 9, 16, 16, 16, 5, 7, 5, 8, 7, 7, 5, 16, 9, 5, 9, 2, 7,
7, 8, 16, 16, 0, 4, 6, 1, 7, 0, 7, 17, 4, 8, 8, 0, 6, 16, 15, 19, 16, 9, 6, 1, 9, 7, 11, 16, 7, 16, 1, 16, 7, 8, 18, 5, 2, 8, 7, 6, 7, 7, 18, 8, 7,
11, 9, 5, 0, 5, 18, 0, 6, 4, 2, 8, 2, 5, 16, 0, 8, 0, 0, 0, 16, 5, 11, 9, 16, 16, 7, 9, 11, 6, 5, 2, 8, 0, 7, 0, 8, 6, 1, 1, 7, 8, 16, 17, 2, 4, 7,
```

(DATA ANALYSIS) 문서 자동 분류 (K-Means) 결과

Category 3 : 대부업

121	대부업 등록취소 관련 질의ㅇ 대부업의 등록 및 금융이용자 보호에 관한 법률 제4조 및 제13조와 관련하여 등록취소"처분에 해당하게 된 경우	3
123	대부업체 광고시 조건게시 의무사항 관련 문의ㅇ 대부업 광고용 전화번호 3개 초과 등록이 법률 위배인지ㅇ 대부업법 시행령(제2조의3, 제2조	3
126	대부업 재직자에 대한 대부업법상 고용제한 규정 적용 관련 문의ㅇ 대부업체가 고용중인 자가 '10. 5.4.에 2심 법원에서 퇴거불응죄(형법 제31	3
140	대부업 등록 제한 관련 질의ㅇ 대부업 취소 및 등록제한 관련 질의ㅇ '09.4.21. 이후에 대부업등의 등록을 신청하는 자가 등록신청시를 기준으	3
143	대부업 등록제한 관련 문의ㅇ 과거 벌금형 선고로 인해 현재 대부업 법인이 등록취소가 되어야 하는데 회사의 지분을 정리하고, 대부업의 재	3
146	연이자율 제한 관련 질의ㅇ 2002년 법 적용 이전 대부업체로부터 고율의 이율(예 연 100% ~ 150 %)로 대부계약을 체결한 경우 대부업법의	3
155	정보제공사실 통지 의무 관련 질의ㅇ 법원의 제출명령 또는 법관이 발부한 영장, 세무서장의 제출요구에 따라 정보를 제공할 경우, 정보를 제	3
213	전자금융결제대행업 인허가 요건 관련 문의 ㅇ 전자금융거래법 제2조에서 정의한 "전자금융보조업자"에 PG업이 포함되는지 ㅇ 전자금융거래	3
349	대부업 등록 제한 사유 관련 질의 ㅇ 거주하고 있는 아파트를 고정사업장으로 하여 대부업을 등록하려고 할 때 위반되는 사항이나 등록이 제	3
371	대부업상 광고 관련 질의ㅇ 대부업법 제9조의3(허위,과장 광고의 금지 등) 1호 가항의 사실과 다르게 광고하거나 사실을 지나치게 부풀리게	3
389	대부계약 체결 시 징구서류 등 관련 질의ㅇ 대부계약 체결시 거래상대방이 법인인 경우 이와 관련하여 시행령 제4조의3 제1항 2호에 따라서	3
634	대부업 광고 등 관련 질의ㅇ 다수의 대부/대부중개업체 들이 자신의 사이트를 개설하여 인터넷 상에서 영업중인바, 이러한 대부/대부중개업	3
669	신용정보 조회서비스 이용 가능 여부 질의 ㅇ 한국도로공사가 신용정보 이용 및 보호에 관한 법률 제2조 및 동법 시행령 제2조의 정의에서 명	3
873	대부업광고 관련 질의ㅇ 대부업 등의 등록 및 금융이용자 보호에 관한 법률 시행령 [별표 3]에서는 법 제9조제2항 및 제3항을 위반하여 광고	3
884	대부업법 및 동법 시행령 입법예고 관련 문의ㅇ 법 2조에서 정한 대부의 범주 내에 (여신전문금융업법에 의하여 설립된)여신금융기관의 개인	3
886	여신전문금융업법 및 대부업법 관련 문의ㅇ 여신전문금융회사는 허가 또는 등록을 받아 리스나 할부금융 등의 여신전문금융업 뿐만 아니라	3

Category 6 : 보험사기

514	권리조사 위탁업무의 재위탁 발생에 대한 법령해석 □ 보험회사는 권리조사기관에 전입세대열람 등 권원보험 대상이 되는 부동산과 관련된	6
523	보험사기혐의가 농후한 건은 개인정보 제공에 대한 사전 동의 없이 ICPS 이력조회가 가능하도록 제도개선이 가능한지 여부□보험회사가 보험	6
531	보험모집업무 위탁 및 재위탁 / 손해사정업무 위탁 및 재위탁 / (약관,신용)대출업무 위탁 및 재위탁 가능여부□ 보험회사가 보험모집업무, 손	6
540	정보처리위탁규정 법령해석요청 회신 금융회사의 정보처리 업무 위탁에 관한 규정 (이하 정보처리 위탁규정') 제7조제1항제6호에 따른 "감독	6
555	보험회사에 대한 실명확인 업무 위탁□ 개정 금융실명거래 및 비밀보장에 관한 법률 (이하 금융실명법') 및 같은 법 시행령의 시행(14.11.29	6
567	보험사기와 관련된 형사고소 정보를 피보험자에게 제공 가능한지 여부 및 제공 가능시 제공범위에 대한 법률적 검토□ 보증보험사가 보험사	6
568	신용정보회사가 파견직 근로자를 통하여 채권추심업무 수행이 가능한지 여부 □ 신용정보회사가 파견직 근로자를 통하여 채권추심업무를 수	6
615	개인신용정보 이용 및 제공내역 조회 시스템 구축 범위□ 신용정보의 이용 및 보호에 관한 법률 (이하 신용정보법') 제35조('16.3.12. 시행)에	6
617	캐피탈과 증권사와의 캐피탈 대출 및 리스 이용 소개업무 위·수탁 계약□ 여신전문금융회사의 대출 및 리스이용 소개업무'는 금융지주회사법	6

(DATA ANALYSIS) 문서 자동 분류 (K-Means) 결과

Category 9 : 신용카드

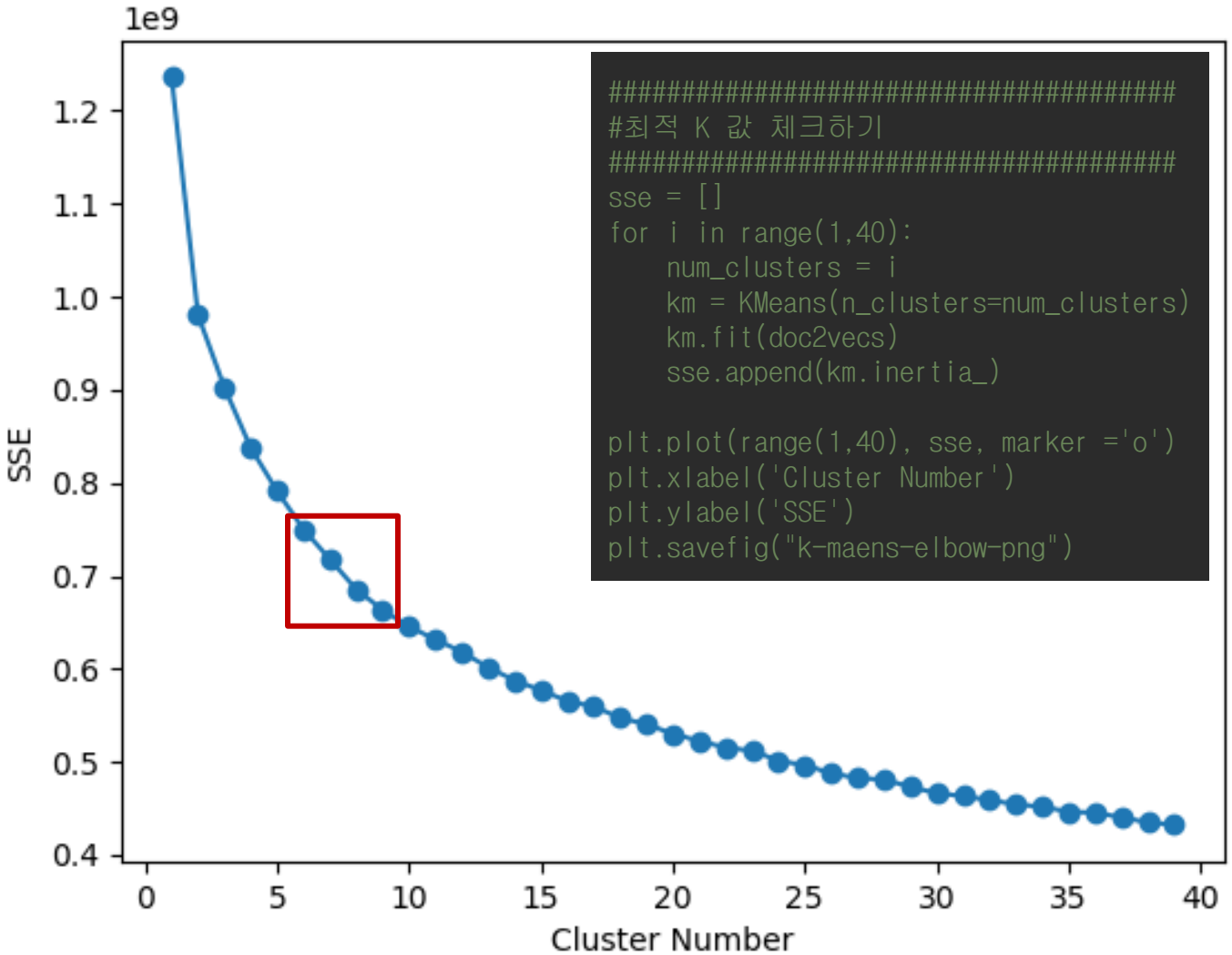
682	신용카드업자가 가맹점에 대해 신용카드 매출채권을 담보로 한 신용카드거래대금 단기 대출서비스를 시행하는 것이 여신전문금융업법상 허	9
692	발급 및 이용조건부 경제적이익 제공□ 인터넷(모바일)상에서 신규회원 대상 이벤트로 신규로 카드를 발급받고 일정금액 이상 이용시 연회비	9
693	지인추천에 대한 혜택 제공□ ㅇㅇ카드 회원이 지인에게 카드를 추천하여 지인이 카드를 발급받는 경우 ㅇㅇ카드가 카드회원에게 경제적 혜	9
707	부가통신업자의 기부행위에 대한 부당한 보상금 해당 여부□ 부가통신업자가 대형가맹점과의 계약조건에 근거하여 대형가맹점의 사회공헌	9
709	신용카드 결제대금 채권의 팩토링이 은행법상 부수업무로서 허용된 팩토링' 업무에 해당하는지 여부□ 은행법 제27조의2 제2항제3호에 따	9
741	신용카드등부가통신업 해당 여부와 인증된 VAN사로부터 복호화 처리된 IC정보를 카드사로 전달하여 서비스를 수행하는데 문제가 없는지의	9
745	신용카드 회원 모집이 가능한 자의 해석범위□ 모집제휴계약을 체결함으로써 제휴 법인의 임직원이 신용카드업자를 위하여 특정 상품에 국한	9

Category 19 : 유사수신

194	유사수신행위 해당 여부 질의ㅇ 상품 자체가 금융법이나 기타 관련된 법에 위반되는 내용인지ㅇ 귀하가 제시한 상품기획 개요를 근거로 판단	19
196	유사수신행위 해당 여부 질의ㅇ 충전금이 일정금액을 초과시 매월 일정 비율의 금액을 사용자에게 지급ㅇ 충전금이 일정금액을 초과시 매월	19
197	유사수신행위 해당 여부 질의ㅇ 일반법인 사업자가 창업투자자를 대리하여 경영전반에 대하여 책임을 지고 계약에 따라 수익금을 투자자와	19
198	유사수신행위 해당 여부 질의ㅇ 상조회사 특약상품의 유사수신행위 여부ㅇ 사업내용이 다음 세가지 요건을 모두 충족하는 경우가 아니라면	19
200	유사수신행위 해당여부 질의ㅇ P2P 금융의 유사수신행위 해당여부ㅇ 사업 내용이 관계 법령에 따른 인 허가 등을 받지 않고 불특정 다수로부	19
201	유사수신행위 해당여부 질의ㅇ 민원인의 사업계획이 유사수신행위에 해당하는지 여부ㅇ 귀하의 사업 내용이 관계 법령에 따른 인 허가 등을	19
373	금전의 대부 등 관련 질의ㅇ 대부업법 제2조 제1호에서 '금전의 대부' 및 '금전의 교부'라고 규정하고 있는데 이 때 '금전'이 유가증권 또는 전	19
470	유사수신행위 해당 여부 질의ㅇ 법인대부업자의 경우 법인 차입(차입자에게 차입증서를 발행해주고 원금에 대해 일정비율의 이자를 매월 지	19
471	유사수신행위 해당 여부 질의ㅇ 원금에 이자 지급을 약정하고 회사 직원을 대상으로 자금을 차입하는 행위가 금융관련법령에 위반되는지 여	19
472	유사수신행위 해당 여부 질의ㅇ 민원인의 사업내용이 유사수신행위에 해당하는지 여부ㅇ 귀하께서 적시하신 내용이 "유사수신행위의규제에	19
473	유사수신행위 해당 여부 질의ㅇ 사안과 같이 미래에 수익을 지급할 것으로 약정한 사업이 유사수신행위에 해당하는지ㅇ 불특정 다수인으로	19
571	영화투자에 관한 유사수신행위 위법 여부 관련 질의ㅇ 수신행위를 업태로 하고 있지 않은 영화제작사나 기타 인터넷 사이트에서 이러한 수신	19

Yeah!

(DATA ANALYSIS) 문서 자동 분류(K-Means) 결과 – 최적 K값 판별

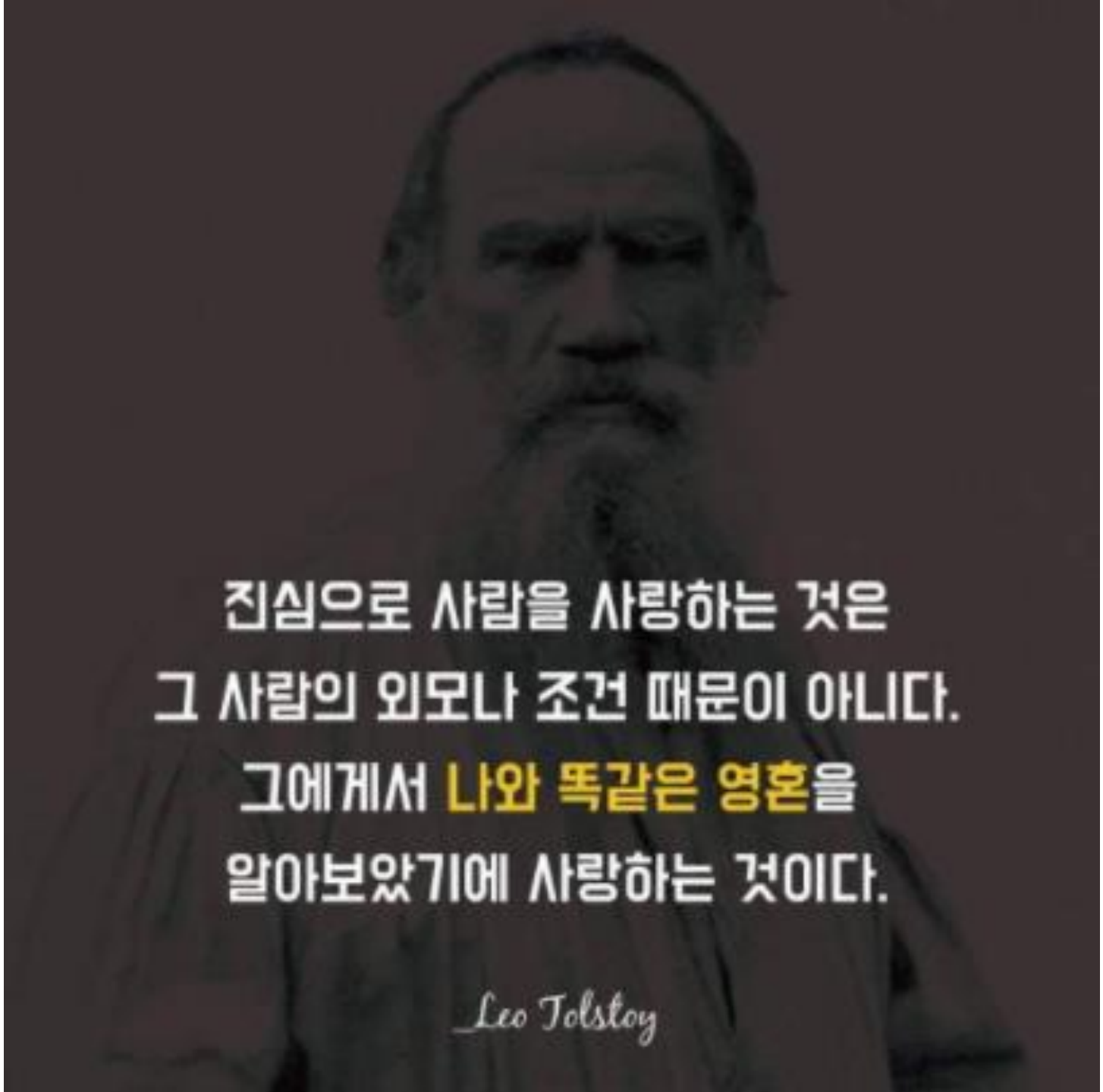



```
#######################################
#최적 K 값 체크하기
#######################################
sse = []
for i in range(1,40):
    num_clusters = i
    km = KMeans(n_clusters=num_clusters)
    km.fit(doc2vecs)
    sse.append(km.inertia_)

plt.plot(range(1,40), sse, marker ='o')
plt.xlabel('Cluster Number')
plt.ylabel('SSE')
plt.savefig("k-maens-elbow-png")
```

Yeah!

Yeah!

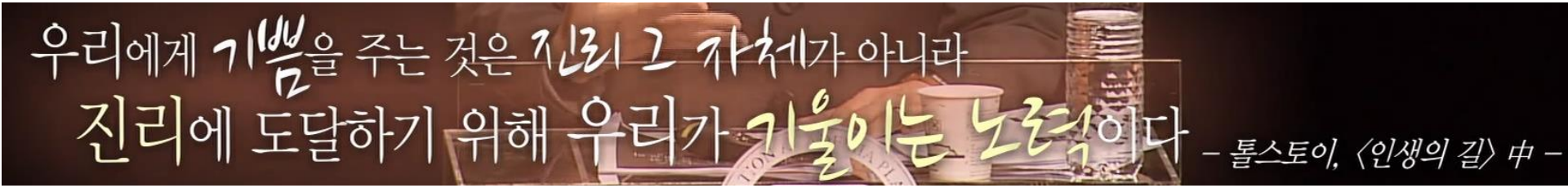

진심으로 사람을 사랑하는 것은
그 사람의 외모나 조건 때문이 아니다.
그에게서 나와 똑같은 영혼을
알아보았기에 사랑하는 것이다.
Leo Tolstoy

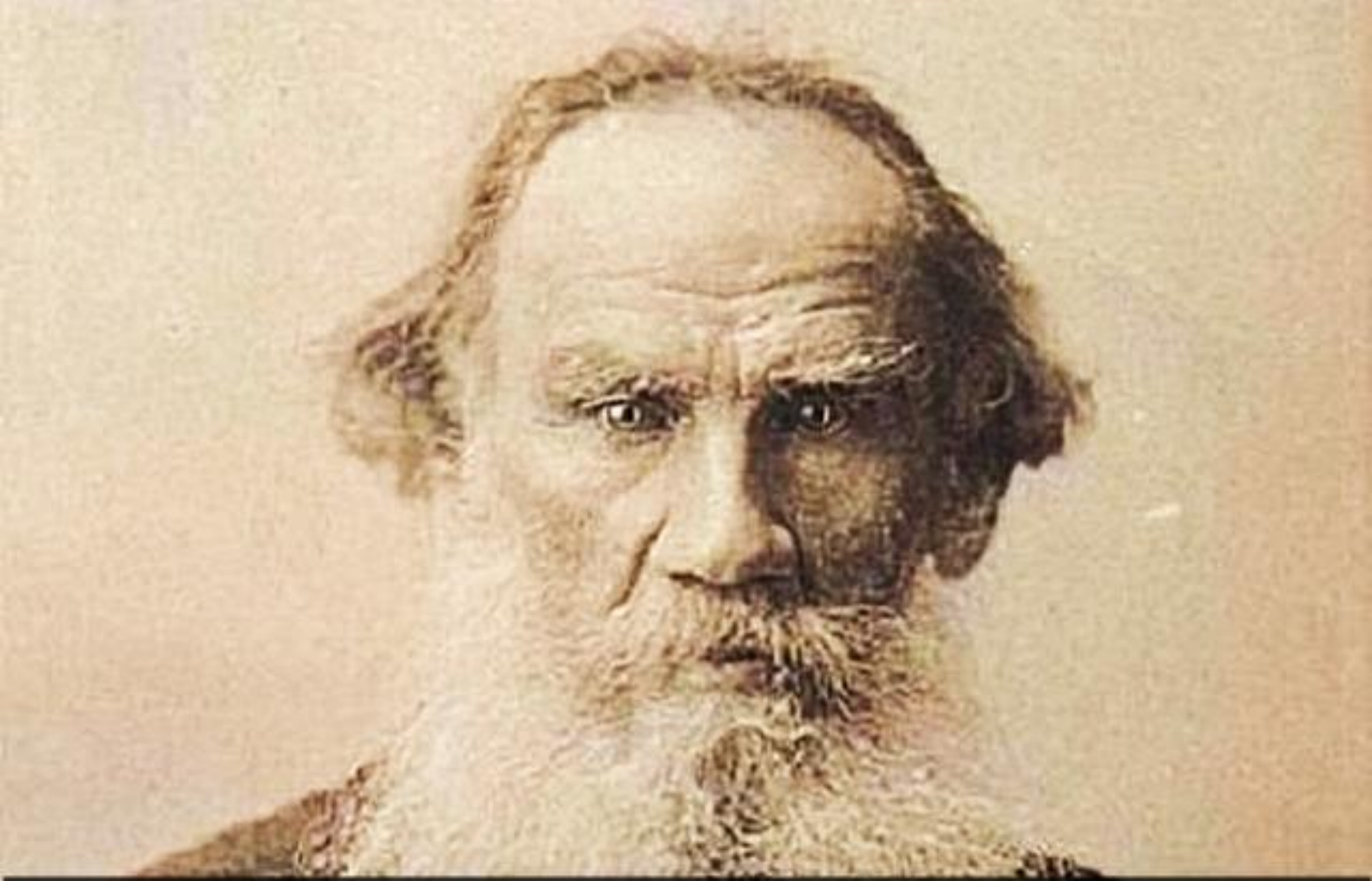
“만약 당신이 미래를 꿈꾸지 않거나 지금 기술개선을 위해 노력하지 않는다면 그건 곧 낙오되고 있는 것이나 마찬가지 입니다.”

그윈 쇼트웰(Gwynne Shtwell, SpaceX CEO, COO)

감사합니다

(facebook.com/sangshik, mikado22001@yahoo.co.kr)


우리에게 기쁨을 주는 것은 진리 그 자체가 아니라
진리에 도달하기 위해 우리가 기울이는 노력이다
– 톨스토이, 〈인생의 길〉 中 –



노력은 수단이 아니라 그 자체가 목적이다. 노력하는 것 자체에 보람을 느낀다면
누구든지 인생의 마지막 시점에서 미소를 지을 수 있을 것이다. - 톨스토이