NLP와 빅데이터, 그리고 머신러닝

2017년 2월 20일

강 승 식
국민대학교 컴퓨터공학부

Topics

- NLP, BigData, and Machine Learning
  - 자연어 처리
  - 빅 데이터 분석
  - 머신러닝 기법

Natural Language Processing

NLP issues and applications

NLP Basics

- Morphological analysis(형태소 분석)
  - Word-level
- Syntactic analysis(구문 분석)
  - Sentence-level
- Semantic analysis(의미 분석)
  - Word-sense disambiguation
- Natural Language Generation(자연어 생성)
- Language Resources(언어 자원)
  - 말뭉치, WordNet, 온톨로지 등

NLP Applications

- Machine Translation, 1950's-now
- Information Retrieval, 1980's-now
  - Text Classification, Information Extraction
  - Text Summarization
  - Text Mining, Opinion Mining
  - Sentiment Classification(감성 분류)
- Natural Language Understanding, 1960-70, 2000's
  - ELIZA: Doctor, Joseph Weizenbaum, MIT, 1965
  - SHRDLU: Robot arm, Terry Winograd, MIT, 1971
  - LUNAR
  - Ask Jeeves(ask.com), 1996
  - Wolfram alpha, 2009

- Speller and grammar checker

- Spam mail filtering, Spam 문자 filtering
- Sentiment analysis(감성 분석)

- 아이폰 시리, IBM 왓슨, 자동통역 시스템
- 텍스트 마이닝, 빅데이터 분석

빅 데이터 분석

빅데이터 소개

- https://www.youtube.com/watch?v=X4hMFym0-uo

- 향후 10년 내 세상을 바꿀 가장 중요한 기술: 네이처(2008)
- 미국 경쟁력을 좌우하는 21세기 원유: 가트너(2011)

- 구글의 독감 트렌드
  - '독감증세' 보이는 사람이 늘면 '독감증세' 단어의 검색량도 증가
- 지난 2년 동안 생산된 정보는 인류 탄생 이후 생산된 정보량보다 많다: IBM(2011)
- 서울시 심야버스, 신생아 질병 감염 등

빅데이터 분석 기술

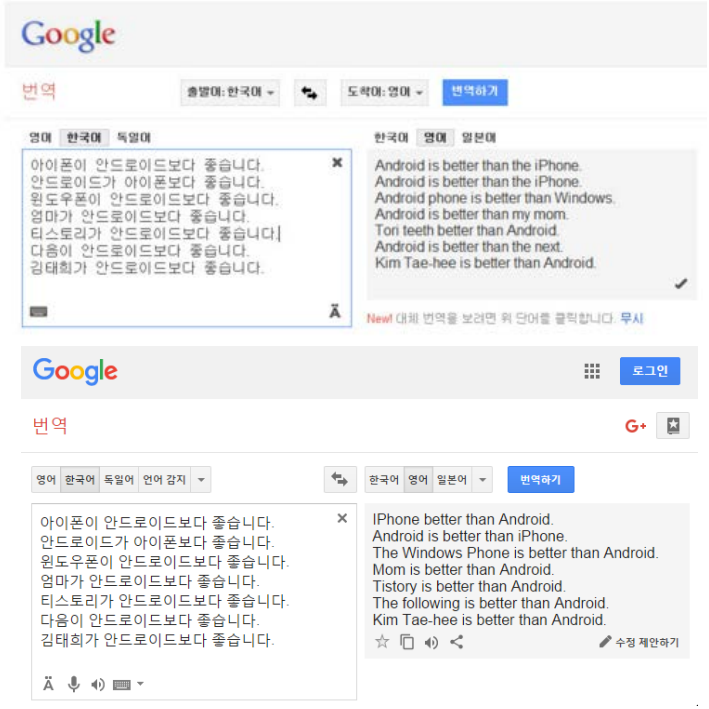
- 데이터 마이닝, 기계 학습, 자연언어 처리, 패턴 인식
- 텍스트 마이닝
  - 비/반정형 텍스트 데이터에서 자연어 처리 기술에 기반하여 유용한 정보를 추출, 가공
- 오피니언 마이닝
  - 소셜미디어 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별
- 소셜 네트워크 분석(SNS)
  - 소셜 네트워크의 연결 구조 및 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정
- 군집 분석(clustering)
  - 비슷한 특성을 가진 개체를 merge하면서 최종적으로 유사 특성의 군집을 발굴

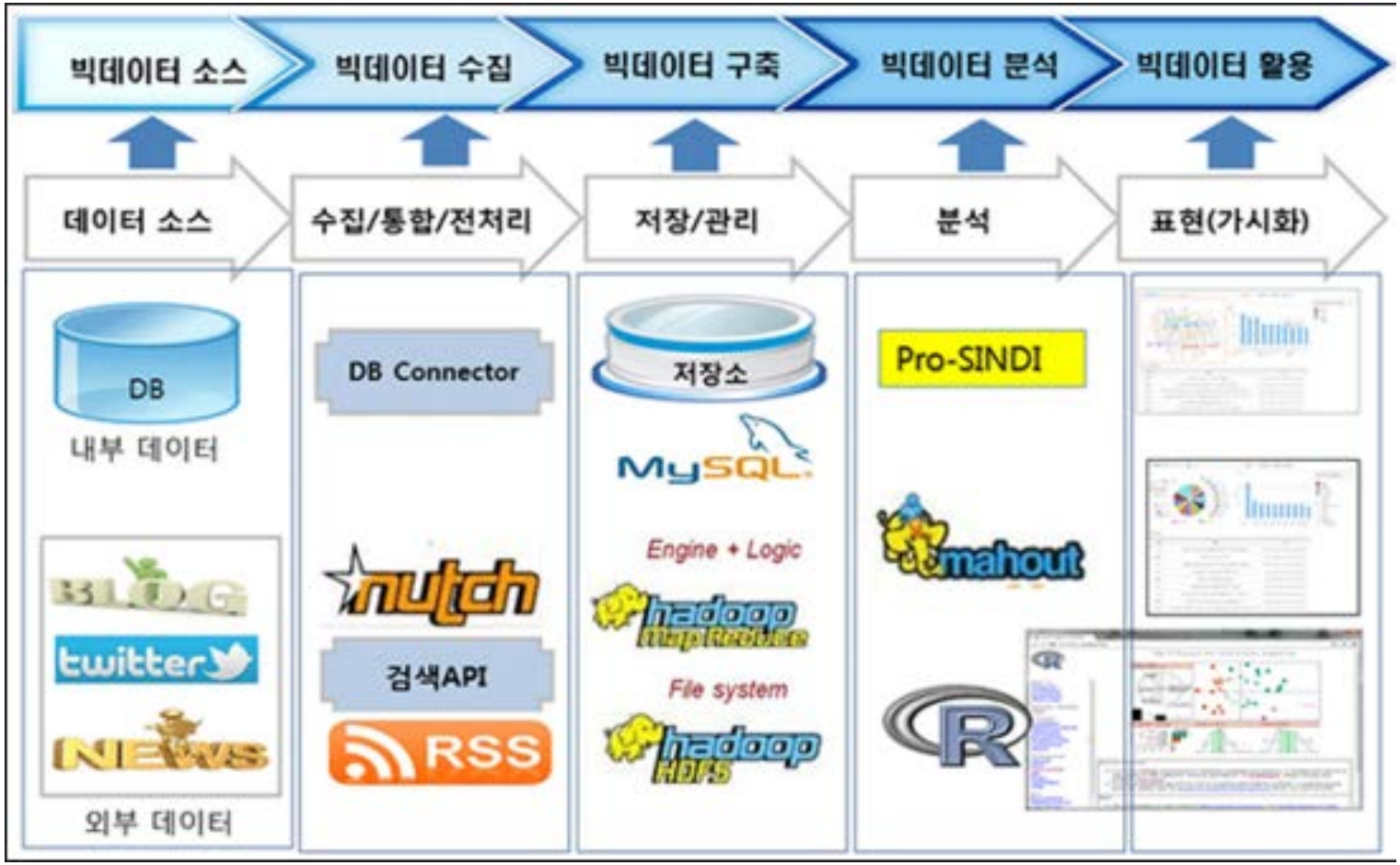
빅데이터 활용 예: 구글 번역

- 기존의 기계번역 방식
  - 변환(transfer) 방식과 피봇(pivot) 방식의 자동 번역 기법
  - 컴퓨터가 명사, 형용사, 동사 등 단어와 어문의 문법적 구조를 인식하여 번역하는 방식

- 구글이 제공하는 자동 번역 서비스인 구글 번역의 특징
  - 통계적 방식: 빅데이터를 활용하는 방법으로 구현
  - 수억 건의 문장과 번역문을 데이터베이스화
  - 번역시 유사한 문장과 어구를 기존에 축적된 데이터를 바탕으로 추론
  - 구글은 수억 건의 자료를 활용하여 전 세계 58개 언어 간의 자동번역 프로그램 개발에 성공

- 데이터 양의 측면에서의 엄청난 차이가 자동 번역 프로그램의 번역의 질과 정확도에 영향을 미침



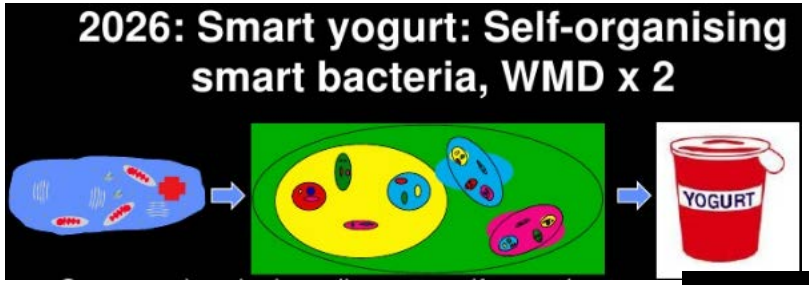

빅데이터 소스
빅데이터 수집
빅데이터 구축
빅데이터 분석
빅데이터 활용
데이터 소스
수집/통합/전처리
저장/관리
분석
표현(가시화)
DB
내부 데이터
BLOG
twitter
NEWS
외부 데이터
DB Connector
nutch
검색API
RSS
저장소
MySQL
Engine + Logic
hadoop
Map Reduce
File system
hadoop
HDFS
Pro-SINDI
mahout
R

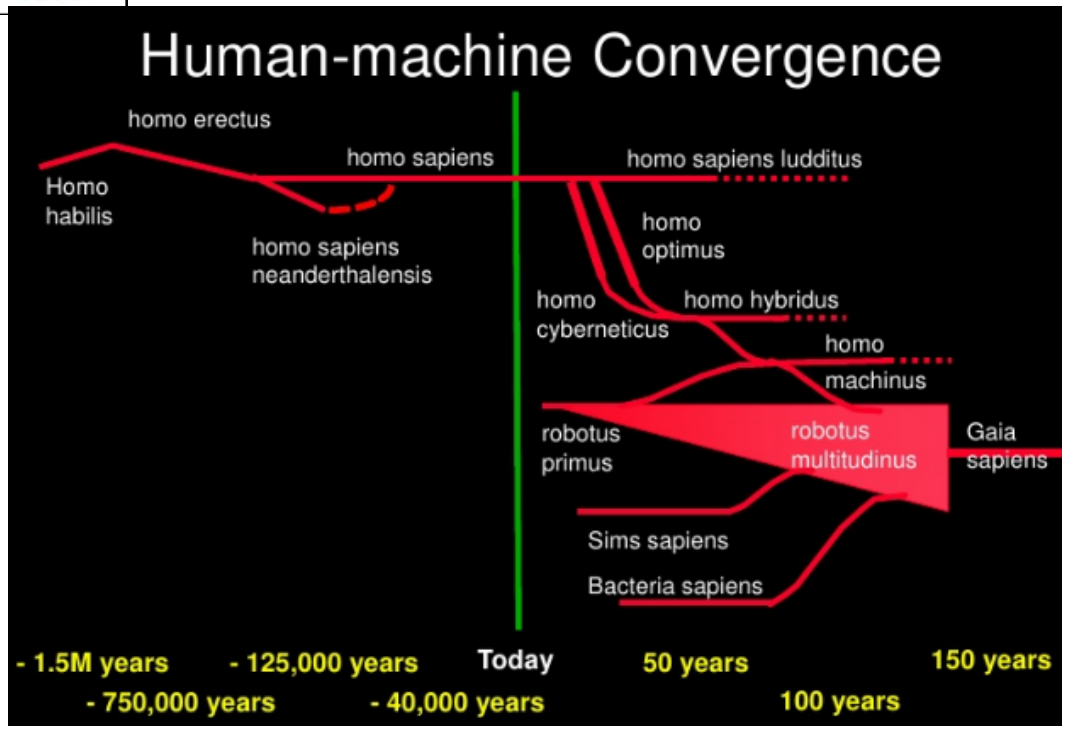
빅데이터 시각화: 남녀간 어휘 분석

- 여자
  - shopping, chocolate, hair, happy, boyfriend
  - 이모티콘
    - ^.^ :)
- 남자
  - youtube, fuck, xbox, league, world_cup, football

미래의 기술은…

미래 기술: Ian Pearson


2026: Smart yogurt: Self-organising smart bacteria, WMD x 2
YOGURT


Human-machine Convergence
homo erectus
homo sapiens
homo sapiens ludditus
Homo habilis
homo sapiens neanderthalensis
homo optimus
homo cyberneticus
homo hybridus
homo machinus
robotus primus
robotus multitudinus
Gaia sapiens
Sims sapiens
Bacteria sapiens
- 1.5M years
- 750,000 years
- 125,000 years
- 40,000 years
Today
50 years
100 years
150 years

IT 기술 발전의 3 단계

1. 모든 정보가 한 개의 칩에 내장되는 단계
2. 별도의 학습 없이도 IT 기술을 이용할 수 있는 '단순성'의 단계
3. 3차원 가상 세계의 순으로 발전

- 인간의 지능을 가진 컴퓨터(2020년)

- 생각하는 컴퓨터 탄생
  - "뇌가 외부의 정보를 받아들이는 방법에 관한 연구"
  - "지각 또는 인식 능력을 감각의 하나로 인식해 컴퓨터에 설계해 넣으려는 시도"

- 컴퓨터가 감정을 느끼는 단계로 발전
  - 항공기가 승객보다 더 추락 사고를 무서워하게 설계돼 추락하지 않기 위해 안간힘을 쓰게 될 것이다.

- 2050년 인간의 뇌에 있는 의식을 슈퍼컴퓨터에 저장

NLP with Emotion

- 친구(사람)보다 컴퓨터(기계)와 보내는 시간이 더 많다.
  - 미래의 컴퓨터는 사용자의 감정 상태까지 감지?
  - 감정 기계 or 감정 computing
- 감정의 종류
  - 기쁨/슬픔, 사랑/미움, 좋음/싫음, 놀라움 등
- 생각하는 컴퓨터, 감정 인식, 추론
  - Intelligent messenger(챗봇)

기술 발전과 진보: 4차 산업혁명?

- Siri on iPhone 4S

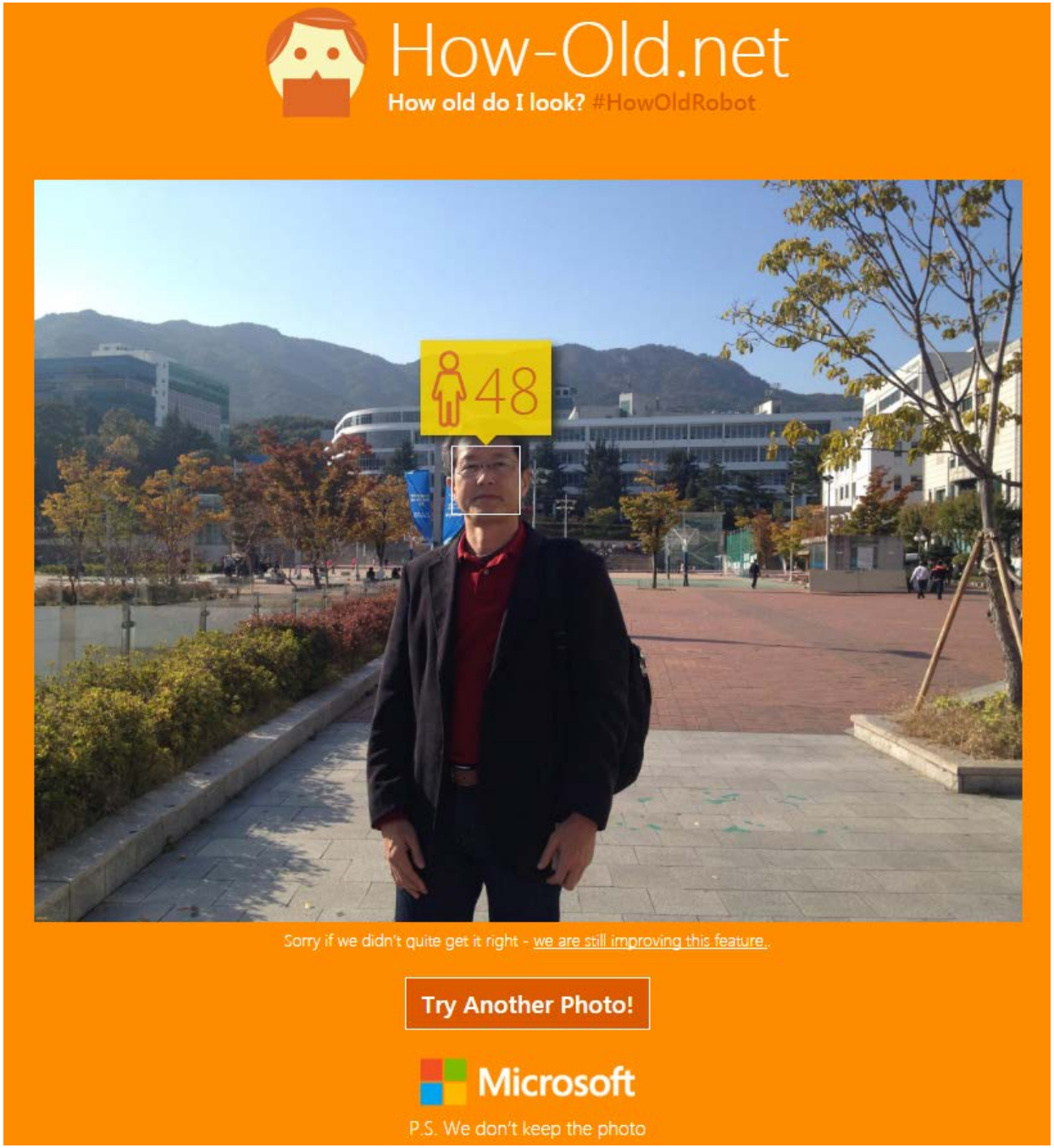
- IBM 딥 블루, 왓슨

Google DeepMind: AlphaGo

- http://www.deepmind.com/alpha-go.html
- 개발자, 아마추어 수준(1년 경력)
  - 16만건의 대국 데이터
- 2015년 10월, Fan Hui 2단(유럽 바둑 챔피언)
- 2016년 3월9-15일, 이세돌 9단(세계랭킹 5위)

알파고의 승리가 논문에 실린 〈네이처〉

AlphaGo


How-Old.net
How old do I look? #HowOldRobot
48
Sorry if we didn't quite get it right - we are still improving this feature.
Try Another Photo!
Microsoft
P.S. We don't keep the photo

http://captionbot.ai

I think it's a large crowd of people at night and they seem 😐😐.

I think it's a bunch of people at night.

Deep Learning

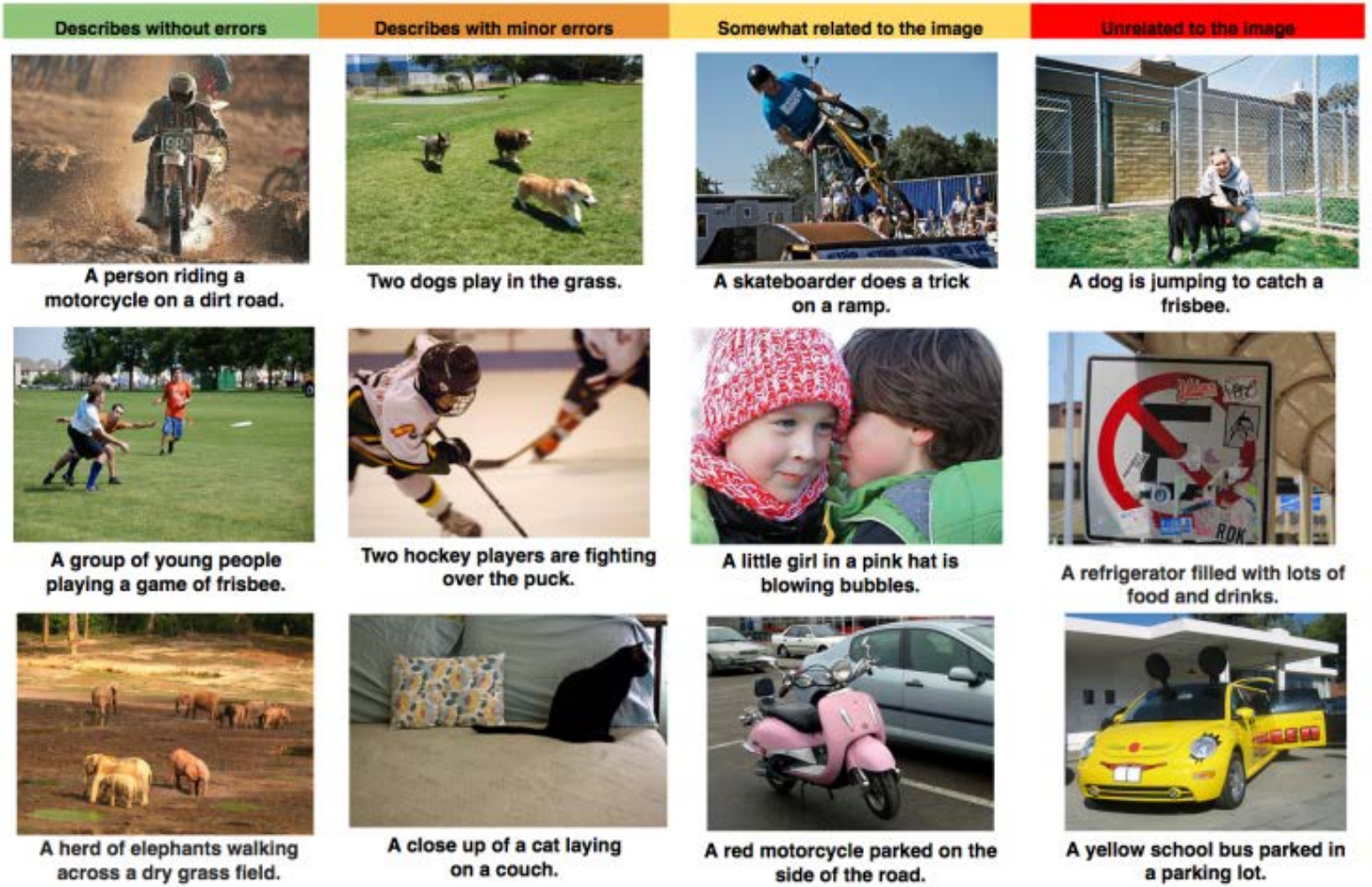
Deep Learning

- 1980년대, Neural Network

- 2006년, Geoffrey E. Hinton(Univ. of Toronto)
  - Generalized backpropagation algorithm for training multi-layer neural nets → deep learning
  - Reducing the dimensionality of data with neural networks.
  - Learning Multiple Layers of Representation
  - Where do features come from?

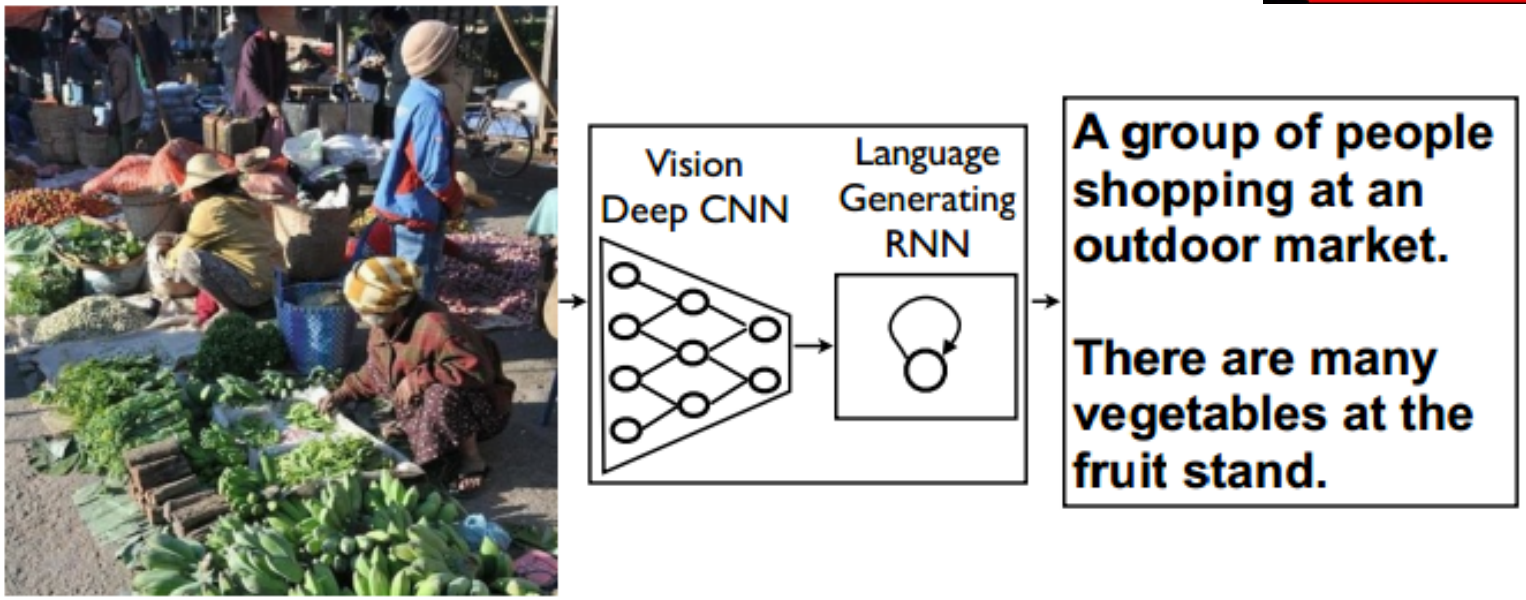
- Motivation: image classification in the Web
  - Automatic feature extraction from images

Google: Auto-Caption Complex Images



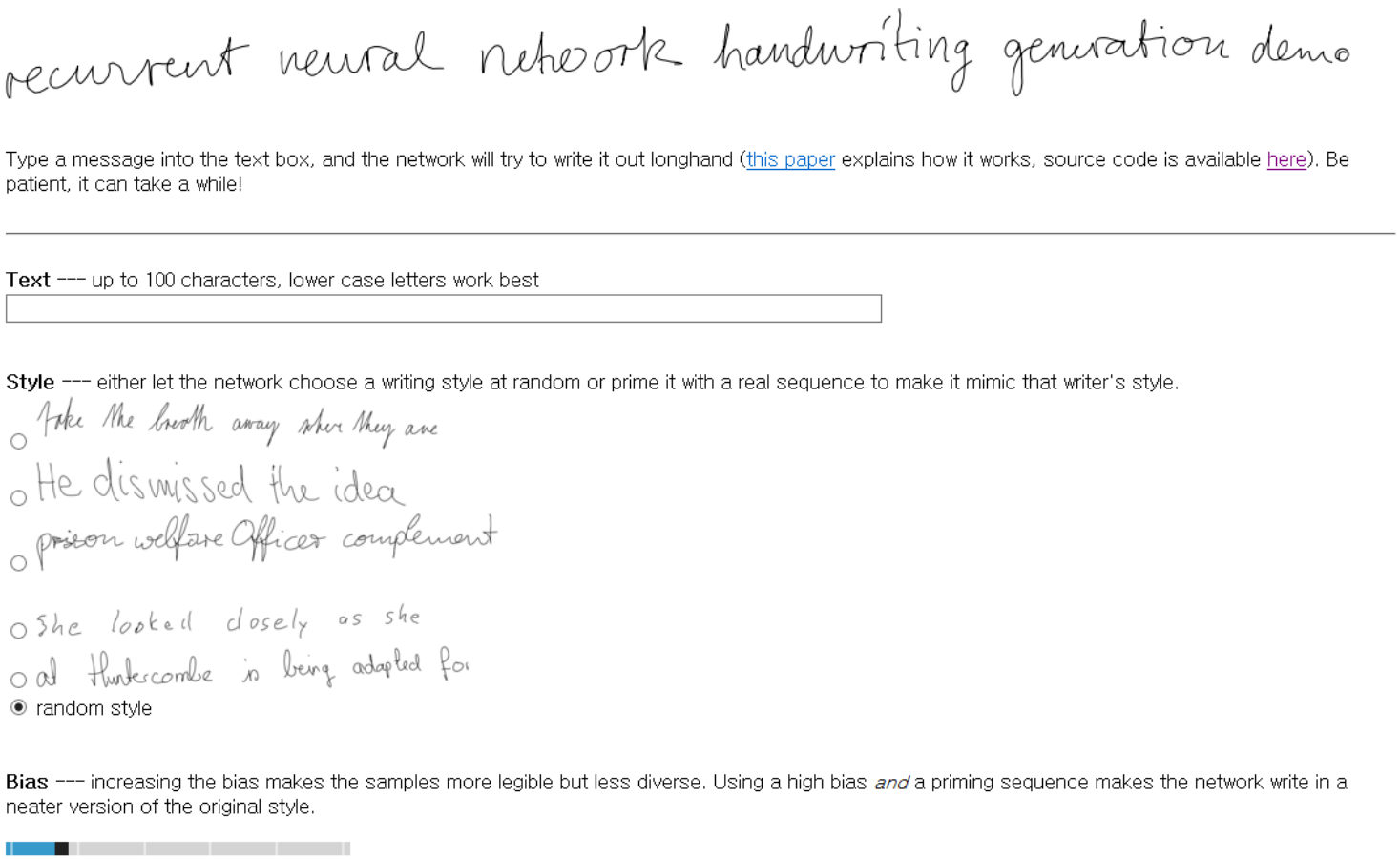
http://techcrunch.com/2014/11/18/new-google-research-project-can-auto-caption-complex-images/

Deep Learning Methodology



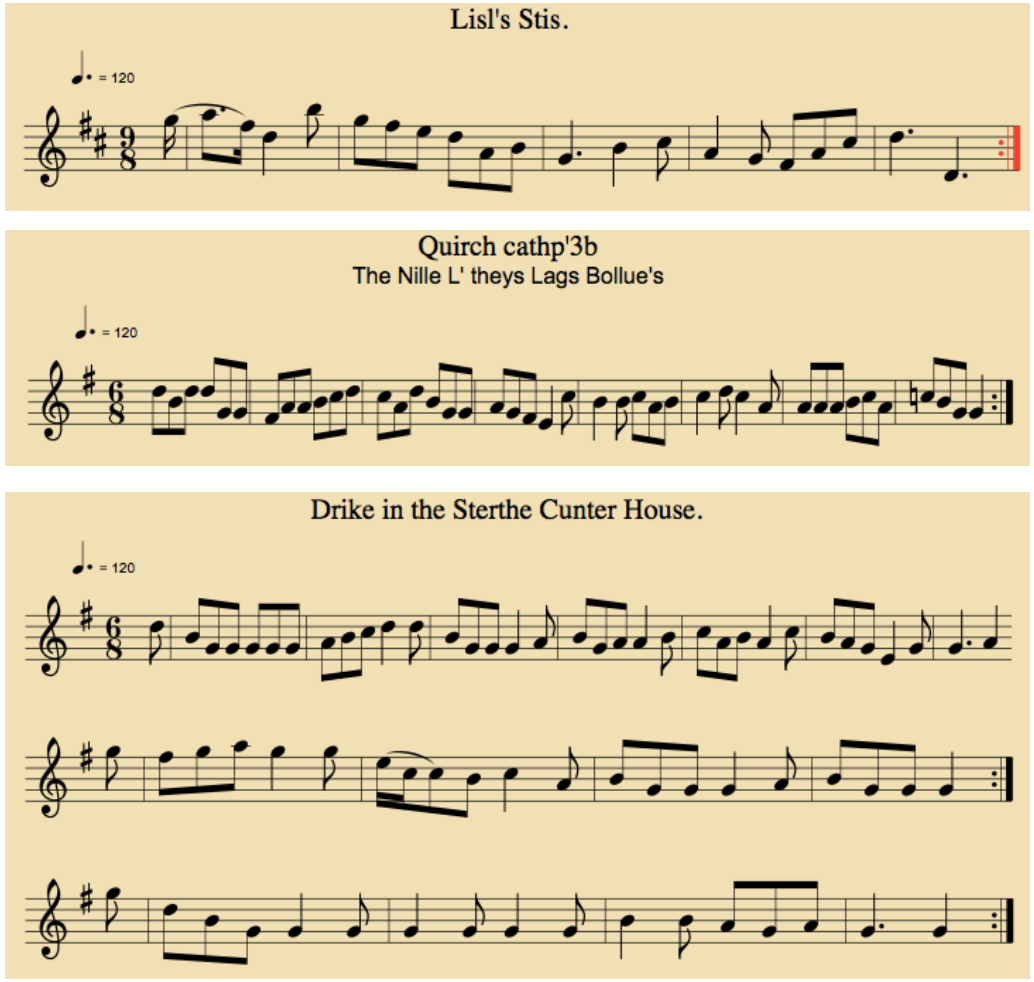
http://techcrunch.com/2014/11/18/new-google-research-project-can-auto-caption-complex-images/

Handwriting Generation by RNN



http://www.cs.toronto.edu/~graves/handwriting.html

Folk Music Generation by RNN

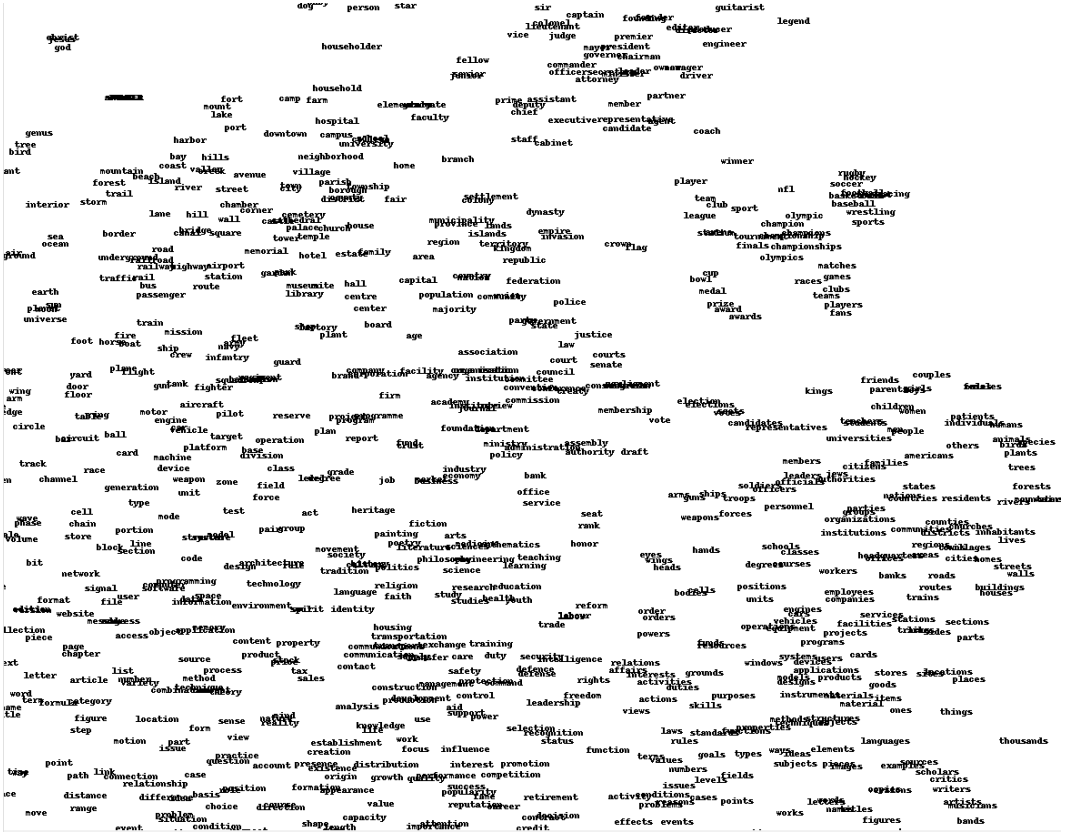


https://highnoongmt.wordpress.com/2015/05/22/lisls-stis-recurrent-neural-networks-for-folk-music-generation/

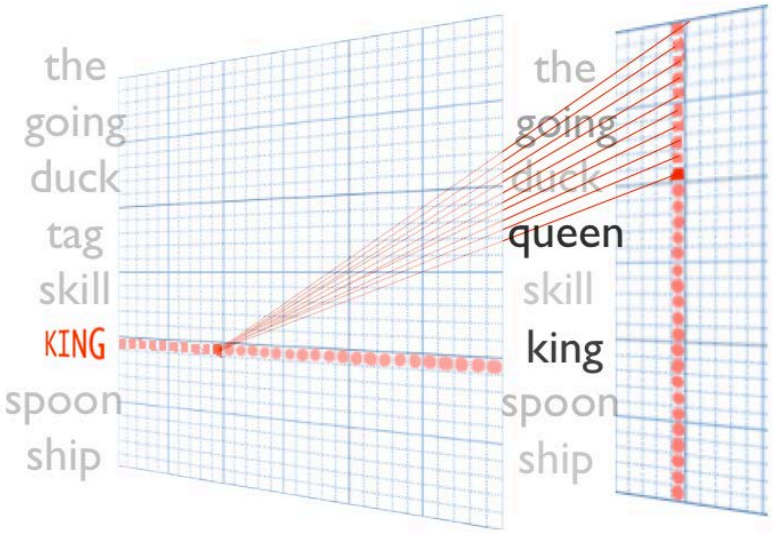
Deep Learning and NLP

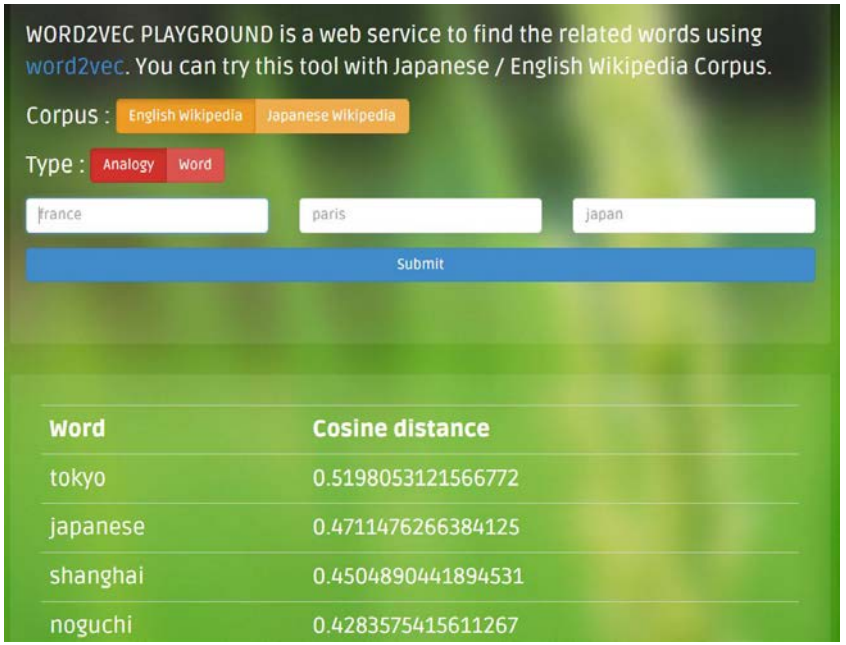
Word2Vec

- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013)
  - “Efficient Estimation of Word Representations in Vector Space”



Word Analogy





France - Paris + Seoul = Korea

King - Boy + Girl = Queen

Neural Machine Translation

- Demo -- http://104.131.78.120/

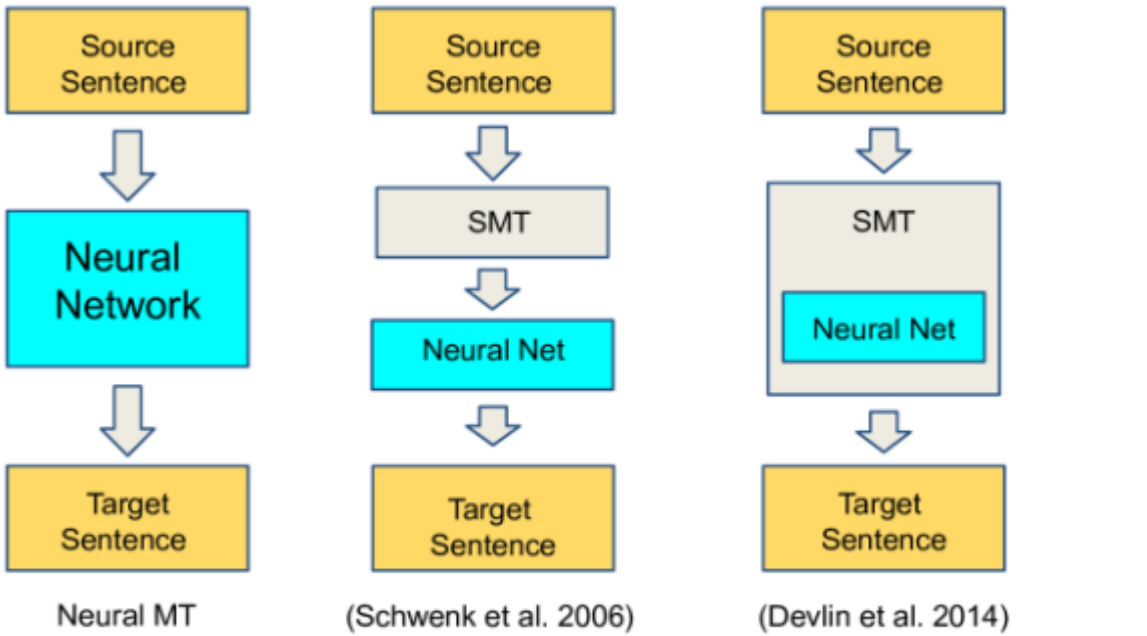


Figure 2. Graphical illustration of Neural MT, SMT+Reranking-by-NN and SMT-NN. From [Bahadanau et al., 2015] slides at ICLR 2015.

https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-with-gpus/

Sentence Completion Challenge

- http://research.microsoft.com/en-us/projects/scc/

- Fill-in-the-blank questions like in SAT

1. One of the characters in Milton Murayama's novel is considered _____ because he deliberately defies an oppressive hierarchical society.
(A) rebellious (B) impulsive (C) artistic (D) industrious (E) tyrannical

2. Whether substances are medicines or poisons often depends on dosage, for substances that are _____ in small doses can be _____ in large.
(A) useless .. effective
(B) mild .. benign
(C) curative .. toxic
(D) harmful .. fatal
(E) beneficial .. miraculous

Deep Learning: Developing Tools

- Theano
  - http://deeplearning.net/software/theano/
- Caffe
  - http://caffe.berkeleyvision.org/
- TensorFlow – Google
  - https://www.tensorflow.org/
- Keras – deep learning library for Theano and Tensorflow
  - http://keras.io/
- Microsoft Azure -- Machine Learning Studio
  - https://studio.azureml.net/

AI, Machine Learning, Deep Learning

AI > machine learning > deep learning

- AI
  - 지식표현, game theory
  - NLP, Q&A, M.T., pattern recognition, expert system, etc
- Machine learning
  - Decision tree, neural net, SVM, Naïve Bayes, ada boost
- Deep learning (deep neural network)
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Network (RNN)
  - Restricted Boltzmann Machine (RBM)

중요한 것은… 인공지능 기술의 실용화

- Apple의 Siri

- IBM의 인공지능 Watson
  - 2011년 2월

- 구글의 Deep Learning 기술
  - 2016년 3월

IBM Watson

- AI phobia
  - http://blog.naver.com/khko7/220780573799
  - 4년 뒤 710만개의 일자리가 사라지고 200만개 생겨남
    - 세계경제포럼, 직업의 미래

- AI 의사: 암 진단 연구, 의사보다 더 정확하다?
- AI 변호사
  - 교통사고 분야

AI applications

- Current trends in AI
  - Feature extraction, classification, decision making
  - Machine learning

- NLP application
  - Apple Siri
  - IBM Watson
  - Google word2vec

- Non-NLP application
  - Alpha Go
  - Caption bot

AI implementing techniques

- Machine Learning
  - Deep learning
  - SVM, CRF, Naïve Bayse
- NLP: analysis and generation
  - POS tagging, named-entity recognition, parsing
  - Semantic analysis, sentiment analysis
  - Natural language generation

Robot Journalism

아래 기사는 누가 썼을까? 인간기자 vs. 로봇

기사 1.

린드블럼이 선발로 등판한 롯데는 박종훈이 나선 SK에게 3:5로 패하며 안방에서 승리를 내주었다. 경기의 승패에 결정적인 영향을 미친 키 플레이어는 브라운이었다. 브라운은 5회초 롯데 린드블럼을 상대로 3점을 뽑아내어 팀의 승리에 결정적으로 기여했다. 롯데는 윤길현을 끝까지 공략하지 못하며 안방에서 SK에 2점차 승리를 내주었다.

기사 2.

SK가 이틀 연속 롯데를 제압했다. SK 와이번스는 6일 부산 사직구장에서 열린 2015 타이어뱅크 KBO리그 롯데 자이언츠와의 경기에서 선발 박종훈의 호투와 앤드류 브라운, 정상호의 홈런에 힘입어 5-3으로 승리했다. 이날 승리로 SK는 2연승을 거두며 시즌 성적 16승 12패를 기록했다. 한화, 넥센이 패하며 5위에서 3위로 도약했다. 반면 롯데는 홈에서 이틀 연속 패배, 시즌 성적 15승 15패가 됐다.

기사 3,

6일 잠실 야구장에서 열린 2015 KBO리그 LG 트윈스와 두산 베어스의 경기에서 LG가 두산에 4-5로 패했다. 이날 니퍼트는 선발로 등판해 6⅔이닝 6피안타 6탈삼진 1볼넷 2실점(1자책)했다. 시즌 3번째 퀄리티 타트(QS)를 달성한 니퍼트는 2승을 달성했다. 한편 LG는 7연패 수렁에 빠졌다.

기사 4.

두산은 6일 열린 홈 경기에서 LG를 5:4, 1점차로 간신히 꺾으며 안방에서 승리했다. 두산은 니퍼트를 선발로 등판시켰고 LG는 임정우가 나섰다. 팽팽했던 승부는 5회말 2아웃에 타석에 들어선 홍성흔에 의해 갈렸다. 홍성흔은 LG 유원상을 상대로 적시타를 터뜨리며 홈으로 주자를 불러들였다. 홍성흔이 만든 2점은 그대로 결승점이 되었다. 두산은 9회에 LG 타선을 맞이해 2점을 실점했지만 최종 스코어 5-4로 두산의 승리를 지켜냈다. 한편 오늘 두산에게 패한 LG는 7연패를 기록하며 수렁에 빠졌다.

기사 5.

프로야구 한화의 초반 상승세를 이끄는 안영명이 4월 최우수선수의 영광을 차지했습니다. KBO는 지난 4일 치러진 출입기자단 투표에서 안영명이 유효 표 28표 가운데 22표를 얻어 다른 후보들을 압도적인 표 차이로 제치고 4월 MVP로 뽑혔다고 밝혔습니다. 시즌 초반 불펜 투수로 활약하다 선발 투수로 보직 변경을 한 안영명은 개막 이후 4월까지 10경기에 등판해 4승, 22탈삼진, 평균 자책점 1.69를 기록했습니다.

인터넷

워싱턴포스트 "올림픽 속보, 로봇이 처리"

경기 스코어-메달 순위 등 실시간 보도키로

김익현 기자 | 입력 : 2016.08.06.07:36 | 수정 : 2016.08.06.09:47

84

미국 워싱턴포스트가 올림픽 경기 소식을 좀 더 빠르고 다양하게 보도하기 위해 로봇을 활용하기로 했다.

워싱턴포스트는 5일(현지 시각) 자체 개발한 헬리오그래프(Heliograf)란 머신러닝 소프트웨어로 올림픽 소식을 자동 보도하기로 했다고 공식 발표했다.

이 회사의 기사 작성 로봇 '헬리오그래프'는 간단한 경기 결과부터 스코어까지 단순한 사실 보도 쪽을 책임질 계획이다. 이렇게 작성된 기사는 워싱턴포스트 웹사이트와 트위터 계정 등에 실시간 업데이트된다.

■ "사람들은 분석 기사-색깔있는 기사에 주력"

이 회사 데이터 과학 부문을 이끌고 있는 제레미 길버트는 "데이터와 머신러닝 기술을 활용한 자동 작성 기사는 워싱턴포스트의 보도 자체를 바꿔놓을 잠재력을 갖고 있다"면서 "앞으로 독자들에게 좀 더 개인맞춤형 뉴스 경험을 제공할 수 있을 것"이라고 강조했다.

이런 실험을 하기엔 올림픽이 가장 적합한 이벤트라는 게 길버트의 주장이다. 그는 "4년 전 런던올림픽 때는 사람 기자들이 엄청나게 많은 시간을 투자해 직접 보도했다"면서 "헬리오그래프는 워싱턴포스트 기자와 편집자들을 (이런 단순 업무로부터) 해방시켜줌으로써 좀 더 분석적이고 색깔있는 기사를 쓸 수 있도록 해 줄 것"이라고 주장했다.

Robot Journalism

- 컴퓨터 소프트웨어를 활용해 자동으로 작성되는 기사 또는 그런 기사에 중점을 둔 저널리즘

- 로봇 저널리즘에 사용되는 소프트웨어
  - 인터넷상의 각종 데이터를 수집, 정리한 후,
  - 알고리즘을 통해 이를 분류하고 의미를 해석하여 기사를 작성

LA times: Quakebot

future tense | THE CITIZEN'S GUIDE TO THE FUTURE | MARCH 17 2014 5:30 PM

The First News Report on the L.A. Earthquake Was Written by a Robot

By Will Oremus

The *LAT*'s Quakebot finished its story in seconds flat.

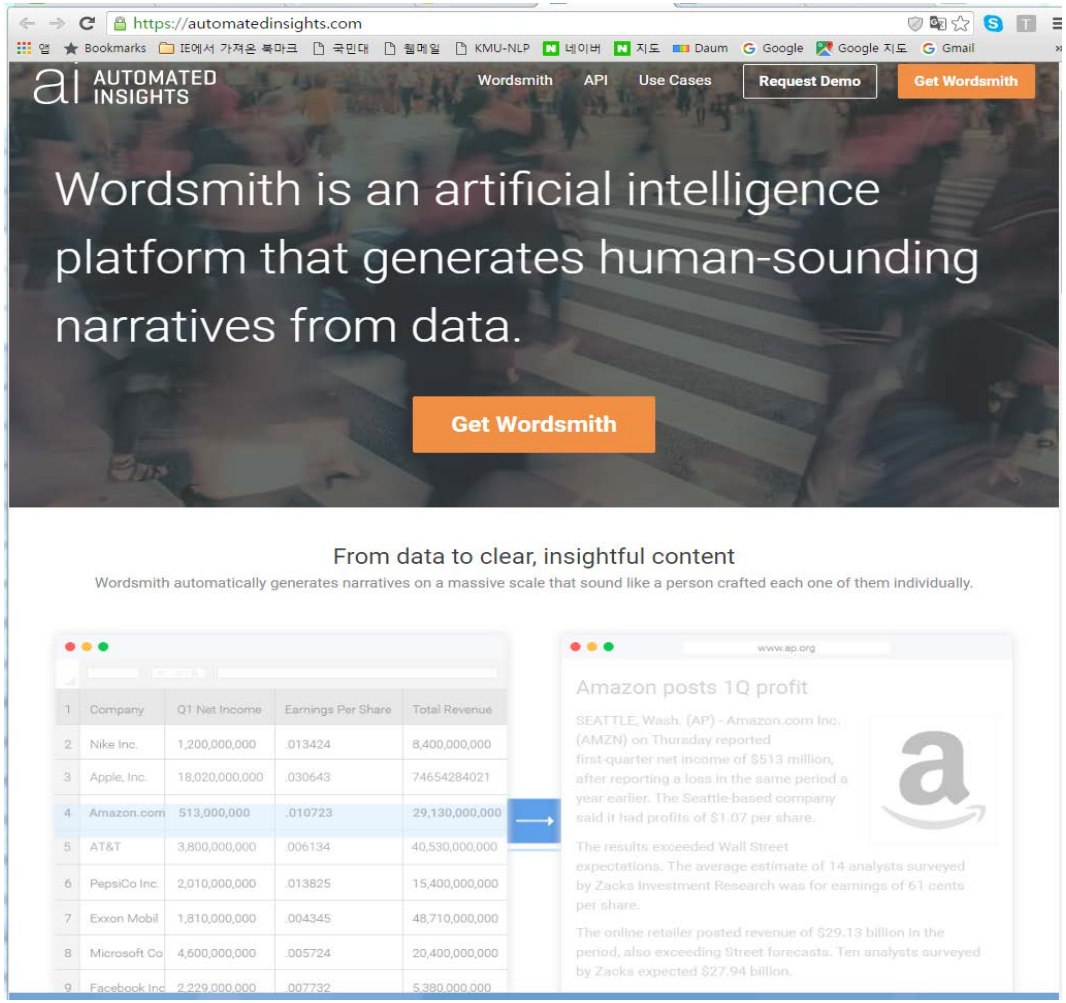
iurii / Shutterstock.com

Here's Monday morning's initial Quakebot report:

> A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles.
>
> According to the USGS, the epicenter was six miles from Beverly Hills, California, seven miles from Universal City, California, seven miles from Santa Monica, California and 348 miles from Sacramento, California. In the past ten days, there have been no earthquakes magnitude 3.0 and greater centered nearby.
>
> This information comes from the USGS Earthquake Notification Service and this post was created by an algorithm written by the author.
>
> Read more about Southern California earthquakes.

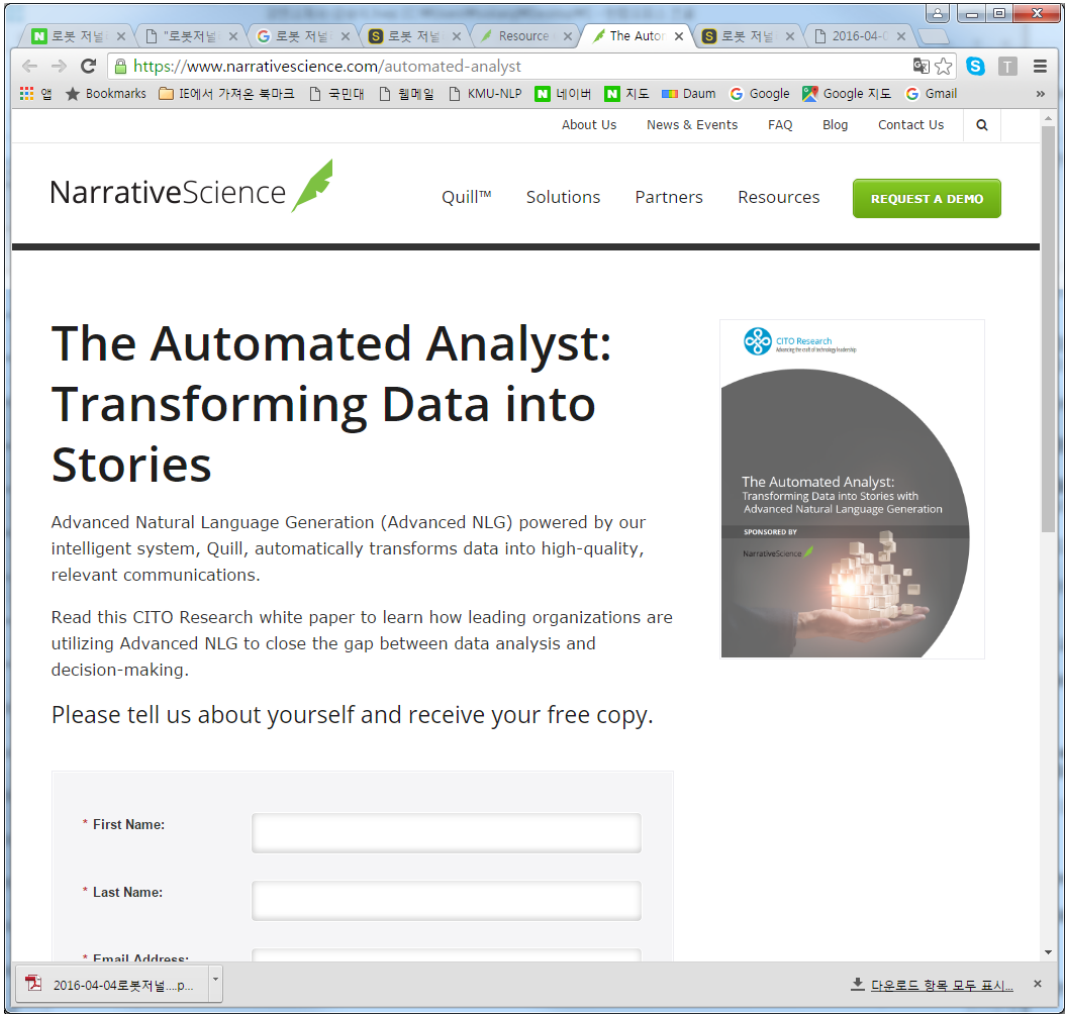
https://automatedinsights.com/

- **Wordsmith**
  - From data
  - To human-sounding narratives



https://www.narrativescience.com/

- NLG tool



Stats Monkey, 2009

- http://infolab.northwestern.edu/projects/

> "9회 2명의 주자가 나가 있었지만, LA 에인절스의 상황은 다소 비관적이었다. 그러나 블라디미르 게레로의 적시타로 에인절스는 지난 일요일 펜웨이파크에서 열린 보스턴 레드삭스와의 경기를 7 대 6으로 승리했다. 게레로는 에인절스 주자 2명을 홈으로 불러들였다. 이로써 게레로는 4타수 2안타를 기록했다."
>
> – 스탯몽키 기사 일부분

> "보스턴 레드삭스는 23년 만에 포스트시즌 경기에 도전한다는 희망을 갖고 있었다. 데이비드 핸더슨이 기념 시구를 던졌다. 핸더슨은 1986년 레드삭스와 에인절스의 아메리칸리그 챔피언십 경기에서 레드삭스가 쳐낸 9회 마지막 공격 역전 홈런의 주인공이다. 그러나 이번에는 레드삭스가 에인절스에 의해 챔피언십 경기에서 탈락했고, 핸더슨은 이번에도 경기는 마지막 순간까지 안심할 수 없다는 것을 증명했다."
>
> – 뉴욕타임스 기사 일부분

- 기자가 쓴 기사
  - 조리 있다(cohrent)
  - 잘 썼다(well written)
  - 명쾌하다(clear)
  - 읽기 편하다(pleasant to read)
  - 읽는 재미가 있다(interesting)
- 소프트웨어가 생산한 기사
  - 설명적이다(descriptive)
  - 이용하기 좋다(useable)
  - 정보가 풍부하다(informative)
  - 지루하다(boring)
  - 정확하다(accurate)
  - 신뢰할 수 있다(trustworthy)
  - 객관적이다(objective)

로봇저널 – 우리나라 사례

- http://www.mediaus.co.kr/news/articleView.html?idxno=62776

[로봇저널리즘, PM 2시 고속도로] 전국교통량 429만대 … 서울→부산 4시간 40분, 울산→서울 4시간 28분 소요

16.07.27 Forecast 기자

[로봇저널=Forecast 기자] 27일 오후 2시 기준, 서울에서 대전까지 1시간 40분, △ 서울~대구 3시간 29분, △ 서울~울산 4시간 28분, △ 서울~부산 4시간 40분이 소요될 것으로 보인다.

또한 서울에서 광주까지 소요시간은 3시간 20분, △ 서울~목포 3시간 30분, △ 서울~강릉 2시간 30분으로 예상된다.

반면, 대전에서 서울까지 1시간 40분, △ 대구~서울 3시간 23분, △ 울산~서울 4시간 28분, △ 부산~서울 4시간 20분, △ 광주~서울 3시간 40분, △ 목포~서울 3시간 30분, △ 강릉~서울 2시간 40분이 소요될 것으로 전망됐다.

한편, 27일 오후 2시 현재 전국교통량은 429만대이며 서울방향 교통량은 39만대, 지방방향 교통량은 400000대다.

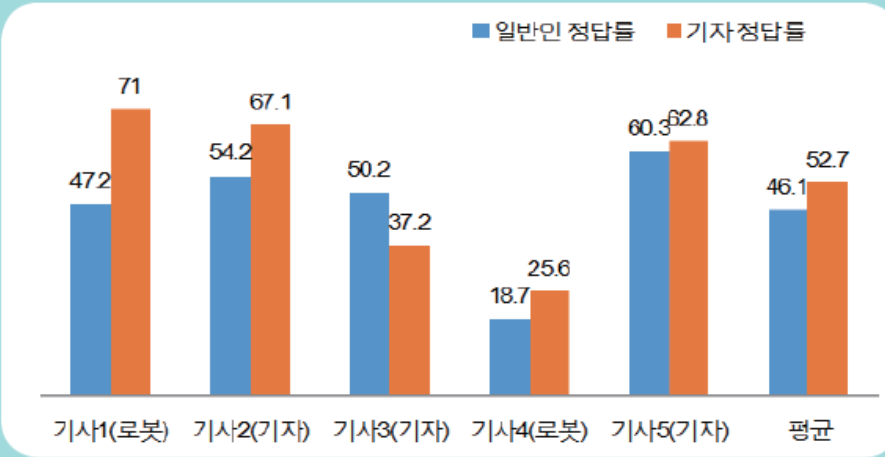
"본 기사는 로봇저널리즘 전문지 로봇저널이 공공 API 를 이용해 제작한 로봇저널리즘입니다."

master@rbjn.kr / Forecast 기자

두산은 6일 열린 홈 경기에서 LG를 5:4, 1점차로 간신히 꺾으며 안방에서 승리했다. 두산은 니퍼트를 선발로 등판시켰고 LG는 임정우가 나섰다. 팽팽했던 승부는 5회말 2아웃에 타석에 들어선 홍성흔에 의해 갈렸다. 홍성흔은 LG 유원상을 상대로 적시타를 터뜨리며 홈으로 주자를 불러들였다. 홍성흔이 만든 2점은 그대로 결승점이 되었다. 두산은 9회에 LG 타선을 맞이해 2점을 실점했지만 최종 스코어 5-4로 두산의 승리를 지켜냈다. 한편 오늘 두산에게 패한 LG는 7연패를 기록하며 수렁에 빠졌다.

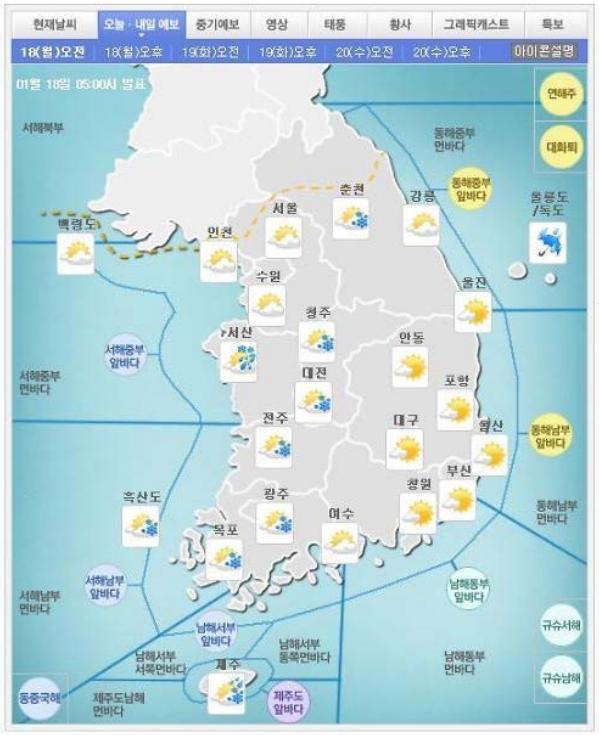
이 기사를 누가 썼을까? 이 기사의 작성자를 묻는 질문에 일반인의 81.4%, 기자의 74.4%가 '인간 기자'라고 답했다. 그런데 그 답은 틀렸다. 이 기사는 인간기자가 아닌 로봇기자, 더 정확히 말하자면 알고리즘이 작성한 기사다. 이 기사 작성의 주체를 '로봇'이라고 맞힌 사람은 일반인은 10명 중 2명, 기자는 3명이 채 안 된다.

〈그림 1〉 기사작성 주체 알아 맞히기



* 실험에 제시된 5건의 기사는 부록으로 별도 첨부하였음.

사례: 일기예보 기사 작성

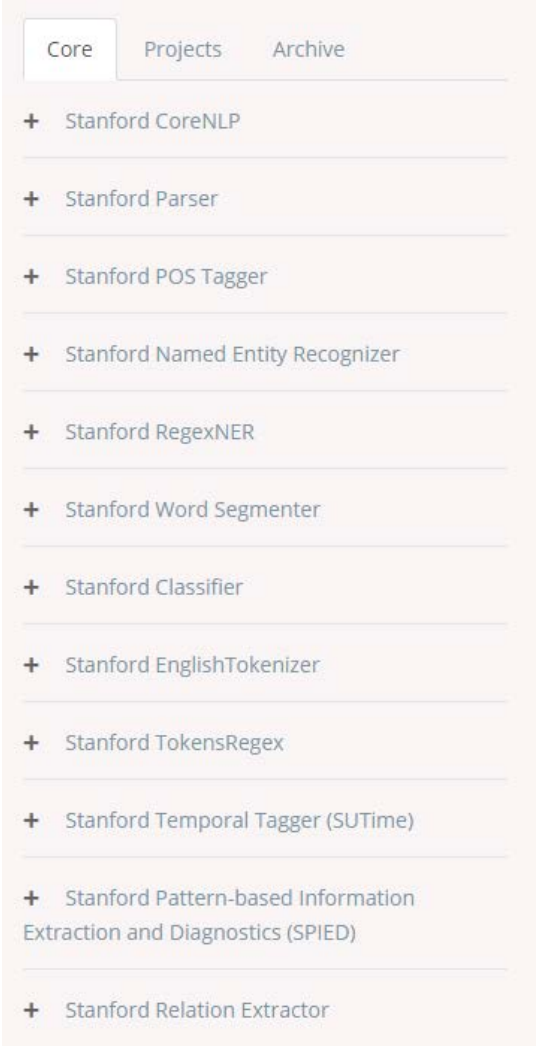

현재날씨
오늘 · 내일 예보
중기예보
영상
태풍
황사
그래픽캐스트
특보
18(월)오전
18(월)오후
19(화)오전
19(화)오후
20(수)오전
20(수)오후
아이콘설명
01월 18일 05:00시 발표
서울
인천
수원
춘천
강릉
청주
서산
대전
전주
대구
안동
포항
울산
부산
광주
여수
목포
제주

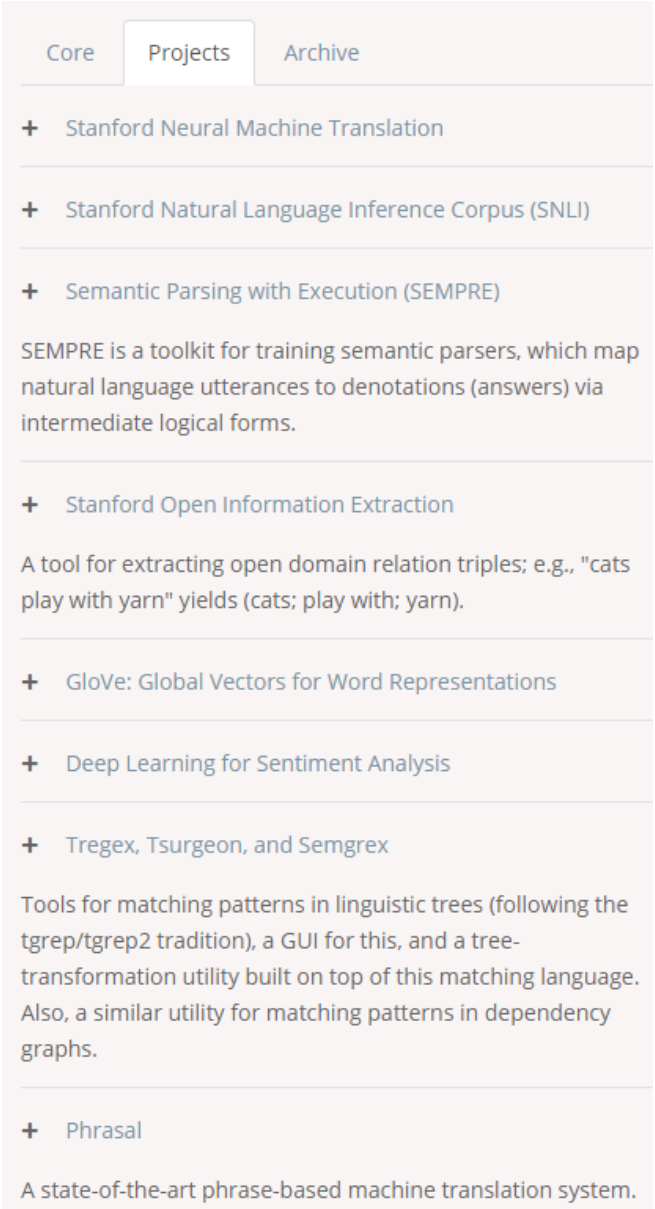

예상강수량 (1일/mm)
5~40
30~80
(100↑)


수 18 22℃
목 16 21
금 13 18
토 12 20
일 12 23
소만 13 25

NLP Resources and NLTK in Python

NLP resources in http://nlp.stanford.edu/





POS-tagging

The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today.

The/DT strongest/JJS rain/NN ever/RB recorded/VBN in/IN India/NNP shut/VBD down/RP the/DT financial/JJ hub/NN of/IN Mumbai/NNP ,/, snapped/VBD communication/NN lines/NNS ,/, closed/VBD airports/NNS and/CC forced/VBD thousands/NNS of/IN people/NNS to/TO sleep/VB in/IN their/PRP$ offices/NNS or/CC walk/VB home/NN during/IN the/DT night/NN ,/, officials/NNS said/VBD today/NN ./.

```
(ROOT
  (S
    (S
      (NP
        (NP (DT The) (JJS strongest) (NN rain))
        (VP
          (ADVP (RB ever))
          (VBN recorded)
          (PP (IN in)
            (NP (NNP India)))))
      (VP
        (VP (VBD shut)
          (PRT (RP down))
          (NP
            (NP (DT the) (JJ financial) (NN hub))
            (PP (IN of)
              (NP (NNP Mumbai)))))
        (, ,)
        (VP (VBD snapped)
          (NP (NN communication) (NNS lines)))
        (, ,)
        (VP (VBD closed)
          (NP (NNS airports)))
        (CC and)
        (VP (VBD forced)
          (NP
            (NP (NNS thousands))
            (PP (IN of)
              (NP (NNS people))))
          (S
            (VP (TO to)
              (VP
                (VP (VB sleep)
                  (PP (IN in)
                    (NP (PRP$ their) (NNS offices))))
                (CC or)
                (VP (VB walk)
                  (ADVP (RB home))
                  (PP (IN during)
                    (NP (DT the) (NN night))))))))))
    (, ,)
    (NP (NNS officials))
    (VP (VBD said)
      (NP (NN today)))
    (. .)))
```

```
det(rain-3, The-1)
amod(rain-3, strongest-2)
nsubj(shut-8, rain-3)
nsubj(snapped-16, rain-3)
nsubj(closed-20, rain-3)
nsubj(forced-23, rain-3)
advmod(recorded-5, ever-4)
partmod(rain-3, recorded-5)
prep_in(recorded-5, India-7)
ccomp(said-40, shut-8)
prt(shut-8, down-9)
det(hub-12, the-10)
amod(hub-12, financial-11)
dobj(shut-8, hub-12)
prep_of(hub-12, Mumbai-14)
conj_and(shut-8, snapped-16)
ccomp(said-40, snapped-16)
nn(lines-18, communication-17)
dobj(snapped-16, lines-18)
conj_and(shut-8, closed-20)
ccomp(said-40, closed-20)
dobj(closed-20, airports-21)
conj_and(shut-8, forced-23)
ccomp(said-40, forced-23)
dobj(forced-23, thousands-24)
prep_of(thousands-24, people-26)
aux(sleep-28, to-27)
xcomp(forced-23, sleep-28)
poss(offices-31, their-30)
prep_in(sleep-28, offices-31)
xcomp(forced-23, walk-33)
conj_or(sleep-28, walk-33)
dobj(walk-33, home-34)
det(night-37, the-36)
prep_during(walk-33, night-37)
nsubj(said-40, officials-39)
root(ROOT-0, said-40)
tmod(said-40, today-41)
```

- This output was generated with the command:
- java -mx200m edu.stanford.nlp.parser.lexparser.LexicalizedParser -retainTMPSubcategories -outputFormat "wordsAndTags,penn,typedDependencies" englishPCFG.ser.gz mumbai.txt

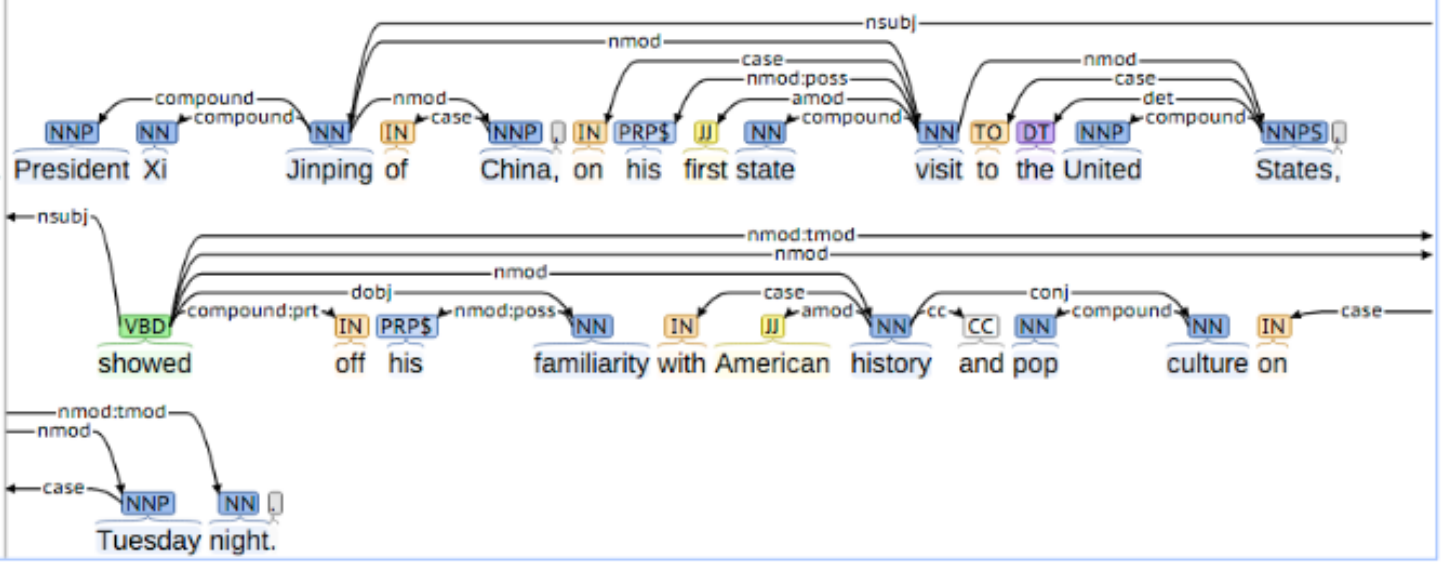
Named Entity Recognition:

1 President Xi Jinping (Person) of China (Loc), on his first (ORDINAL) state visit to the United States (Location), showed off his familiarity with American (Misc) history and pop culture on Tuesday (Date) night (Time).

Coreference:

1 President Xi Jinping of China, on his first state visit to the United States (Mention), showed off his (M) familiarity with American history and pop culture on Tuesday night.

Basic Dependencies:



NLP Took Kit: NLTK

- Natural Language Toolkit
  - http://www.nltk.org/
- Suite of classes for several NLP tasks
  - Parsing, POS tagging, classifiers…
- Easy-to-use interfaces to over 50 corpora and lexical resources
  - http://www.nltk.org/nltk_data/

Installing NLTK

- http://www.nltk.org/install.html

- Mac/Unix
  1. Install Setuptools
  2. Install Pip
  3. Install Numpy(optional)
  4. Install PyYAML and NLTK
  5. Test installation

- Windows
  1. Install Python
  2. Install Numpy(optional)
  3. Install Setuptools
  4. Install Pip
  5. Install PyYAML and NLTK
  6. Test installation

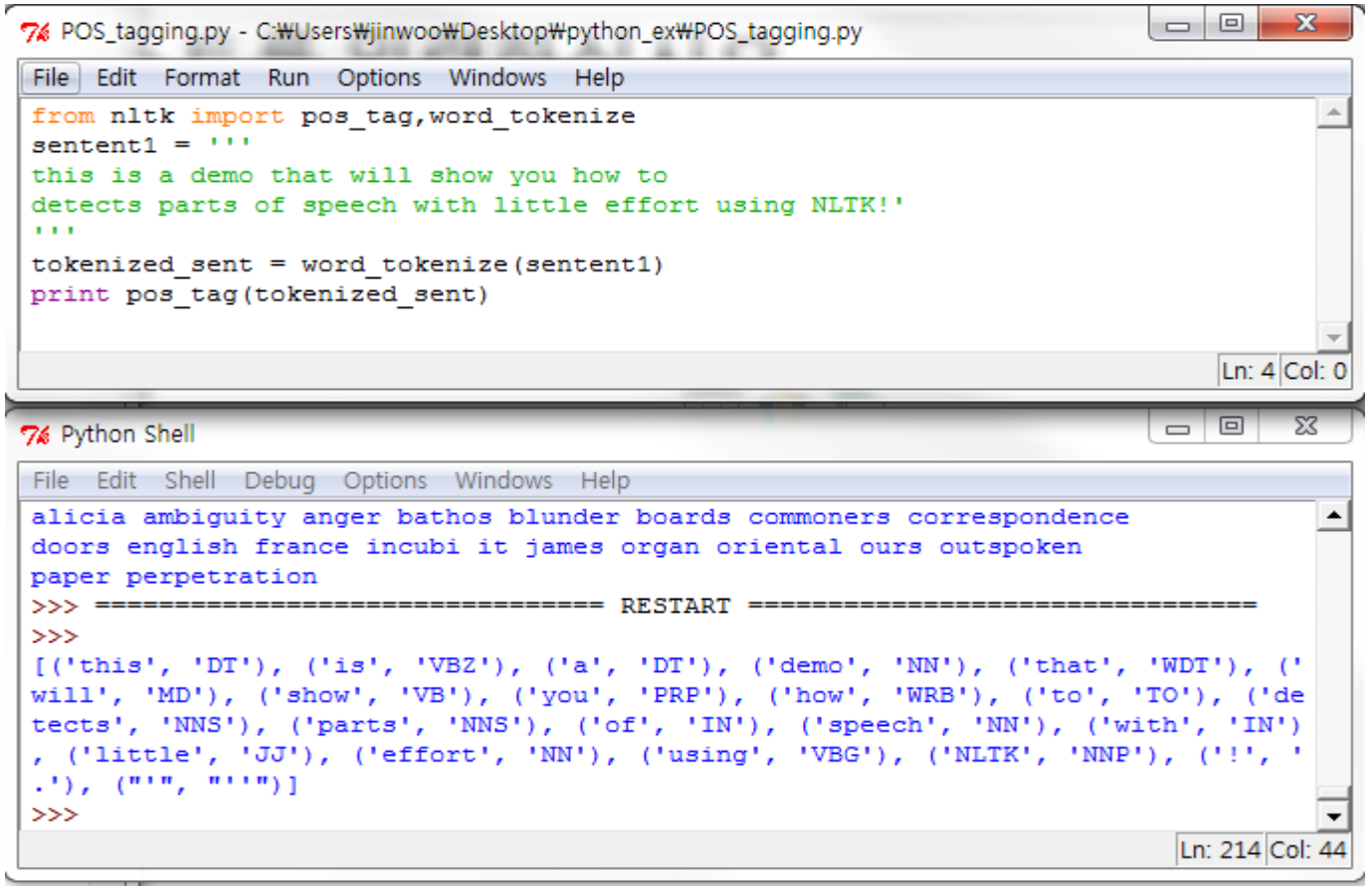
Modules

- The NLTK modules include:
  - nltk.token : processing individual elements of text, such as words or sentences
  - nltk.tagger : tagging tokens with supplemental information, such as POS or wordnet sense tags
  - nltk.parser : high-level interface for parsing texts
  - nltk.classify : classify text into categories
  - nltk.corpus : access (tagged)corpus data

  …….

- http://www.nltk.org/py-modindex.html#

Example: POS-tagging



Example: Parsing

nounphrase_chunker.py - C:₩Users₩jinwoo₩Desktop₩python_ex₩nounphrase_chunker.py

File Edit Format Run Options Windows Help

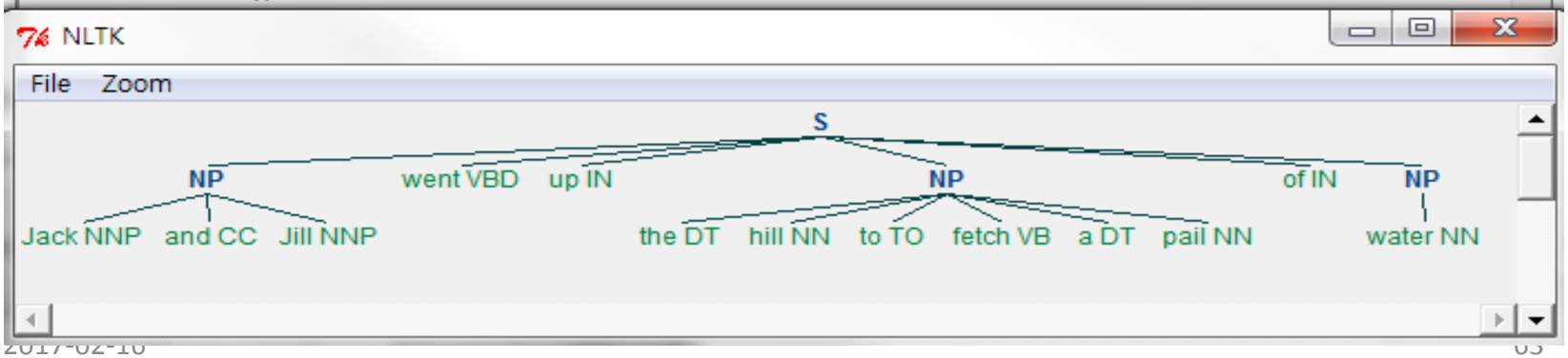
```
from nltk.chunk import *
from nltk.chunk.util import *
from nltk.chunk.regexp import *
from nltk import word_tokenize
from nltk import pos_tag

text = '''
Jack and Jill went up the hill to fetch a pail of water
'''

tokens = pos_tag(word_tokenize(text))

chunk = ChunkRule("<.*>+", "Chunk all the text")
chink = ChinkRule("<VBD|IN|\.>", "Leave verbs and prepositions out of this")
split = SplitRule("<DT><NN>", "<DT><NN>","Chunk on sequences of determiner+noun phr

chunker = RegexpChunkParser([chunk, chink, split],chunk_node='NP')
chunked = chunker.parse(tokens)
chunked.draw()
```



Example: WordNet

similarity.py - C:₩Users₩jinwoo₩Desktop₩python_ex₩similarity.py

File Edit Format Run Options Windows Help

```
from nltk.corpus import wordnet as wn

Aword = 'language'
Bword = 'barrier'

synsetsA = wn.synsets(Aword)
synsetsB = wn.synsets(Bword)

similars = []

for sseta in synsetsA:
        for ssetb in synsetsB:
                path_similarity = sseta.path_similarity(ssetb)

                if path_similarity is not None:
                        similars.append({
                                'path':path_similarity,
                                'wordA':sseta,
                                'wordB':ssetb,
                                'wordA_definition':sseta.definition,
                                'wordB_definition':ssetb.definition
                        })

similars = sorted(similars, key=lambda item: item['path'],reverse=True)

for item in similars:
        print item['wordA'],"\n",item['wordA_definition']
        print item['wordB'],"\n",item['wordB_definition']
        print 'Path similarity - ',item['path'],"\n"
```

Ln: 30 Col: 0

Python Shell

File Edit Shell Debug Options Windows Help

```
>>> ================================ RESTART ================================
>>>
Synset('linguistic_process.n.02')
the cognitive processes involved in producing and understanding linguistic commu
nication
Synset('barrier.n.02')
any condition that makes it difficult to make progress or to achieve an objectiv
e
Path similarity -  0.111111111111

Synset('language.n.05')
the mental faculty or power of vocal communication
Synset('barrier.n.02')
any condition that makes it difficult to make progress or to achieve an objectiv
e
Path similarity -  0.111111111111

Synset('language.n.01')
a systematic means of communicating by the use of sounds or conventional symbols
Synset('barrier.n.02')
any condition that makes it difficult to make progress or to achieve an objectiv
e
Path similarity -  0.1

Synset('language.n.01')
a systematic means of communicating by the use of sounds or conventional symbols
Synset('barrier.n.03')
anything serving to maintain separation by obstructing vision or access
Path similarity -  0.1

Synset('language.n.01')
a systematic means of communicating by the use of sounds or conventional symbols
Synset('barrier.n.01')
a structure or object that impedes free movement
Path similarity -  0.0909090909091
```

Ln: 431 Col: 35

For more details

- NLTK
  - http://www.nltk.org/index.html
- NLTK demo site
  - http://text-processing.com/demo/

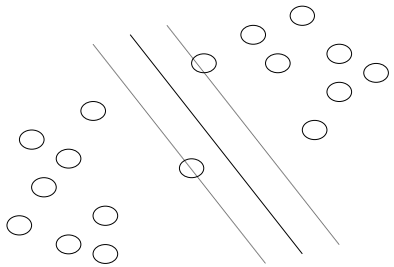
NLP & Machine Learning

Machine Learning for NLP

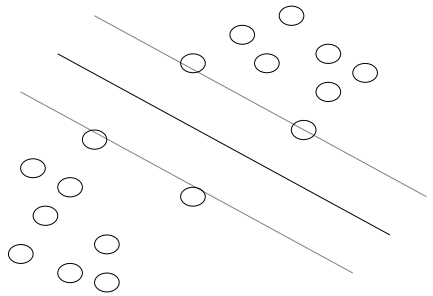
- HMM, MEM(Maximum Entropy Model)
- kNN(k-Nearest Neighbor)
- Naïve Bayse
- SVM(Support Vector Machine)
- CRF++ (Conditional Random Field)
- Neural Network

Support Vector Machine (SVM)

- Support Vector Machine (SVM)
  - 이원(binary) 패턴 인식 문제를 해결하기 위해 제안된 학습 방법
  - 두 클래스 사이에 가장 최적의 결정면(벡터 평면)을 찾는 것이 목적



smaller margin



maximal margin

SVM: binary classifier

- SVM light
  - Thorsten Joachims <thorsten@joachims.org>
    - Cornell University Department of Computer Science
  - An implementation of the SVMs in C.

- SVM 엔진 다운로드
  - http://svmlight.joachims.org/
  - source code: http://download.joachims.org/svm_light/current/svm_light.tar.gz
  - Binary versions are also available for the various systems.

SVM: Install and compile

- Create a new directory
  - $ mkdir svm_light
- Move svm_light.tar.gz into svm_light and decompress
  - $ tar xzf svm_light.tar.gz
- Compile
  - $ make
- Two executables will be created.
  - svm_learn (learning module)
  - svm_classify (classification module)

Learning Module

- svm_learn [options] example_file model_file
  - options: Refer help messages using “-?” option
  - example_file: Input file for training examples.
    - Format for classification mode
      - <Target> <Feature1>:<Value1> <F2>:<V2>…<Fn>:<Vn>
        - Target: +1 | -1 | 0
        - Feature: <integer>, Value: <float>
          - Feature/value pairs MUST be ordered by increasing feature number.
      - For example
        - -1 1:0.43 3:0.12 9284:0.2 --- Negative example
        - 1 1:0.1 10:0.45 --- Positive example
        - 0 1:0.34 5:0.13 189:0.5 --- Unknown example
  - model_file: Result of svm_learn is the model which is learned from the training examples.

Classification Module

- svm_classify [options] example_file model_file output_file
  - options: Refer help messages using “-?” option
  - example_file: Test examples in the same format as the training examples.
  - model_file: The model_file from svm_learn.
  - output_file
    - The result of svm_classify which has the predicted values.
    - The predicted values are result of the decision function for each examples.
    - The sign of the predicted value is the predicted class.
    - The zero indicates unknown

SVM 실행 예

- Example
  - http://download.joachims.org/svm_light/examples/example1.tar.gz

  - The task is to learn which Reuters articles are about "corporate acquisitions“.
  - 9947 features : Each feature corresponds to a word stem.
  - train.dat : 1000 positive and 1000 negative examples
  - test.dat : 600 test examples
  - words : A set of word stems. Features correspond to the line numbers. (9947 lines)

- 학습 모델 생성 및 실행

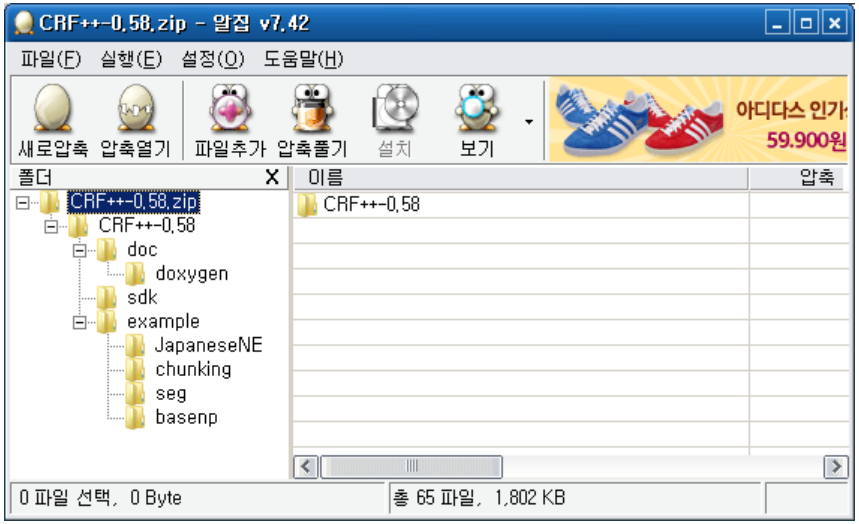
  $ svm_learn train.dat model

  $ svm_classify test.dat model predictions

CRF++

- http://crfpp.googlecode.com/svn/trunk/doc/index.html#download

- CRF++-0.58.tar.gz -- Source
- CRF++-0.58.zip
  - Binary for MS-Windows



CRF 통합 가능한 언어

- C++, Java, Python, Perl, Ruby 등

언어	설치 Directory	설명	비고
C++	CRF++-0.58/sdk	C++에서 CRF++라이브러리 연동 방법 제공	swig를 이용한 스크립트언어 C++ 라이브러리 인터페이스
JAVA	CRF++-0.58/java	JAVA 에서 CRF++라이브러리 연동 방법 제공	
Python	CRF++-0.58/python	Python 에서 CRF++라이브러리 연동 방법 제공	
Perl	CRF++-0.58/perl	Perl 에서 CRF++라이브러리 연동 방법 제공	
Ruby	CRF++-0.58/ruby	Ruby 에서 CRF++라이브러리 연동 방법 제공	

CRF++-0.58/example/basenp/

```
[taeseok@localhost CRF++-0.58]$ cd example/basenp/
exec.sh template test.data train.data
[taeseok@localhost python]$ ../../crf_learn -c 10.0 template train.data model
…
iter=33 terr=0.00000 serr=0.00000 act=32970 obj=19.70277 diff=0.00019
iter=34 terr=0.00000 serr=0.00000 act=32970 obj=19.70237 diff=0.00002
iter=35 terr=0.00000 serr=0.00000 act=32970 obj=19.70003 diff=0.00012
iter=36 terr=0.00000 serr=0.00000 act=32970 obj=19.69958 diff=0.00002
iter=37 terr=0.00000 serr=0.00000 act=32970 obj=19.69887 diff=0.00004
iter=38 terr=0.00000 serr=0.00000 act=32970 obj=19.69855 diff=0.00002

Done!0.15 s

[taeseok@localhost python]$ ../../crf_test  -m model test.data > output.txt
…
of      IN      O       O
Columbus        NNP     B       B
,       ,       O       O
Ohio    NNP     B       B
,       ,       O       O
grew    VBD     O       O
3.8     CD      B       B
%       NN      I       I
.       .       O       O

[taeseok@localhost python]$ ./conlleval.pl -d "\t" < output.txt
processed 19172 tokens with 5051 phrases; found: 4978 phrases; correct: 4285.
accuracy:  93.67%; precision:  86.08%; recall:  84.83%; FB1:  85.45
              : precision:  86.08%; recall:  84.83%; FB1:  85.45  4978
              : precision:  86.08%; recall:  84.83%; FB1:  85.45  4978
```

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

U23:%x[0,1]

# Bigram
B
```

http://www.cnts.ua.ac.be/conll2000/chunking/output.html

마지막으로…

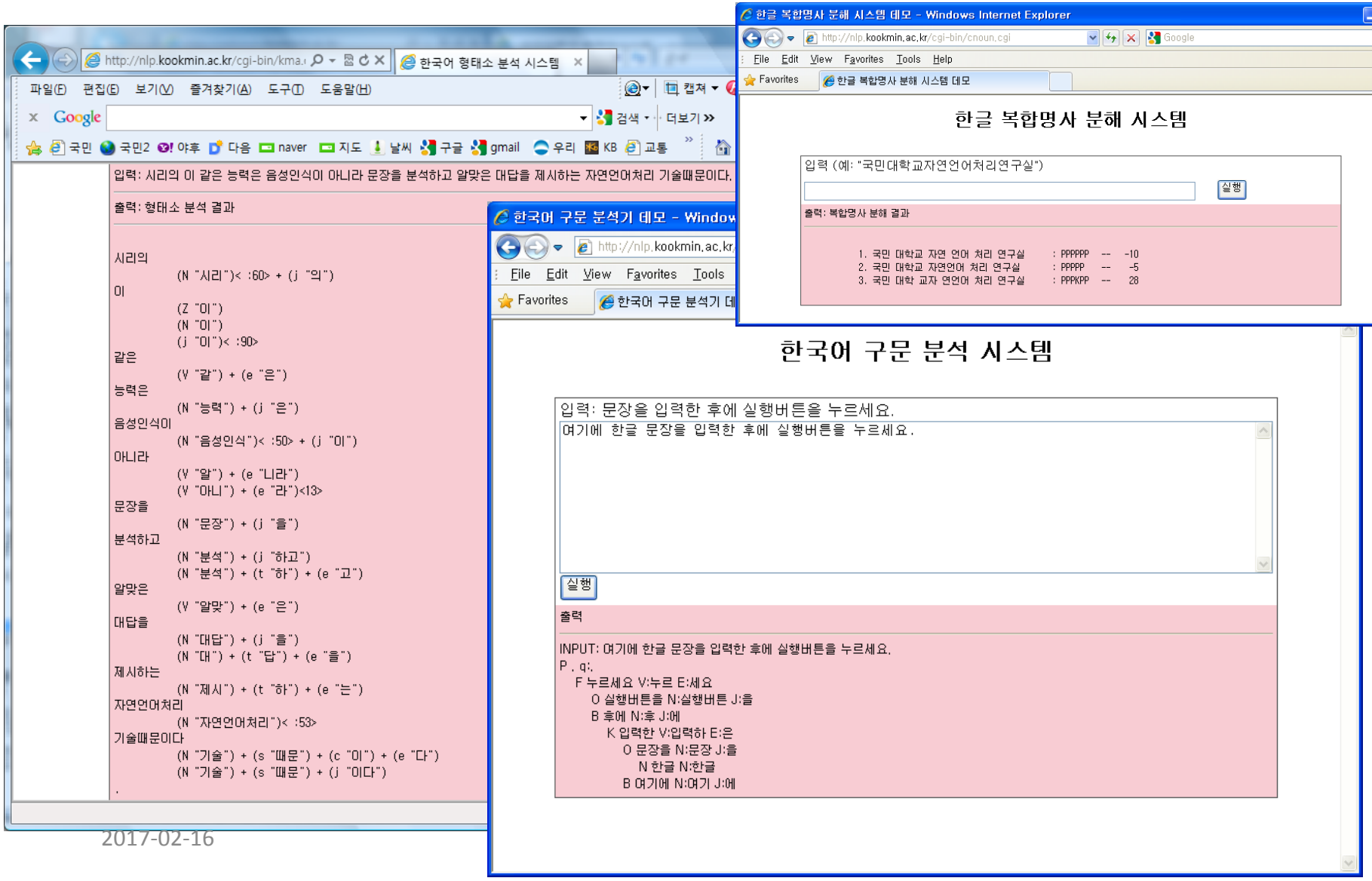
http://nlp.kookmin.ac.kr

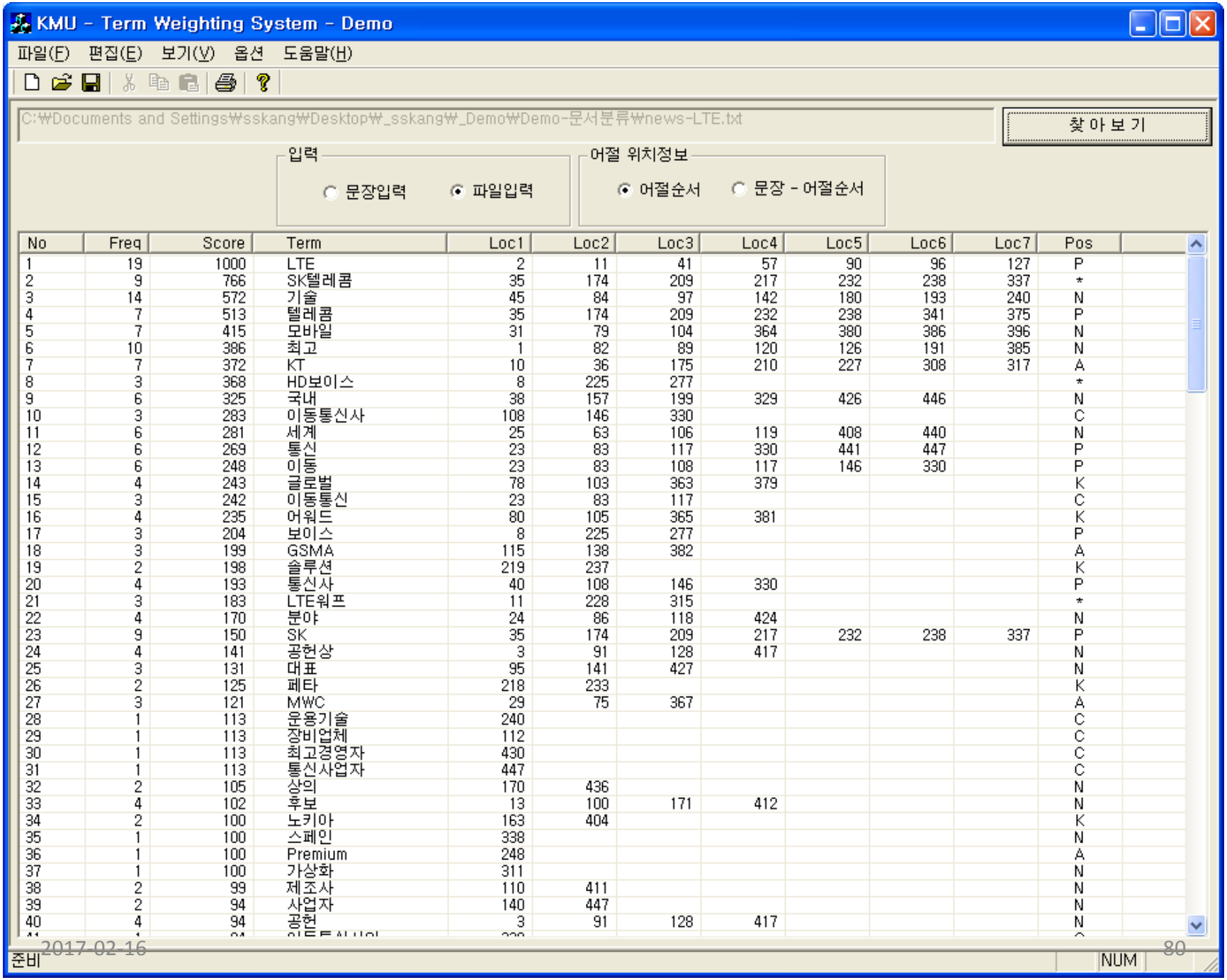
- 한국어 형태소 분석, 구문 분석
- 색인어 추출 및 가중치 계산

- 복합명사 분해
- 맞춤법 검사 및 교정

- 자동 문서 분류, 자동 띄어쓰기 등

형태소 분석과 구문분석


한국어 형태소 분석 시스템
입력: 시리의 이 같은 능력은 음성인식이 아니라 문장을 분석하고 알맞은 대답을 제시하는 자연언어처리 기술때문이다.
출력: 형태소 분석 결과
시리의
(N "시리")< :60> + (j "의")
이
(Z "이")
(N "이")
(j "이")< :90>
같은
(V "같") + (e "은")
능력은
(N "능력") + (j "은")
음성인식이
(N "음성인식")< :50> + (j "이")
아니라
(V "알") + (e "니라")
(V "아니") + (e "라")<13>
문장을
(N "문장") + (j "을")
분석하고
(N "분석") + (j "하고")
(N "분석") + (t "하") + (e "고")
알맞은
(V "알맞") + (e "은")
대답을
(N "대답") + (j "을")
(N "대") + (t "답") + (e "을")
제시하는
(N "제시") + (t "하") + (e "는")
자연언어처리
(N "자연언어처리")< :53>
기술때문이다
(N "기술") + (s "때문") + (c "이") + (e "다")
(N "기술") + (s "때문") + (j "이다")
.
한글 복합명사 분해 시스템 데모 - Windows Internet Explorer
한글 복합명사 분해 시스템
입력 (예: "국민대학교자연언어처리연구실")
실행
출력: 복합명사 분해 결과
1. 국민 대학교 자연 언어 처리 연구실 : PPPPPP -- -10
2. 국민 대학교 자연언어 처리 연구실 : PPPPP -- -5
3. 국민 대학 교자 연언어 처리 연구실 : PPPKPP -- 28
한국어 구문 분석기 데모 - Windows Internet Explorer
한국어 구문 분석 시스템
입력: 문장을 입력한 후에 실행버튼을 누르세요.
여기에 한글 문장을 입력한 후에 실행버튼을 누르세요.
실행
출력
INPUT: 여기에 한글 문장을 입력한 후에 실행버튼을 누르세요.
P , q:,
F 누르세요 V:누르 E:세요
O 실행버튼을 N:실행버튼 J:을
B 후에 N:후 J:에
K 입력한 V:입력하 E:은
O 문장을 N:문장 J:을
N 한글 N:한글
B 여기에 N:여기 J:에


KMU - Term Weighting System - Demo
파일(F) 편집(E) 보기(V) 옵션 도움말(H)
C:₩Documents and Settings₩sskang₩Desktop₩_sskang₩_Demo₩Demo-문서분류₩news-LTE.txt
찾 아 보 기
입력
문장입력
파일입력
어절 위치정보
어절순서
문장 - 어절순서
No Freq Score Term Loc1 Loc2 Loc3 Loc4 Loc5 Loc6 Loc7 Pos
1 19 1000 LTE 2 11 41 57 90 96 127 P
2 9 766 SK텔레콤 35 174 209 217 232 238 337 *
3 14 572 기술 45 84 97 142 180 193 240 N
4 7 513 텔레콤 35 174 209 232 238 341 375 P
5 7 415 모바일 31 79 104 364 380 386 396 N
6 10 386 최고 1 82 89 120 126 191 385 N
7 7 372 KT 10 36 175 210 227 308 317 A
8 3 368 HD보이스 8 225 277 *
9 6 325 국내 38 157 199 329 426 446 N
10 3 283 이동통신사 108 146 330 C
11 6 281 세계 25 63 106 119 408 440 N
12 6 269 통신 23 83 117 330 441 447 P
13 6 248 이동 23 83 108 117 146 330 P
14 4 243 글로벌 78 103 363 379 K
15 3 242 이동통신 23 83 117 C
16 4 235 어워드 80 105 365 381 K
17 3 204 보이스 8 225 277 P
18 3 199 GSMA 115 138 382 A
19 2 198 솔루션 219 237 K
20 4 193 통신사 40 108 146 330 P
21 3 183 LTE워프 11 228 315 *
22 4 170 분야 24 86 118 424 N
23 9 150 SK 35 174 209 217 232 238 337 P
24 4 141 공헌상 3 91 128 417 N
25 3 131 대표 95 141 427 N
26 2 125 페타 218 233 K
27 3 121 MWC 29 75 367 A
28 1 113 운용기술 240 C
29 1 113 장비업체 112 C
30 1 113 최고경영자 430 C
31 1 113 통신사업자 447 C
32 2 105 상의 170 436 N
33 4 102 후보 13 100 171 412 N
34 2 100 노키아 163 404 K
35 1 100 스페인 338 N
36 1 100 Premium 248 A
37 1 100 가상화 311 N
38 2 99 제조사 110 411 N
39 2 94 사업자 140 447 N
40 4 94 공헌 3 91 128 417 N
준비
NUM