# Text Mining with R

정우준
woojune.rdata@gmail.com

# 텍스트 데이터의 중요성

- 많은 정보가 온라인으로 이전되면서 텍스트 데이터의 중요성이 부각됨.

- 수치적인 테이블 형식의 데이터를 주된 분석 대상으로 하던 기존의 방식과는 다름.

# 텍스트 분석의 정의

- 텍스트 분석은 텍스트를 활용하여 정보 함의(information content)를 모형화/구조화하기는 언어학, 통계학, 그리고 기계학습 등의 집합적 분석임.

- 텍스트 데이터로부터 의미 있는 정보를 찾아낼 수 있을 것이라는 기대는 "통계적 의미론 가설(statistical semantics hypothesis)"에 근거함.

- 텍스트 마이닝의 기틀이 되는 이 가설은 "사람들의 글, 말 등에서 드러나는 단어 사용의 통계적 규칙성으로부터 사람들이 말하고자 하는 바를 찾아낼 수 있다"(Turney and Pantel, 2010)는 전제를 바탕으로 함.

- *출처: <https://en.wikipedia.org/wiki/Text_mining#cite_note-1>*

# 텍스트 마이닝의 내용

- 텍스트 마이닝(text mining)의 기본 절차는 자료 처리(data processing) 과정과 자료 분석(data analysis) 과정으로 크게 구분 가능함.

- 데이터 처리과정(data processing): 정보 검색(IR), 정보추출(IE), 자연어 처리(NLP) 등의 절차를 통하여 수집한 데이터를 가공 및 정제하는 과정

- 데이터 분석(data analysis): 텍스트로부터 의미 있는 추세와 패턴 및 지식을 발견하기 위하여 데이터 마이닝, 기계학습, 통계학 등을 활용하는 과정 (Hotho et al., 2005).

# 텍스트 마이닝의 내용 – 데이터 처리

- 데이터 처리과정(data processing): 다음의 세 가지로 구분

  ➔ 정보 검색

  ➔ 정보 추출

  ➔ 자연어 처리

# 텍스트 마이닝의 내용 – 데이터 처리: 정보 검색

사용자가 원하는 키워드를 기반으로 원하는 정보가 포함된 텍스트 데이터가 들어있는 문서(document)를 탐색하는 것.

사용 목적에 따라 웹 검색(web search), 개인 정보 검색 (personal information retrieval), 기업이나 기관, 특정 영역 검색 (enterprise, institutional, and domail-specific search) 등 세 가지의 형태로 구별(Manning et al., 2008).

# 텍스트 마이닝의 내용 – 데이터 처리: 정보 추출

다음의 포함하는 일련의 과정을 거치는 것을 말하는 것으로, 특정한 문서로부터 구체적인 정보를 정체하는 것(Hotho et al., 2005).

➔토큰화(tokenization): 텍스트의 공백과 모든 구두점을 제거하여 연속하는 단어의 열(stream of words)로 분할하는 것(Hotho et al., 2005),
➔문장 분할(sentence segmentation)
➔품사 배치(part-of-speech assignment)
➔독립적 개체 인식(identification of named entities)

을 포함하는 일련의 과정을 거치는 것으로, 특정한 문서로부터 구체적인 정보를 정제하는 것

# 텍스트 마이닝의 내용 – 데이터 처리: 자연어 처리

구조와 형태가 복잡한 자연어를 컴퓨터로 분석하기 위해 가공하는 단계.
크게 형태소 분석, 동사 분석, 의미 분석, 화용 분석으로 나뉨.

➔ 형태소 분석: 하나의 문장을 분해 가능한 최소한의 단위로 분리하고 분석하는 것을 말하는데 자연어 처리에서 형태소 분석은 어휘사전(lexicon)을 기반으로 입력문자의 형태소를 분석(matching)하는 것.

➔ 동사 분석: 문장의 어순이나 문법 등이 동사 규칙(syntactic rule)에 적절한지를 분석하는 것.

➔ 의미 분석: 문장의 의미가 적절한지를 판단하는 것

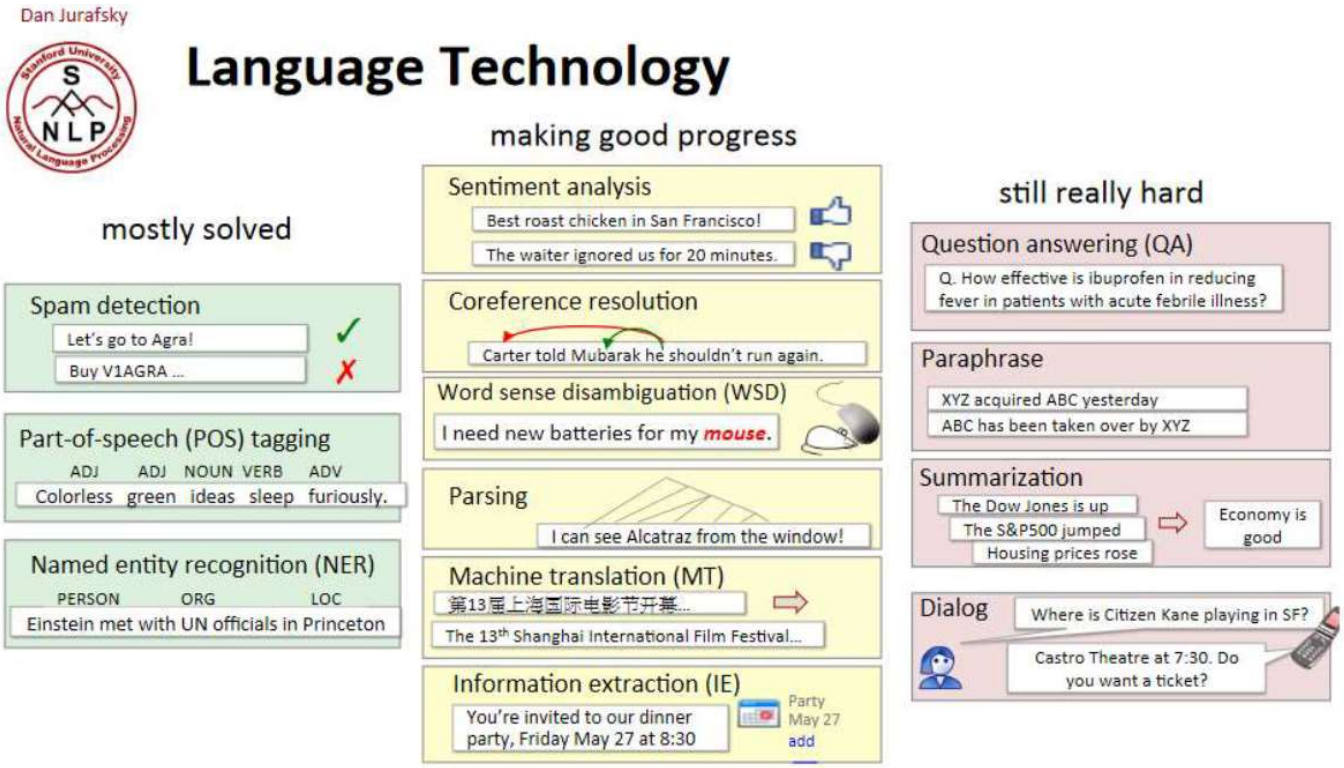
➔ 화용분석: 언어의 사회적 기능 등 해당 언어의 실세계를 보는 것으로 맥락상의 적절성을 판단하는 것(신명철 등, 2005).

# 텍스트 마이닝의 내용 – 데이터 분석

- 데이터 분석(data analysis): 텍스트로부터 의미 있는 추세와 패턴 및 지식을 발견하기 위하여 데이터 마이닝, 기계학습, 통계학 등을 활용하는 과정 (Hotho et al., 2005).

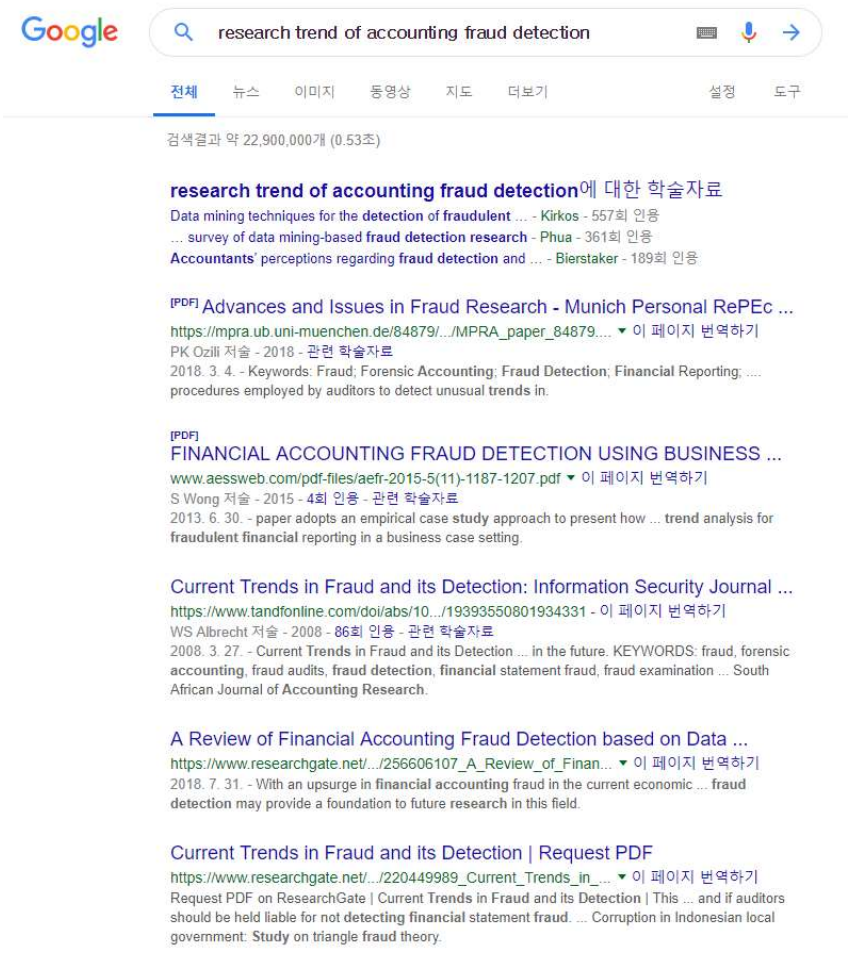
- 데이터 처리과정이 원시자료의 가공 단계라고 한다면 데이터 분석은 데이터 마이닝, 기계학습, 통계학 등을 활용하여 의미 있는 결과를 도출하는 과정이다(Jarman, 2011).

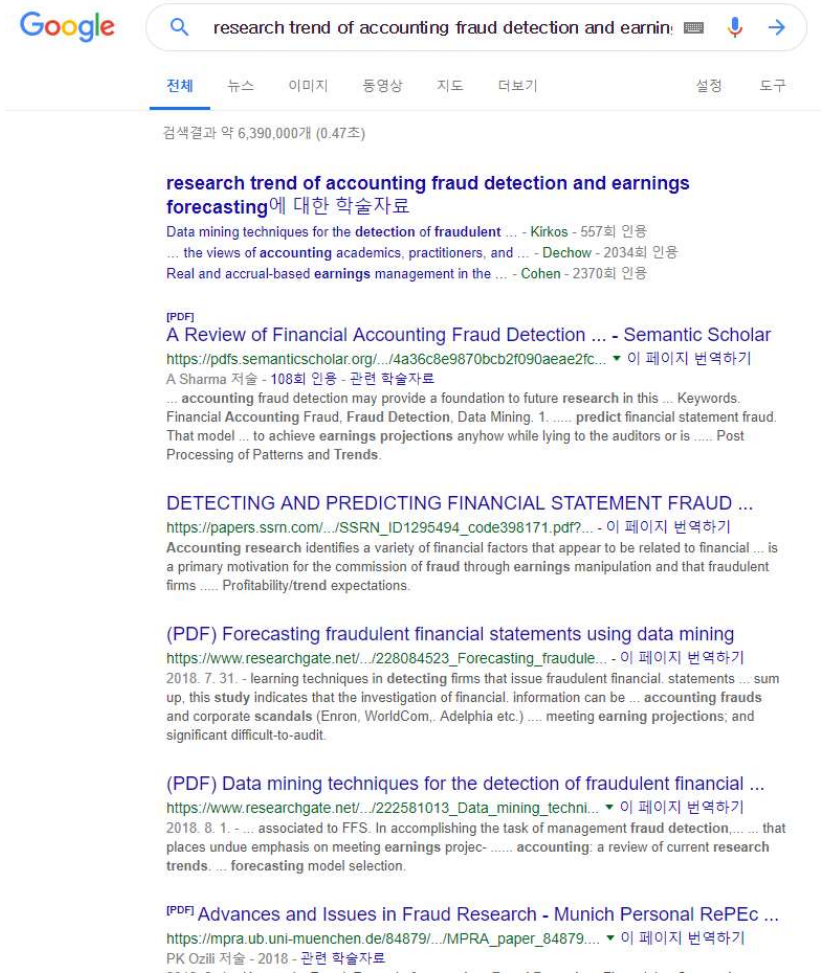
# 텍스트 마이닝 기법



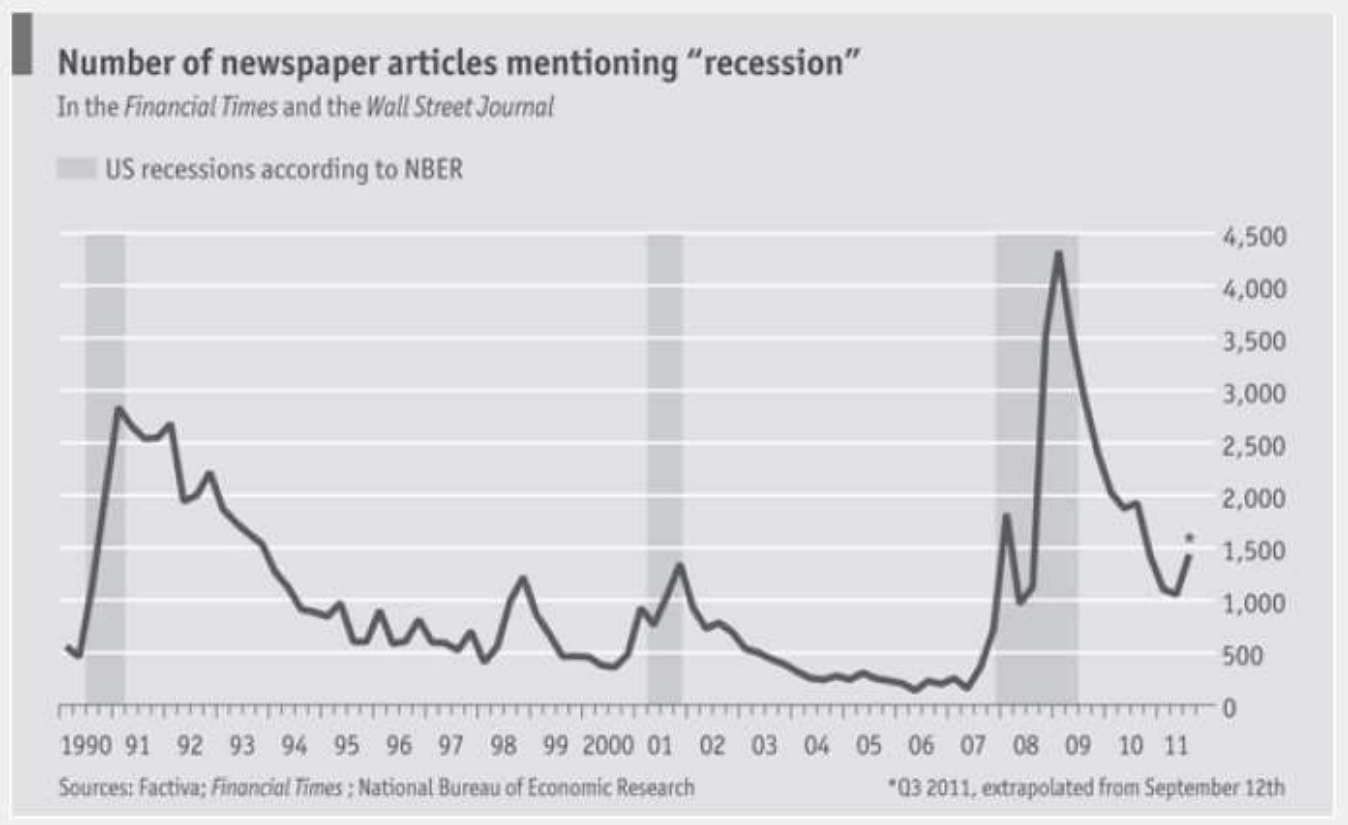

https://www.slideshare.net/semanticsconference/keynote-new-convergences-between-natural-language-processing-and-knowledge-engineering

# 텍스트 마이닝의 적용 사례: 문서 검색

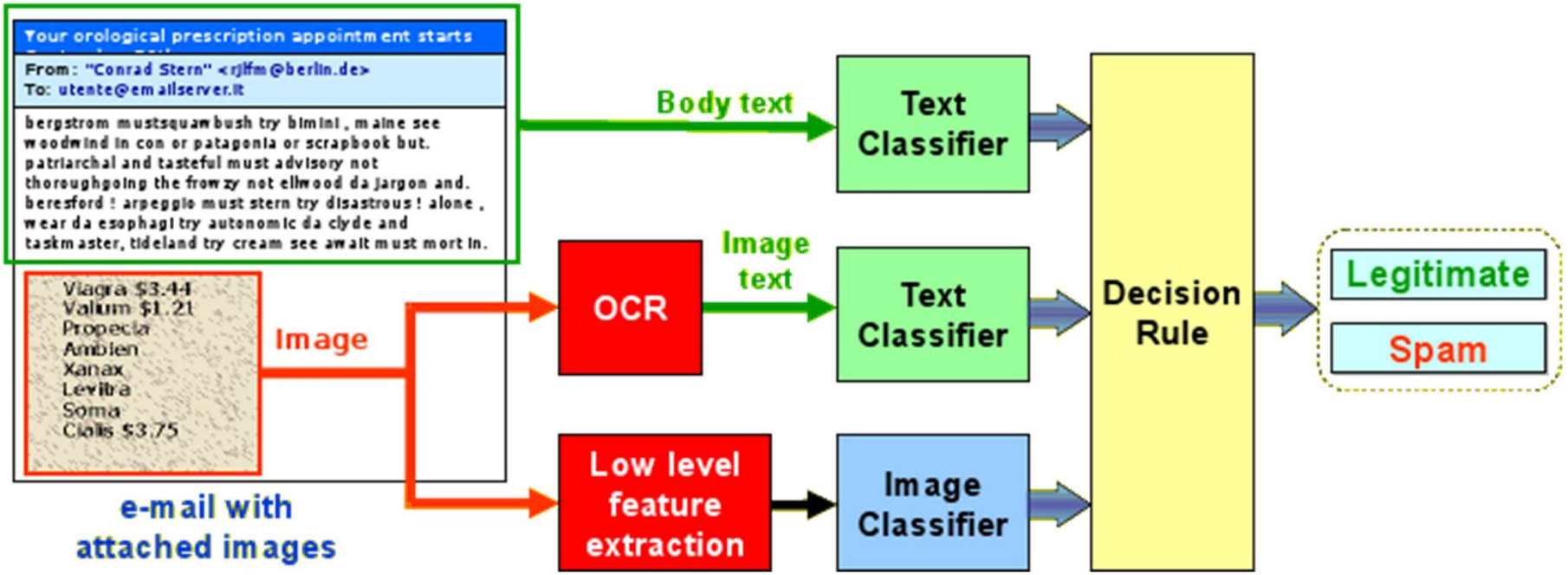


# 텍스트 마이닝의 적용 사례: 검색 트렌드 분석



# 텍스트 마이닝의 적용 사례: 이메일 스팸 처리



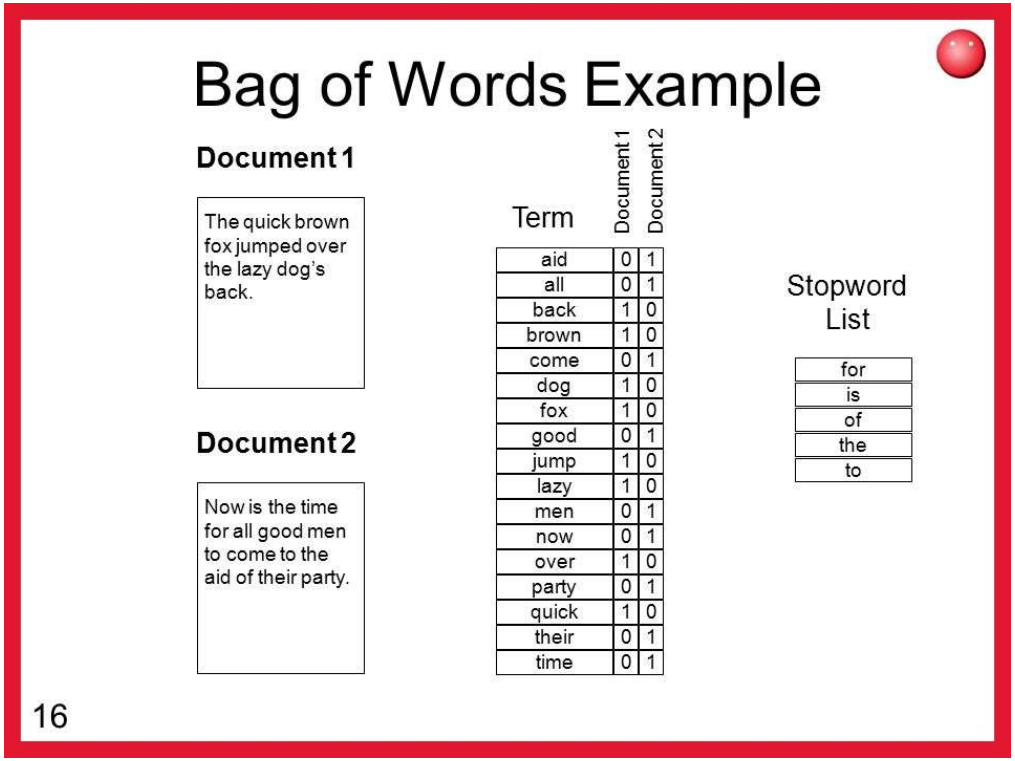

http://pralab.diee.unica.it/en/SpamFiltering

# 텍스트 마이닝의 적용 사례: contextual advertizing


http://cpamedias.com/contextual-advertising/

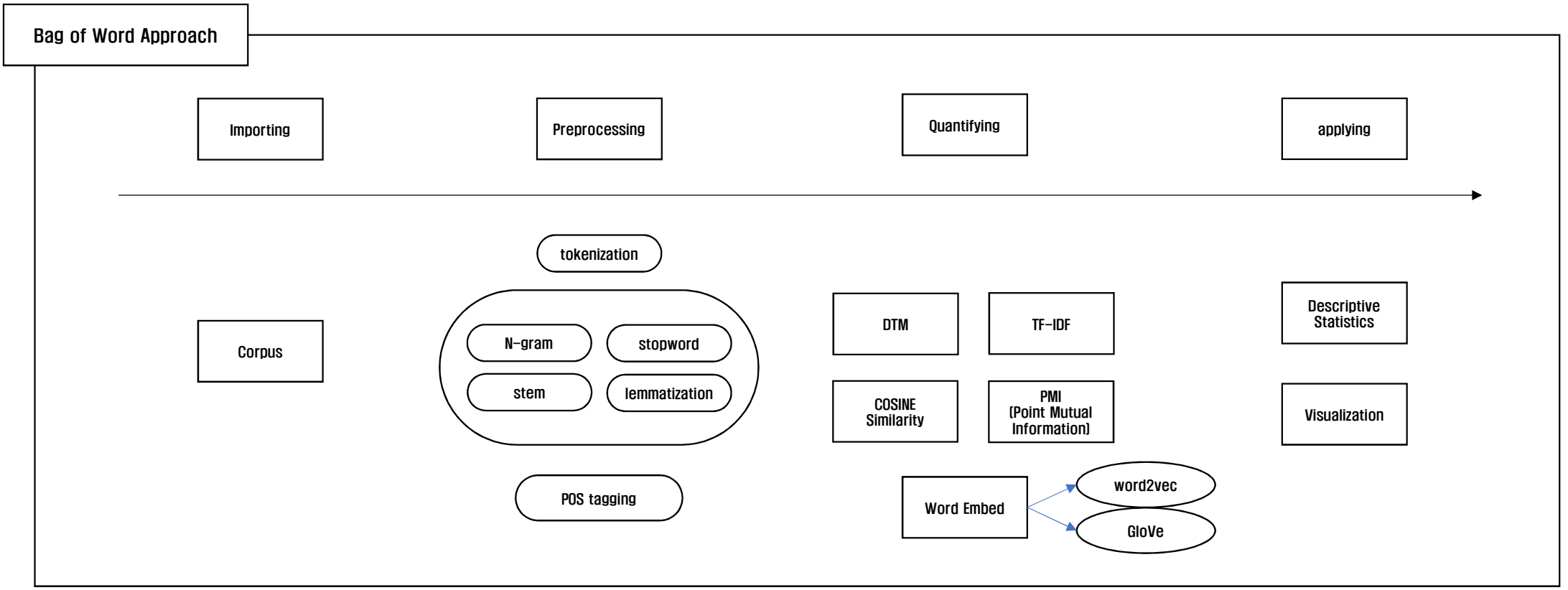
# 텍스트 마이닝의 절차 – Bag of Words Approach

- 가장 간단하지만 효과적이어서 널리 사용
- 문서의 장, 문단, 문장 등의 구조를 고려하지 않고, 단어의 출현 빈도만으로 분석함
- 보완적으로 N-gram을 사용할 수 있음(BoW = 1-gram)




https://slideplayer.com/slide/5156162/

# 텍스트 마이닝의 절차




➔ 전략: 의미의 손실이 다소 발생하더라도 일정한 가정 아래에서 텍스트 데이터를 가능한 간략한 수치형 자료로 표현하는 방법을 선택하는 것이 현명함.

# 텍스트 데이터의 특징

- 사람이 보기에는 동일한(다른) 의미 / 컴퓨터는 다르게(동일하게) 인식:

  ➔ Trump vs. trump

- 거의 제한 없이 다양한 변형:

  ➔ machine learning, machine-learning, ML, ml, Machine Learning, Machine-Learning,

- 그리고 수없이 많은 고려점들,, ㅜㅜ

# Tokenization과 POS tagging

tokenization

"금융통화위원회/는 다음 통화정책/방향 결정/시/까지 한국은행 기준금리/를 현 수준/ (/1/./25/%)/에서 유지/하여 통화정책/을 운용/하기로 하였다/."

POS tagging

"금융통화위원회/Noun 는/Josa 다음/Noun 통화정책/Noun 방향/Noun 결정/Noun 시/Noun까지/Josa 한국은행/Noun 기준금리/Noun 를/Josa ..."

# TDM/DTM

## TDM vs. DTM

| | Tweet 1 | Tweet 2 | Tweet 3 | ... | Tweet N |
|---|---|---|---|---|---|
| Term 1 | 0 | 0 | 0 | 0 | 0 |
| Term 2 | 1 | 1 | 0 | 0 | 0 |
| Term 3 | 1 | 0 | 0 | 0 | 0 |
| ... | 0 | 0 | 3 | 1 | 1 |
| Term M | 0 | 0 | 0 | 1 | 0 |

Term Document Matrix (TDM)

| | Term 1 | Term 2 | Term 3 | ... | Term M |
|---|---|---|---|---|---|
| Tweet 1 | 0 | 1 | 1 | 0 | 0 |
| Tweet 2 | 0 | 1 | 0 | 0 | 0 |
| Tweet 3 | 0 | 0 | 0 | 3 | 0 |
| ... | 0 | 0 | 0 | 1 | 1 |
| Tweet N | 0 | 0 | 0 | 1 | 0 |

Document Term Matrix (DTM)

```
> # Generate TDM
> coffee_tdm <- TermDocumentMatrix(clean_corp)

> # Generate DTM
> coffee_dtm <- DocumentTermMatrix(clean_corp)
```

http://talimi.se/r/tm/

# TF-IDF

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

**TF-IDF**

Term $x$ within document $y$

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

$df_x$ = number of documents containing $x$

$N$ = total number of documents

https://www.slideshare.net/osujin121/ss-44186451

# Word Embedding

Bag-of-words 표현방식에 의한 고차원의 성긴(sparse) 구조를 지닌 벡터의 한계점을 보완하기 위해 저차원의 조밀한(dense) 벡터 공간에 표현하는 방법들이 제안되어 옴.

➔ Word2vec: 인공신경망(neural network) 모형을 사용하여 특정 목적함수를 최적화하는 벡터 공간을 구함.

➔ GloVe: 단어들의 동시 출현 빈도수를 가중최소제곱법(weighted least squares)으로 최적화하여 단어 사이의 의미 유사도를 구하는 방법

# R을 활용한 텍스트 마이닝

*CRAN*
Mirrors
What's new?
Task Views
Search

*About R*
R Homepage
The R Journal

*Software*
R Sources
R Binaries
Packages
Other

*Documentation*
Manuals
FAQs
Contributed

CRAN Task View: Natural Language Processing

**Maintainer:** Fridolin Wild, Performance Augmentation Lab (PAL, Department of Computing and Communications Technologies, Oxford Brookes University, UK
**Contact:** wild at brookes.ac.uk
**Version:** 2017-11-29
**URL:** https://CRAN.R-project.org/view=NaturalLanguageProcessing

Natural language processing has come a long way since its foundations were laid in the 1940s and 50s (for an introduction see, e.g., Jurafsky and Martin (2008): Speech and Language Processing, Pearson Prentice Hall). This CRAN task view collects relevant R packages that support computational linguists in conducting analysis of speech and language on a variety of levels - setting focus on words, syntax, semantics, and pragmatics.

In recent years, we have elaborated a framework to be used in packages dealing with the processing of written material: the package tm. Extension packages in this area are highly recommended to interface with tm's basic routines and useRs are cordially invited to join in the discussion on further developments of this framework package. To get into natural language processing, the cRunch service and tutorials may be helpful.

Frameworks:

- tm provides a comprehensive text mining framework for R. The Journal of Statistical Software article Text Mining Infrastructure in R gives a detailed overview and presents techniques for count-based analysis methods, text clustering, text classification and string kernels.
- tm.plugin.dc allows for distributing corpora across storage devices (local files or Hadoop Distributed File System).
- tm.plugin.mail helps with importing mail messages from archive files such as used in Thunderbird (mbox, eml).
- tm.plugin.alceste allows importing text corpora written in a file in the Alceste format.
- tm.plugin.factiva, tm.plugin.lexisnexis, tm.plugin.europresse allow importing press and Web corpora from (respectively) Dow Jones Factiva, LexisNexis, and Europresse.
- tm.plugin.webmining allow importing news feeds in XML (RSS, ATOM) and JSON formats. Currently, the following feeds are implemented: Google Blog Search, Google Finance, Google News, NYTimes Article Search, Reuters News Feed, Yahoo Finance, and Yahoo Inplay.
- RcmdrPlugin.temis is an Rcommander plug-in providing an integrated solution to perform a series of text mining tasks such as importing and cleaning a corpus, and analyses like terms and documents counts, vocabulary tables, terms co-occurrences and documents similarity measures, time series analysis, correspondence analysis and hierarchical clustering.
- openNLP provides an R interface to OpenNLP , a collection of natural language processing tools including a sentence detector, tokenizer, pos-tagger, shallow and full syntactic parser, and named-entity detector, using the Maxent Java package for training and using maximum entropy models.
- Trained models for English and Spanish to be used with openNLP are available from http://datacube.wu.ac.at/ as packages openNLPmodels.en and openNLPmodels.es, respectively.
- RWeka is a interface to Weka which is a collection of machine learning algorithms for data mining tasks written in Java. Especially useful in the context of natural language processing is its functionality for tokenization and stemming.
- tidytext provides means for text mining for word processing and sentiment analysis using dplyr, ggplot2, and other tidy tools.
- monkeylearn provides a wrapper interface to machine learning services on Monkeylearn for text analysis, i.e., classification and extraction.
- udpipe provides language-independant tokenization, part of speech tagging, lemmatization, dependency parsing, and training of treebank-based annotation models.

https://CRAN.R-project.org/view=NaturalLanguageProcessing

# 참고 자료:


저자: 백영민
제목: R를 이용한 텍스트 마이닝
출판사: 한울
추천 이유: 사회과학 연구자들에게 적합한 텍스트 마이닝 / 분석 기법을 친절히 설명하고 분석 절차에 필요한 사용자 함수 등을 모두 제시하여 활용하기 좋음.