텍스트마이닝

핀인사이트
데이터분석가 김현진

텍스트마이닝

텍스트 마이닝

텍스트 마이닝이란?

텍스트마이닝

텍스트 마이닝(Text Mining)이란?

언어학, 통계학, 기계 학습 등을 기반으로 한 자연언어 처리 기술을 활용하여 반정형 및 비정형 텍스트 데이터를 정형화하고, 특징을 추출하기 위한 기술과 추출된 특징으로부터 의미 있는 정보를 발견할 수 있도록 하는 기술

출처 : wiki.hash.kr 텍스트마이닝

즉, 비정형 데이터에서 분석 도구(Python, R)를 이용하여 새롭고 유용한 정보를 찾아내는 과정 또는 기술

- 비정형 텍스트 데이터로 부터 의미있는 패턴(특징)과 정보를 찾아낸다.

그림 출처 : https://towardsdatascience.com/organizing-your-first-text-analytics-project-ce350dea3a4a

텍스트마이닝

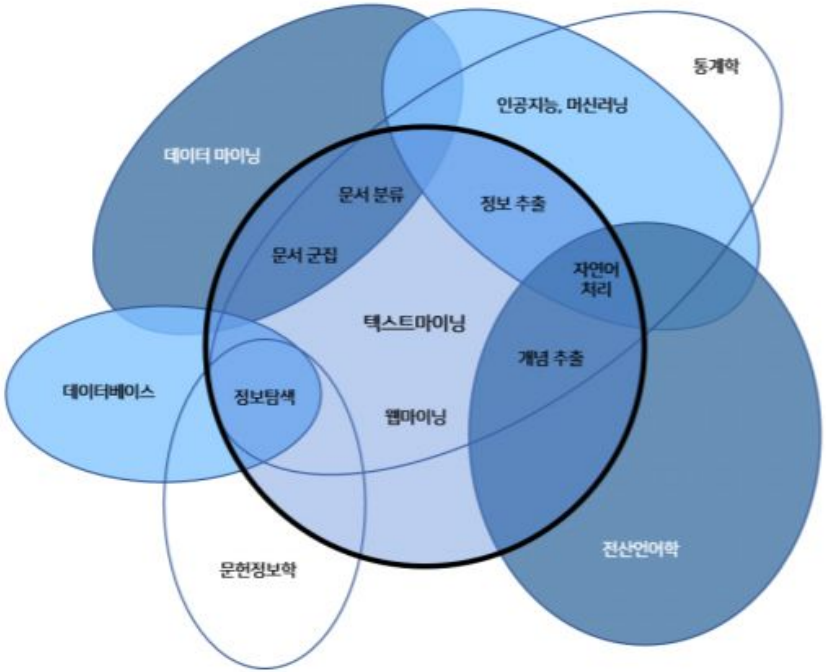


그림 출처 : 재정정보 분야 텍스트마이닝 활용 방안 연구(2019, 한국재정정보원)

텍스트 마이닝(Text Mining)이란?

- 이메일
- 고객의 리뷰
- 피드백
- 연구보고서
- 소셜네트워크 서비스 게시물 등의

텍스트 데이터를 통해 의사결정에 유용한 정보나

텍스트 패턴을 도출하는 과정으로,

인공지능, 통계학, 빅데이터 분석을 아우르는

여러분야가 융합된 분석 방법

데이터 마이닝 VS 텍스트 마이닝

- 데이터 마이닝

  - 정형 데이터에서 의미 있는 정보를 추출하는 기술

  - 고급 통계분석과 모델링 기법을 적용하여 데이터 안의 패턴과 관계를 찾아내는 과정

  - 데이터 마이닝의 전처리 과정에는 데이터 정제, 정규화, 병합

데이터 마이닝 VS 텍스트 마이닝

- 텍스트 마이닝

  - 텍스트 문서에서 의미있는 정보를 추출하는 기술

  - 비정형 텍스트 데이터를 정형화하고 특징을 추출하는 과정 필요

  - 컴퓨터가 인식해 처리하는 자연어 처리(NLP) 기술에 기반을 두고 데이터를 가공하는 기술

예시)

마침표, 대문자, 숫자, 특수문자, 띄어쓰기, 표 등 텍스트의 하위요소들로 규칙을 찾아낼 수 있으며,
글의 문단, 제목, 날짜, 저자 이름, 헤더, 각주 등의 하위요소들로 부터 규칙을 찾는다.

텍스트 마이닝 VS 데이터 마이닝

	텍스트 마이닝	데이터 마이닝
대상	텍스트	수치 또는 범주화된 데이터
구조	비정형 또는 정형의 텍스트 데이터	관계형 데이터 구조
목적	적합한 정보를 획득하고 의미를 정제하고 범주화함	미래 상황 결과의 예견 및 예측
방법	단어 빈도분석, 군집분석, 토픽모델링, 감성분석, 연관어 분석 등	기계학습

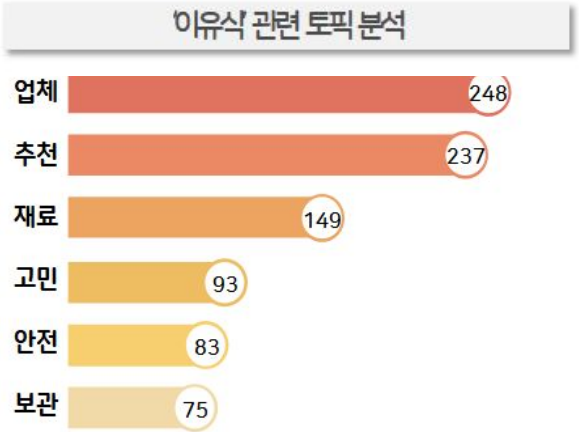
텍스트 마이닝 기법

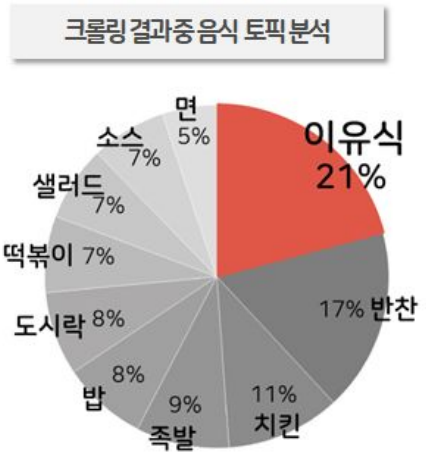
텍스트 마이닝 종류

단어 빈도분석

텍스트 데이터를 분석할 때 가장 보편적으로 활용되는 방법

- 데이터의 흐름을 파악하는 기초단계
- 수치 데이터를 분석할 때 평균, 분산, 표준편차, 최대값, 최소값을 살펴보는 것과 비슷함

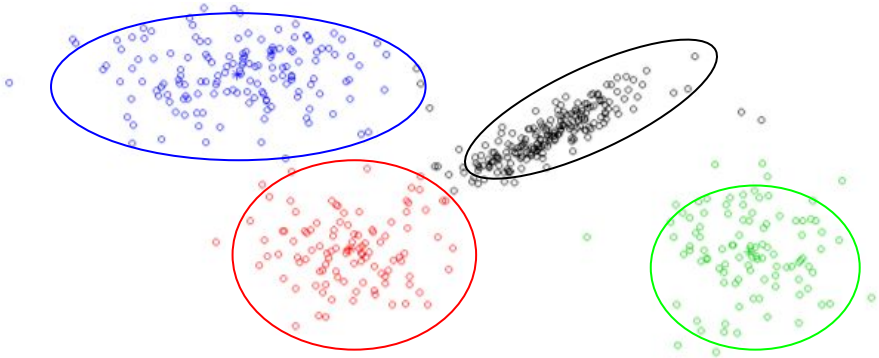


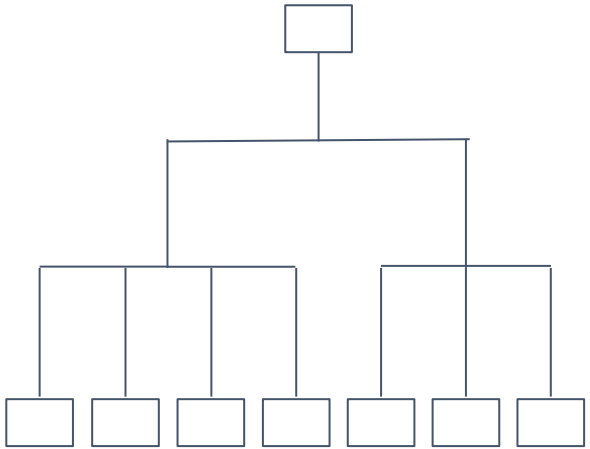


군집분석

유사한 데이터들을 서로 묶어주는 분석

- 대량의 텍스트 데이터들을 성격이 비슷한 데이터들끼리 묶어 줄 수 있다.
- 분할 군집 분석 : k-mean
- 구조적 군집 분석





토픽모델링

구조화되지 않은 방대한 문헌집단에서 주제를 찾아내기 위한 알고리즘

- 맥락과 관련된 단서들을 이용하여 의미를 가진 단어들을 클러스터링하여 주제를 추론함

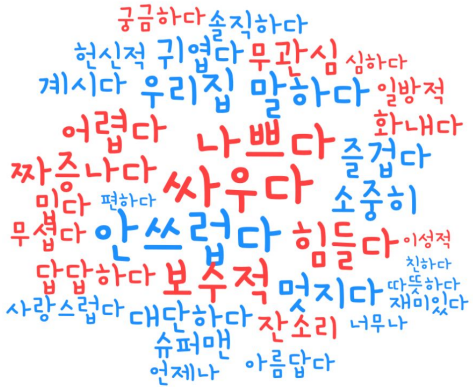
Q 어떤 주제 일까요?

참고 : BOK 경제연구 2019년 1월 논문 topic modeling

감정분석

텍스트에 나타난 주관성 요소를 탐지하여 긍정과 부정의 요소 및 그 정도를 판별하여 정량화하는 방법

- 문서 내 감성표현을 분석함으로써 이에 나타난 의견, 평가, 태도 등의 특징을 정량화된 자료로 제시한다.
- 텍스트 간의 비교우위를 밝힘으로써, 상대적 비교를 할 수 있다.



출처 : http://doc.mindscale.kr/blog/2016/1/25/introduction-to-sentiment-analysis

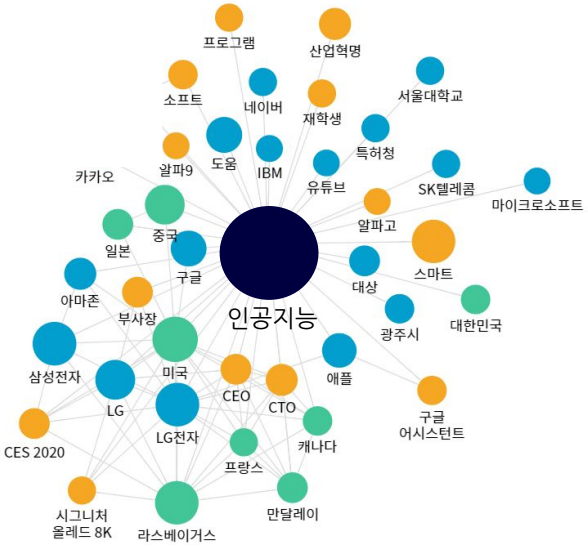
연관어 분석

두 개의 단어가 주어진 문맥(문서, 문단, 문장)에서 서로 얼마나 연관되어 있는지에 대한 분석

- 단어간의 연관도를 살펴보기 위한 분석
- how? 두 단어가 같은 문서에서 함께 출현하는 횟수를 센다.

  통계적 방법으로 유사도 산출

  딥러닝으로 유사도 산출(word2vec)



출처 : https://www.bigkinds

텍스트 마이닝 활용사례

텍스트 마이닝 활용 사례

리스크 관리(Risk Management)

어떤 업종이든지 간에 불충분한 리스크 분석은 실패의 원인으로 작용되어 왔다.

수 페타바이트에 이르는 문서들을 완벽하게 관리하고, 서로 관련 있는 정보들을 이어주며, 꼭 필요한 정보를 제때에 찾아주는 등 텍스트 마이닝을 기반으로 한 리스크 관리 소프트웨어를 사용했을 때 금융업과 같은 산업에서의 리스크 관리 능력이 극적으로 향상된다.

출처 : https://expertsystem.com/10-text-mining-examples/

지식 경영(Knowledge management)

방대한 양의 텍스트 문서에서 중요한 정보를 빨리 찾아내는 것은 항상 어려운 작업이다.

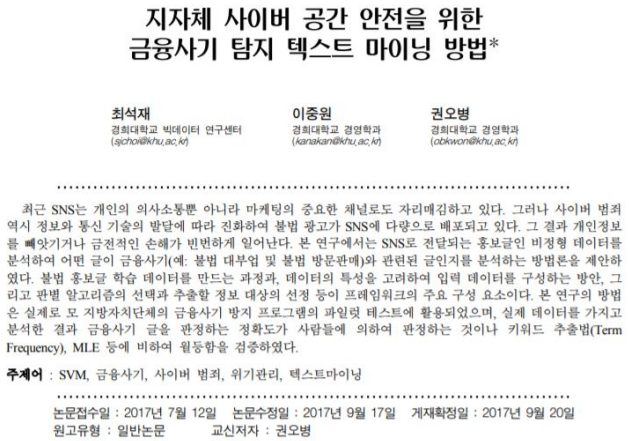
의료 업계와 같은 기관들은 많은 양의 정보를 관리해야 하는 어려움에 처해 있다. 하지만 텍스트 마이닝을 활용하면 수십 년 간의 유전자 연구 자료뿐 아니라 수 많은 임상 환자 데이터에 이르기까지, 장기적으로 이익의 중심점이 될 수 있는 신제품 개발에 도움을 줄 수 있는 데이터를 효율적으로 관리할 수 있다.

출처 : https://expertsystem.com/10-text-mining-examples/

사이버 범죄 예방(Cybercrime prevention)

인터넷이 지닌 익명성이라는 속성은 인터넷 기반 범죄의 위험을 가중시켜 왔다.

오늘날, 텍스트 마이닝을 이용한 범죄 예방 어플리케이션 등은 개인 및 기관을 대상으로 한 인터넷 범죄 예방에 도움을 주고 있다.



지자체 사이버 공간 안전을 위한
금융사기 탐지 텍스트 마이닝 방법*

최석재
경희대학교 빅데이터 연구센터
(sjchoi@khu.ac.kr)

이중원
경희대학교 경영학과
(kanakan@khu.ac.kr)

권오병
경희대학교 경영학과
(obkwon@khu.ac.kr)

최근 SNS는 개인의 의사소통뿐 아니라 마케팅의 중요한 채널로도 자리매김하고 있다. 그러나 사이버 범죄 역시 정보와 통신 기술의 발달에 따라 진화하여 불법 광고가 SNS에 다량으로 배포되고 있다. 그 결과 개인정보를 빼앗기거나 금전적인 손해가 빈번하게 일어난다. 본 연구에서는 SNS로 전달되는 홍보글인 비정형 데이터를 분석하여 어떤 글이 금융사기(예: 불법 대부업 및 불법 방문판매)와 관련된 글인지를 분석하는 방법론을 제안하였다. 불법 홍보글 학습 데이터를 만드는 과정과, 데이터의 특성을 고려하여 입력 데이터를 구성하는 방안, 그리고 판별 알고리즘의 선택과 추출할 정보 대상의 선정 등이 프레임워크의 주요 구성 요소이다. 본 연구의 방법은 실제로 모 지방자치단체의 금융사기 방지 프로그램의 파일럿 테스트에 활용되었으며, 실제 데이터를 가지고 분석한 결과 금융사기 글을 판정하는 정확도가 사람들에 의하여 판정하는 것이나 키워드 추출법(Term Frequency), MLE 등에 비하여 월등함을 검증하였다.

주제어 : SVM, 금융사기, 사이버 범죄, 위기관리, 텍스트마이닝

논문접수일 : 2017년 7월 12일 논문수정일 : 2017년 9월 17일 게재확정일 : 2017년 9월 20일
원고유형 : 일반논문 교신저자 : 권오병

출처 : https://expertsystem.com/10-text-mining-examples/

고객 관리 서비스(Customer care service)

자연어 처리 뿐 아니라 텍스트 마이닝은 고객 관리에 빈번하게 활용된다.

설문조사나 서비스 질 향상을 위한 고객 전화 문제해결의 효용성이나 속도 개선을 위해 활용된다. 설문조사나 전화응답 결과를 텍스트 분석을 통해, 고객에게 빠르고 자동화된 응답을 제공하기 위해 활용된다다. 이는 콜센터 직원에게 의지하던 일을 획기적으로 감소시켜 준다.

출처 : https://expertsystem.com/10-text-mining-examples/

고객 클레임 분석을 통한 부정행위 탐지 (Fraud detection through claims investigation)

보험 회사는 텍스트 분석과 정형 데이터를 결합해 사기를 방지하고 빠르게 클레임을 처리할 수 있습니다.

참고기사 : http://www.itworld.co.kr/news/99635

출처 : https://expertsystem.com/10-text-mining-examples/

콘텐츠 강화(Content enrichment)

텍스트로 이루어진 정보를 처리할 때 아직까지 인간의 노력을 필요로 하는 것이 사실이지만 텍스트 마이닝 기술은 방대한 양의 정보를 다룰 때에 특히 유용하다.

텍스트 마이닝 기술은 정보를 층층이 쌓으며 내용을 더욱 풍부하게 해주고, 다양한 목적에 따라 그에 적합한 내용으로 정리하고 요약하는데 활용이 가능하다.

출처 : https://expertsystem.com/10-text-mining-examples/

소셜 미디어 데이터 분석 (Social media data analysis)

오늘날 많은 기업들이 인지하고 있는 바와 같이 소셜 미디어는 비정형 데이터가 가장 활발히 생산되고 있는 원천 중 하나이다. 소셜 미디어는 시장 및 고객 정보를 파악하는데 있어서 점차 그 중요성이 높아지고 있다.

그러므로 텍스트 마이닝은 많은 양의 비정형 데이터를 분석함으로써 해당 브랜드나 제품에 대한 다양한 의견과 감성반응을 살펴볼 수 있다.

출처 : https://expertsystem.com/10-text-mining-examples/

텍스트 마이닝 프로세스

Text-mining Process

텍스트 마이닝 절차

데이터 수집 단계	텍스트 전처리 단계	텍스트 분석 단계	시각화 단계
데이터 수집 Data Collection	토큰화 Tokenizing	주제어 찾기 Topic modeling	
데이터 정제 Data Cleansing	품사 부착 POS Tagging	문서 요약 Summarize	
	원형 복원 Stemming, Lemmatizing	문서 분류 Category classification	
	불용어 처리 Stopword	감성 분석 Sentiment analysis	

자연어 처리 텍스트 분석 절차

Q. 내가 적접 인스타그램에서 “보헤미안 랩소디" 해시태그관련 정보를 수집해서 시장반응을 분석해야 한다면 어떻게 할 것인가?

인스타그램에서 “#보헤미안랩소디" 해시태그를 입력하고 포스트 검색결과를 수집	데이터 수집 단계
포스트 내용을 일관된 포맷으로 정리	텍스트 전처리 단계
각 포스트 내용을 읽어보고 핵심 키워드, 긍정/부정/중립을 판단	텍스트 분석 단계
정리한 내용을 보고서로 정리	시각화 단계

자연어 처리를 위한 텍스트 수집

자연어 처리 텍스트 분석 절차

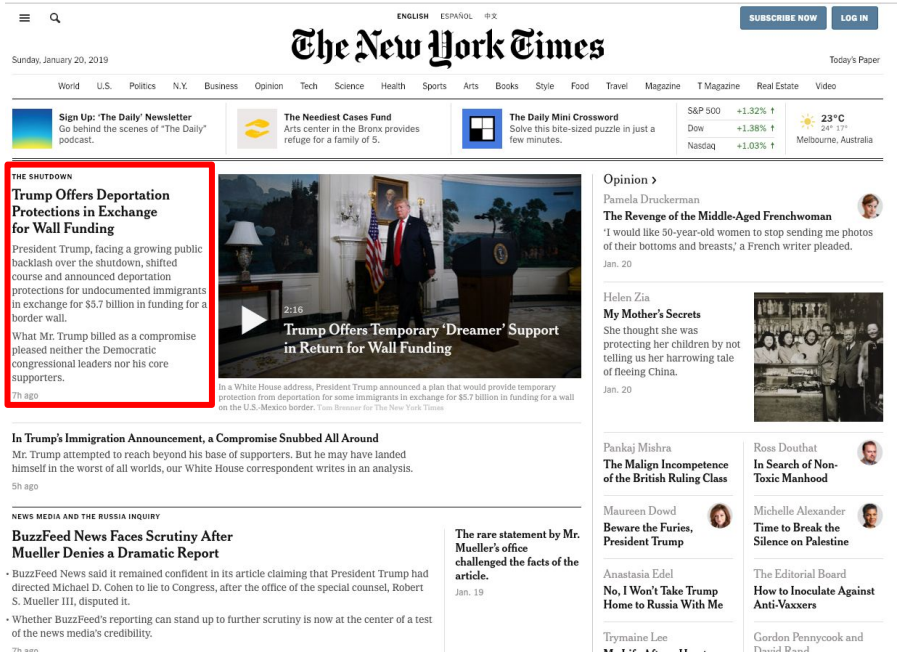
데이터 수집 (Data Collection)

필요한 데이터를 선별하고 수집하여 저장하는 것

- 현장에서 직접 입수
- 인터넷 검색
- 공개 API 활용
- 웹크롤링 (Web Crawling, Web Scraping)

데이터 수집 (Data Collection)

필요한 데이터를 선별하고 수집하여 저장하는 것



데이터 정제 (Data Cleasing)

데이터를 쉽게 사용할 수 있도록 불필요한 부분을 제거

자연어 처리를 위한 텍스트 전처리

자연어 처리 텍스트 분석 절차

데이터 수집 단계	텍스트 전처리 단계	텍스트 분석 단계	시각화 단계
데이터 수집 Data Collection	토큰화 Tokenizing	주제어 찾기 Topic modeling	
데이터 정제 Data Cleansing	품사 부착 POS Tagging	문서 요약 Summarize	
	원형 복원 Stemming, Lemmatizing	문서 분류 Category classification	
	불용어 처리 Stopword	감성 분석 Sentiment analysis	

텍스트 전처리 개요

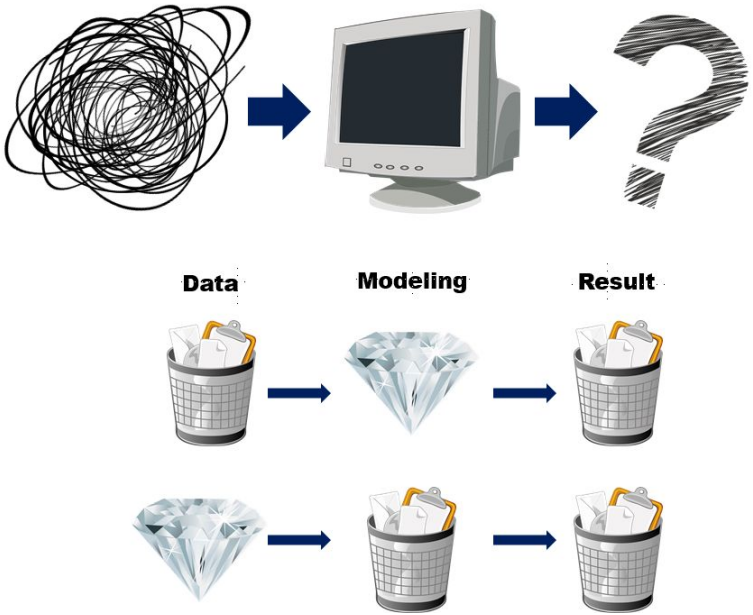
분석 하기 전 텍스트를 분석에 적합한 형태로 변화하는 작업을 **텍스트 전처리**라 한다.

전처리 단계는 텍스트를 토큰화하고 자연어 처리에 필요없는 조사, 특수문자, 단어(불용어)의 제거과정을 포함한다.

전처리는 분석결과와 모델 성능에 직접 영향을 미치기 때문에 전처리 단계는 매우 중요하다.

1. 토큰화	구두점으로 문서를 문장으로 분리하는 과정이다. 문장부호를 제거하거나 영어의 경우 대문자를 소문자로 변환하는 작업을 할 수 있다.
2. 형태소 분석	뜻을 가진 가장 작은 단위인 형태소로 문장을 분리하고 품사정보를 부착하는 작업을 뜻한다.
3. 개체명 인식	개별 개체로 인식되어야하는 단어를 구별하는 과정이다. 여기서 개별 개체로 인식되어야 하는 단어는 인물명, 지명, 약자 등을 의미한다.
4. 원형 복원	단어 기본 형태인 어간을 추출하는 과정이다. Stemming방식과 Lemmatizing방식이 있다.
5. 불용어 처리	분석에 불필요한 단어나 방해되는 단어를 제거하는 과정이다.

텍스트 전처리 개요



자연어 텍스트 전처리

1. 토큰화(Tokenization)
2. 형태소 분석- 품사부착(Pos-tagging)
3. 개체명인식(NER, Named Entity Recognition)
4. 원형복원 - 어간추출(Stemming)

   원형복원 - 표제어추출(Lemmatizing)
5. 불용어처리 (StopWord)

토큰화(Tokenization)

텍스트 전처리

토큰화(Tokenization)

토큰화란?

- 텍스트를 자연어 처리를 위해 분리하는 것
- 문서를 단어 또는 문장으로 분리하는 과정

1) 단어 토큰화(Word Tokenization)
2) 문장 토큰화(Sentence Tokenization)

토큰화(Tokenization)

단어 토큰화(Word Tokenization)

- 단어(word)를 기준으로 토큰화
- 영문의 경우 공백으로 분리하면 유의미한 토큰화 가능
- 한글의 경우 품사를 고려한 토큰화 필요

토큰화(Tokenization)

단어 토큰화 예시) 영문

> 형태소 분석기 : nltk - word_tokenize

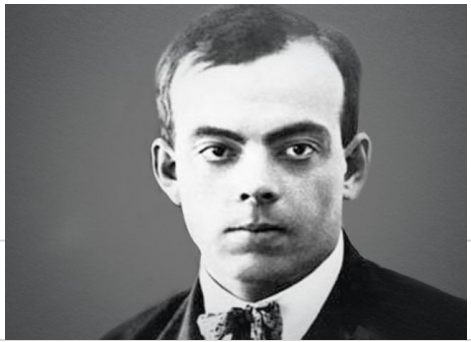
A day without friends is like honey without honey.

A	day	without	friends	is	like	honey	without	honey	.

토큰화(Tokenization)

단어 토큰화 예시) 한글

> 형태소 분석기 : Twitter 사용



계획없는 목표는 한낱 꿈에 불과하다.

계획	없는	목표	는	한낱	꿈	에	불과하다	.

단어 토큰화 고려사항

- 특수문자가 있는 경우

  > 구두점 및 특수문자를 단순하게 제외하면 안된다.

특수문자	원문	토큰화 예제 1	토큰화 예제 2
‘	Don’t	Do / n’t	Don / ’ / t
-	State-of-the-art	State / of / the / art	State-of-the-art

- 단어 내 띄어쓰기가 있는 경우

특수문자	원문	토큰화 예제 1	토큰화 예제 2
공백	New York	New / York	New York

토큰화(Tokenization)

문장 토큰화(Sentence Tokenization)

- 문장(sentence)을 기준으로 토큰화
- 온점(.), 느낌표(!), 물음표(?)로 분리하면 해결 될 것으로 생각됨
- 하지만, 단순하게 분리할 경우 정확한 분리가 어려움.

> Why ?

토큰화(Tokenization)

문장 토큰화 예시)

My name is Minho Lee. Just call me Mr.Lee

My name is Minho Lee.

Just call me Mr.Lee

> 만약 단순하게 점(.)으로 문장을 분리한다면?

Just call me Mr.

Lee

형태소 분석(Pos-Tagging)

텍스트 전처리

형태소 분석과 품사부착(Pos-tagging)

형태소 분석이란?

뜻을 가진 가장 작은 단위인 형태소로 문장을 분리하는 것

품사부착이란?

- 각 토큰에 품사 정보를 추가
- 분석시에 불필요한 품사(조사, 접속사)를 제거하거나 필요한 품사를 필터링하기 위해 사용

품사부착(Pos-tagging)

품사부착 예시) 영어

> 형태소 분석기 : nltk - word_tokenize

A day without friends is like honey without honey.

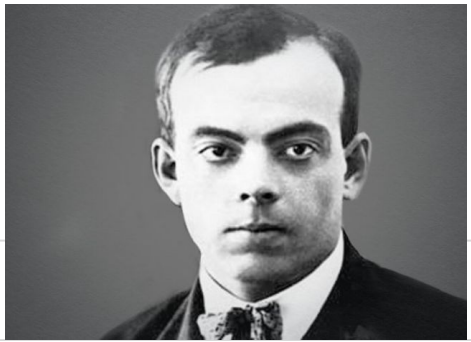
A/ DT	day/ NN	without/ IN	friends/ NNS	is/ VBZ	like/ IN	honey/ NN	without/ IN	honey/ NN	./ .

http://kkma.snu.ac.kr/documents/index.jsp?doc=postag

품사부착(Pos-tagging)

품사부착 예시) 한글

> 형태소 분석기 : Twitter 사용



계획없는 목표는 한낱 꿈에 불과하다.

계획/ Noun	없는/ Adjective	목표/ Noun	는/ Verb	한낱/ Adverb	꿈/ Noun	에/ Josa	불과하다/ Adjective	.

개체명인식(Named Entity Recognition)

텍스트 전처리

개체명인식(NER)

개체명인식이란?

- 사람, 조직, 지역, 날짜, 숫자 등의 개체 유형을 식별하는 것
- 텍스트가 무엇과 관련되어 있는지 구분하기위해 사용됨

ex) apple  

- 검색 엔진 색인(index)에 활용

검색을 빠르게 하기 위해 데이터를 저장하는 장소
즉, 특정 장소(문서)에 데이터를 저장하는 과정을 색인(인덱싱, indexing)이라 함.

개체명인식(NER)

청킹(chunking)이란?

- 자연어 처리 기법중의 하나로 정보를 의미 있는 단위로 묶어주는 기술

  예) ‘data mining’

개체명인식(NER)

개체명인식 예시)

> nltk의 ne_chunk 사용

Barack Obama likes fried chicken very much.

Barack / NNP / PERSON	Obama / NNP / ORGANIZATION	likes / VBZ	fried / VBN	chicken / JJ	very / RB	much / RB	.

원형복원

텍스트 전처리

원형복원 - Stemming

원형복원이란?

- 단어의 기본형태인 어간을 추출하는 과정

Stemming이란?

- 어간추출
- 각 토큰의 원형을 복원함으로써 토큰을 표준화하여 불필요한 데이터 중복을 방지한다.
- 단어의 수를 줄일 수 있어 연산의 효율이 높아진다.
- 품사를 무시하고 규칙에 기반해 어간을 추출한다.(https://tartarus.org/martin/PorterStemmer/def.txt)

원형복원 - Lemmatization

Lemmatization이란?

- 표제어추출
- 각 토큰의 원형을 복원함으로써 토큰을 표준화하여 불필요한 데이터 중복을 방지한다.
- 단어의 수를 줄일 수 있어 연산의 효율이 높아진다.
- 품사정보를 유지하여 표제어를 추출한다.

원형복원

Stemming & Lemmatization 비교

Stemming	Lemmatization
어간추출	표제어추출
규칙기반	사전기반
새로운 단어 처리 가능	사전에 없는 단어 처리 불가 (Out of vacabulary)

원형복원

원형 복원 예시)

		Stemming	Lemmatization
am	→	am	be
the listening	→	the listen	the listening
having	→	hav	have

불용어 처리(StopWord)

텍스트 전처리

불용어 처리(StopWord)

불용어 처리란?

- 분석에 불필요한 단어
- 방해되는 단어
- 자주사용되지만 특별한 의미가 없는 단어를 제거하는 과정
- 불필요한 토큰을 제거와 불필요한 품사를 제거함
- 예를들면, 영어에서 ‘the’ , ‘a’, ‘an’과 같은 관사는 많이 사용되지만 특별한 의미를 가지고 있지 않다.

불용어 처리(StopWord)

불용어 처리 예시)

> 형태소 분석기 : Twitter 사용

계획없는 목표는 한낱 꿈에 불과하다.

계획/ Noun | 없는/ Adjective | 목표/ Noun | 는/ Verb | 한낱/ Adverb | 꿈/ Noun | [illegible] | 불과하다/ Adjective | [illegible]

실습 1 - 영문전처리

NLTK(Natural Language Toolkit) 자연어처리 패키지를 사용하여
토큰화 / 품사테깅 / 개체명인식 을 해봅시다

실습 2 - 한글전처리

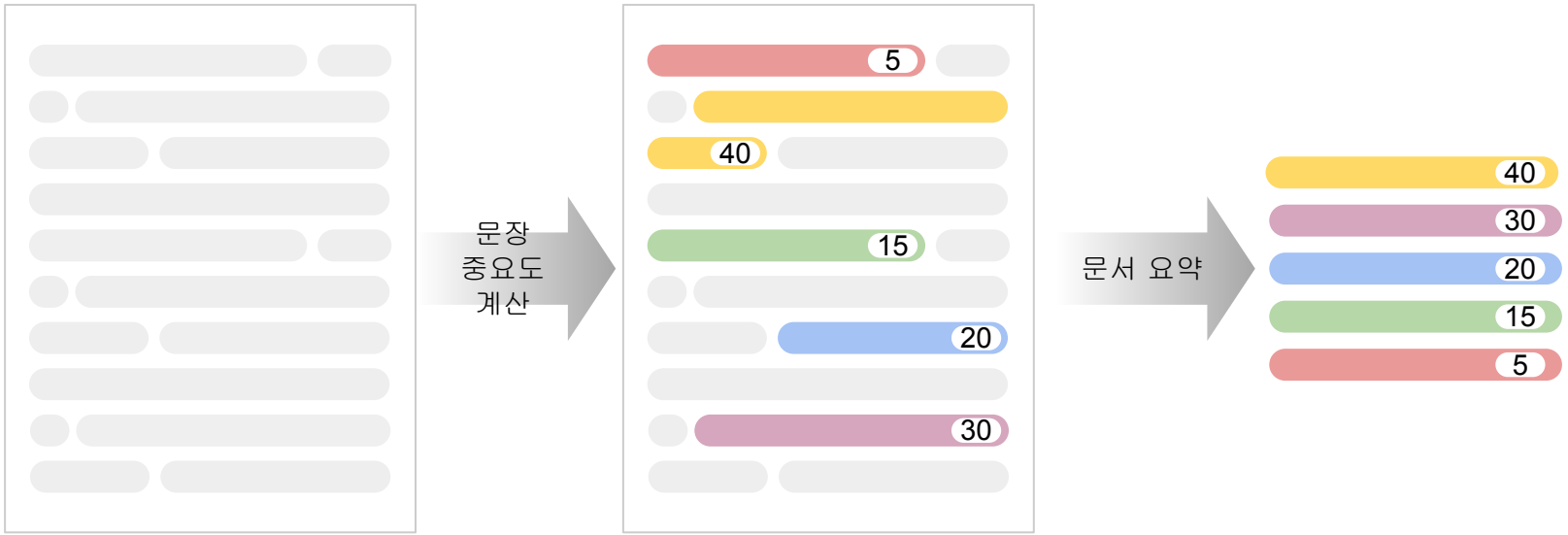
KoNLPy 한국어 자연어처리 패키지를 사용하여
토큰화 / 품사테깅 / 불용어처리 를 해봅시다

자연어 처리를 위한 텍스트 분석

텍스트 분석 단계

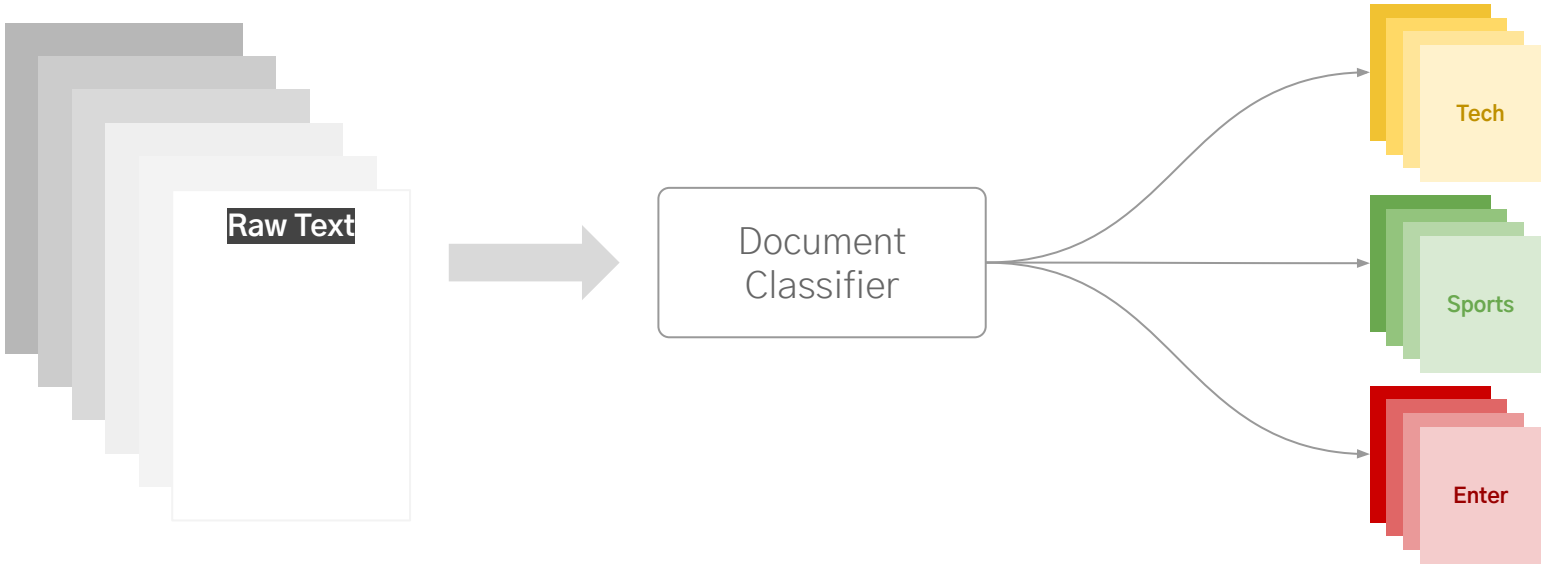
문서 요약 (Text Summarize)

문서 내에서 주요 문장을 찾아 요약



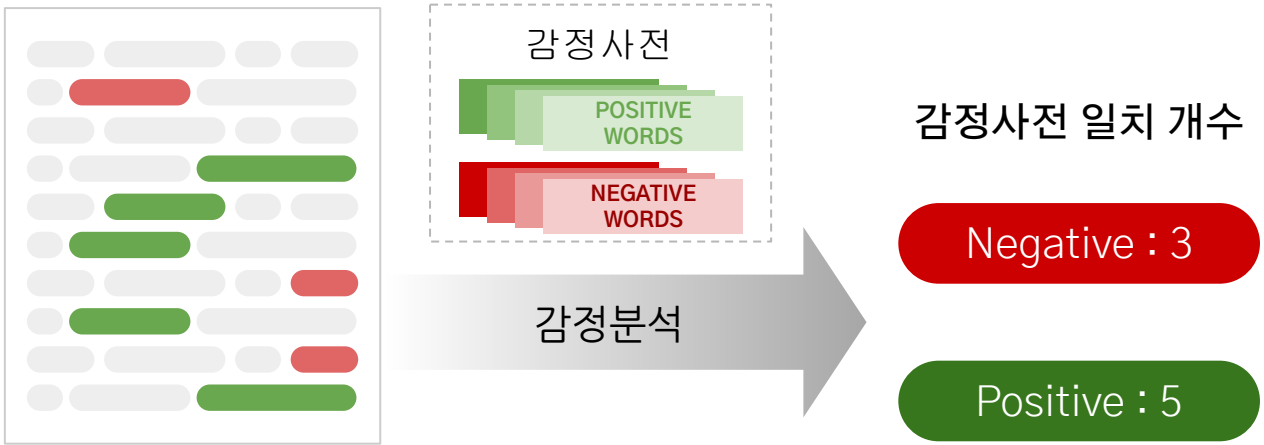
문서 분류 (Category Classification)

문서 내 단어 혹은 문장을 분석하여 문서를 분류



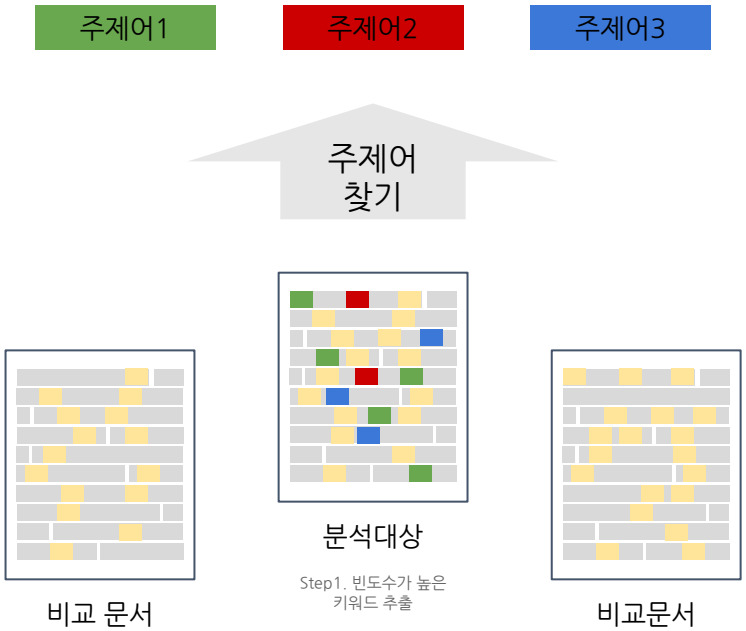
감성 분석 (Sentiment Analysis)

문서 내 나타난 사람들의 태도, 의견, 성향 같은 주관성을 분석



주제어 찾기 (Topic Modeling)

문서 내에서 주제를 발견하기 위한 모델



자연어 처리를 위한 텍스트 시각화

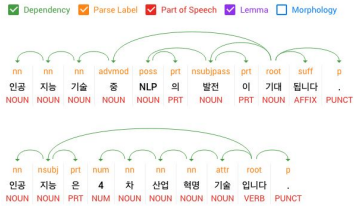
텍스트 전처리 단계

시각화

데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정

데이터 수집
Data Collection

텍스트 전처리
Pre-processing

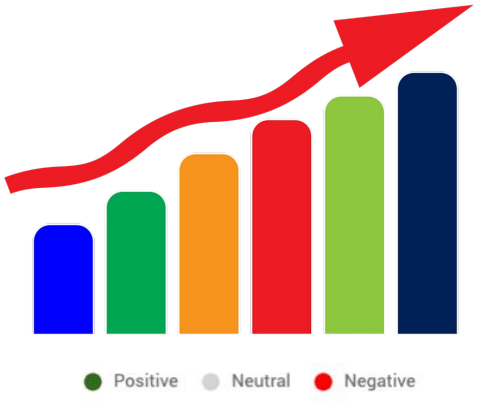


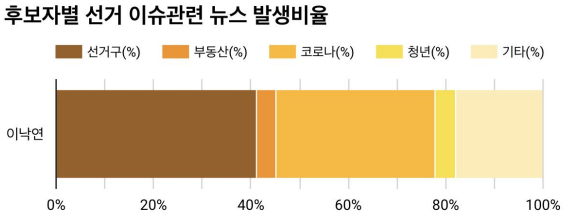
텍스트 분석
Text Analytics

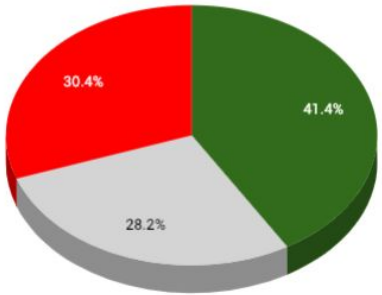
- 주제어 찾기
- 문서 요약
- 문서 분류
- 감성 분석

시각화
Visualization

시각화 예시







시각화 예시