

# Text Mining

텍스트 마이닝이란?

# 텍스트 마이닝

## 텍스트 마이닝(Text Mining)이란?

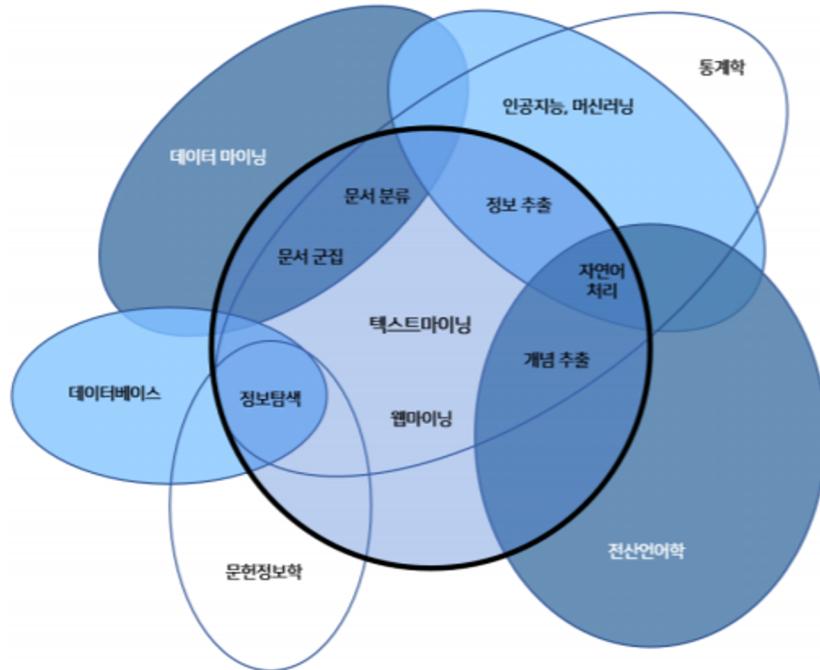
언어학, 통계학, 기계 학습 등을 기반으로 한 자연언어 처리 기술을 활용하여 반정형 및 비정형 텍스트 [데이터](#)를 정형화하고, 특징을 추출하기 위한 기술과 추출된 특징으로부터 의미 있는 정보를 발견할 수 있도록 하는 기술



출처 : wiki.hash.kr 텍스트마이닝

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

# 텍스트 마이닝



## 텍스트 마이닝(Text Mining)이란?

텍스트 데이터를 통해 의사결정에 유용한 정보나  
텍스트 패턴을 도출하는 과정으로,  
인공지능, 통계학, 빅데이터 분석을 아우르는  
여러분야가 융합된 분석 방법

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

# 데이터 마이닝 VS 텍스트 마이닝

---

## ● 데이터 마이닝

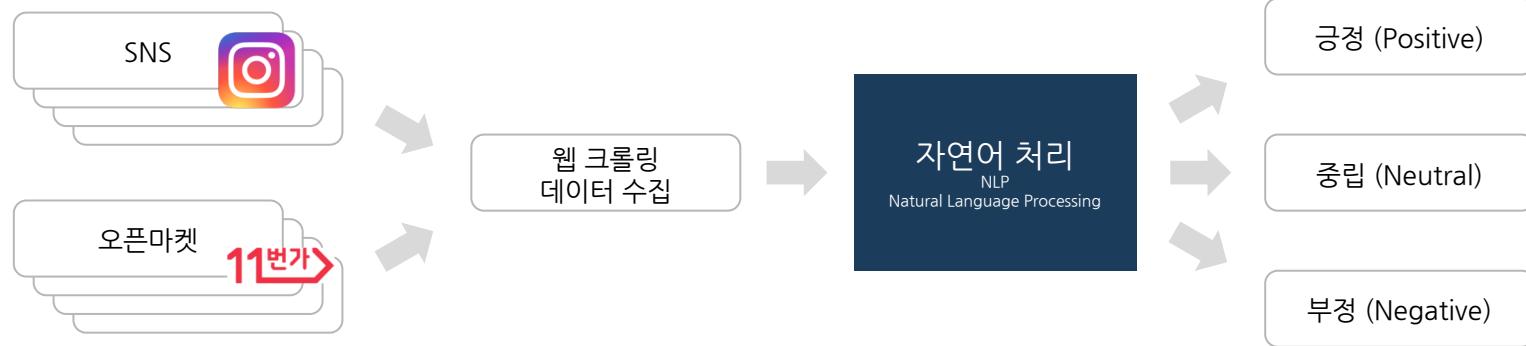
- 정형 데이터에서 의미 있는 정보를 추출하는 기술
- 고급 통계분석과 모델링 기법을 적용하여 데이터 안의 패턴과 관계를 찾아내는 과정
- 데이터 마이닝의 전처리 과정에는 데이터 정제, 정규화, 병합

# 텍스트 마이닝 활용 사례

자연어 처리 시작하기

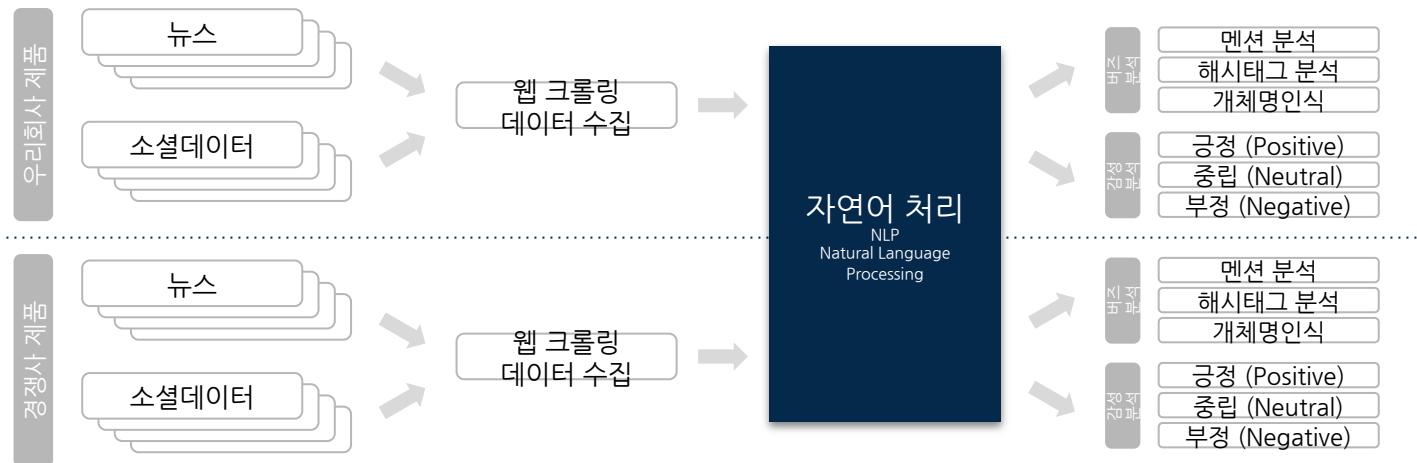
# 실시간 리뷰 모니터링

- 고객의 92%는 온라인 리뷰를 읽고, 86%는 5개의 별 중 3개 미만의 제품을 구매하지 않음
- 온라인 미디어(소셜, 오픈마켓 등)에서 제품관련 리뷰 정보를 수집하여 시장반응을 모니터링
- 대량의 정보를 취합하여 빠르게 시장반응을 파악하고 대응할 수 있음



# 경쟁사 분석

- 전략기획 담당자 혹은 마케터에게 전략을 수립함에 있어 경쟁사 분석은 필수
- 우리 회사와 경쟁 회사를 비교, 어떤 부분에 우위와 열세를 가지고 있는지 객관적으로 판단
- 우리 회사의 현황과 경쟁 회사 실시간 분석 가능



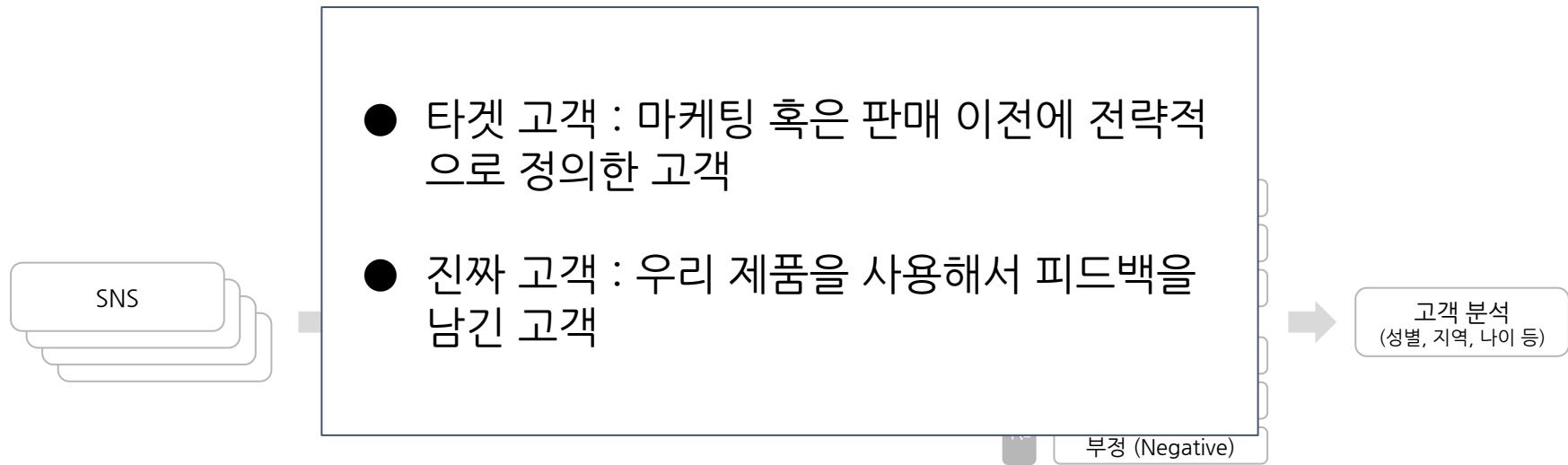
# 뉴스레터 메일링 / SNS 자동 포스팅

- 뉴스레터 담당자는 정기적으로 시장 동향을 정리해서 뉴스레터를 발송
- 산업 관련 키워드를 검색하여 결과를 취합, 정리하여 메일로 발송
- 동일한 방법을 소셜미디어에 자동 포스팅 가능



# 고객분석

- 우리의 진짜 고객은 누구일까요?
- 우리 제품에 대한 리뷰를 남긴 고객을 인구통계학적으로 분석해 본다면 추정이 가능



# 자연어처리 텍스트처리 프로세스

# 자연어 처리 활용 텍스트 데이터 분석 절차

# 자연어 처리 텍스트 분석 절차



# 자연어 처리 텍스트 분석 절차

Q. 내가 직접 인스타그램에서 “코로나 마스크”에 대한  
시장반응을 분석해야 한다면 어떻게 할 것인가?

인스타그램에서 “#마스크 #코로나 #코로나 마스크” 해시태그를 입력하고 포스트 검색결과를  
수집

데이터 수집 단계

포스트 내용을 일괄된 포맷으로 정리

텍스트 전처리 단계

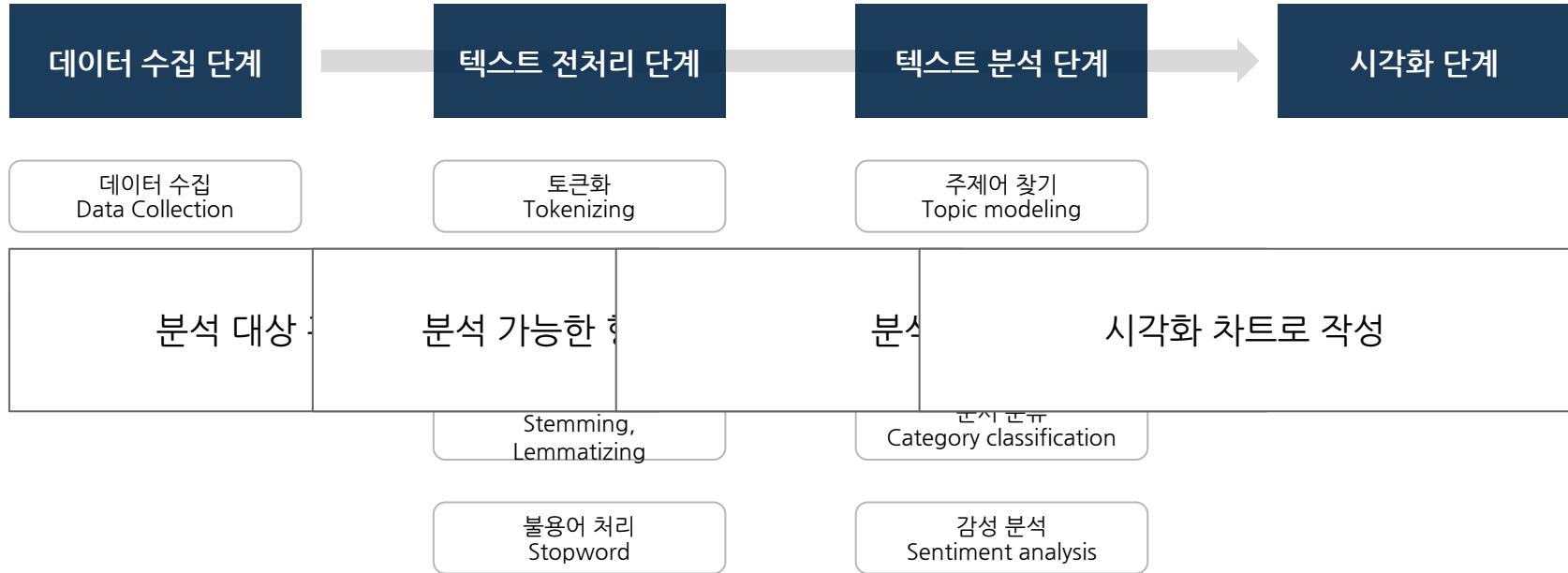
각 포스트 내용을 읽어보고 핵심 키워드, 긍정/부정/중립을 판단

텍스트 분석 단계

정리한 내용을 보고서로 정리

시각화 단계

# 자연어 처리 텍스트 분석 절차



# 데이터 수집 단계

자연어 처리 분석 절차

# 자연어 처리 텍스트 분석 절차



# 데이터 수집 (Data Collection)

필요한 데이터를 선별하고 수집하여 저장하는 것

The New York Times

Sunday, January 20, 2019

ENGLISH ESPAÑOL 中文

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Sign Up: 'The Daily' Newsletter Go behind the scenes of "The Daily" podcast.

The Neediest Cases Fund Arts center in the Bronx provides refuge for a family of 5.

The Daily Mini Crossword Solve this bite-sized puzzle in just a few minutes.

S&P 500 +1.32% ↑ Dow +1.38% ↑ Nasdaq +1.03% ↑ 23°C 24° 17° Melbourne, Australia

**THE SHUTDOWN**

**Trump Offers Deportation Protections in Exchange for Wall Funding**

President Trump, facing a growing public backlash over the shutdown, shifted course and announced deportation protections for undocumented immigrants in exchange for \$5.7 billion in funding for a border wall.

What Mr. Trump billed as a compromise pleased neither the Democratic congressional leaders nor his core supporters.

7h ago

**In Trump's Immigration Announcement, a Compromise Snubbed All Around**

Mr. Trump attempted to reach beyond his base of supporters. But he may have landed himself in the worst of all worlds, our White House correspondent writes in an analysis.

5h ago

**NEWS MEDIA AND THE RUSSIA INQUIRY**

**BuzzFeed News Faces Scrutiny After Mueller Denies a Dramatic Report**

BuzzFeed News said it remained confident in its article claiming that President Trump had directed Michael D. Cohen to lie to Congress, after the office of the special counsel, Robert S. Mueller III, disputed it.

Whether BuzzFeed's reporting can stand up to further scrutiny is now at the center of a test of the news media's credibility.

7h ago

**The rare statement by Mr. Mueller's office challenged the facts of the article.**

Anastasia Edel  
No, I Won't Take Trump Home to Russia With Me

Jan. 19

**Pankaj Mishra**  
**The Malign Incompetence of the British Ruling Class**

Maureen Dowd  
Beware the Furies, President Trump

Trymaine Lee

**Ross Douthat**  
**In Search of Non-Toxic Manhood**

Michelle Alexander  
Time to Break the Silence on Palestine

The Editorial Board  
How to Inoculate Against Anti-Vaxxers

Gordon Pennycook and David Rand

Opinion >

Pamela Druckerman  
**The Revenge of the Middle-Aged Frenchwoman**  
"I would like 50-year-old women to stop sending me photos of their bottoms and breasts," a French writer pleaded.

Jan. 20

Helen Zia  
**My Mother's Secrets**  
She thought she was protecting her children by not telling us her harrowing tale of fleeing China.

Jan. 20

**Trump Offers Temporary 'Dreamer' Support in Return for Wall Funding**

2:16

In a White House address, President Trump announced a plan that would provide temporary protection from deportation for some immigrants in exchange for \$5.7 billion in funding for a wall on the U.S.-Mexico border. Tom Brenner for The New York Times

# 데이터 정제 (Data Cleaning)

데이터를 쉽게 사용할 수 있도록 불필요한 부분을 제거

메뉴

▶  
기사  
검색

▶  
기사  
검색

불필요  
문구

사람은 흔적을 남기고…흔적은 기회를 낳는다

당신의 흔적에 기회가 있다

필 사이먼 지음 / 장명재·이유진 옮김 / 한국경제신문 / 380쪽 / 1만8000원



폭풍우가 다가온다는 기상 예보를 접했을 때 사람들은 어떤 물건을 살까. 배터리, 동조림 제품, 생수 등 구호 물품이 잘 팔릴 것이라는 점은 쉽게 예상할 수 있다. 하지만 그뿐이 아니다. 임마트는 2004년 허리케인과 폭풍우 예보에 앞서 회사의 과거 데이터를 분석한 결과 딸기맛 칼라트르초(과자의 일종)가 평상시 판매량보다 7배나 더 많이 판매됐다는 사실을 발견했다. 허리케인이 다가오는 시장에서 가장 많이 팔린 것은 애주었다.

직관은 영감을 가져다주지만 이에 기반한 결정이 항상 옳은 것은 아니다.

당신의 흔적에

실시간 인기 기사

- 1. 마이클이 가수 김해인 뇌강질 친고양고 남편에게
- 2. 이성이 미안해하는 이유는? “여기저기”
- 3. 수소경제, 차세대 에너지로 손수소차니…
- 4. [정교노미] 정부가 추구하는 경쟁의 세기, 뭐야?
- 5. 페인드 페트리 투자에 아름다운 한국인 얼굴은…
- 6. [PlutoTV] 350 Coupon
- 7. Webhouse News API Tool

Look forward to a secure investment with DHA.

[Find out more](#)

Take advantage of our secure investment opportunities with DHA.

DHA  
Defence Housing Australia  
Look forward

FIN INSIGHT  
Copyright FIN INSIGHT. All Right Reserved

가치를 높이는 금융 인공지능 실무교육  
**Insight** campus

# 텍스트 전처리 단계

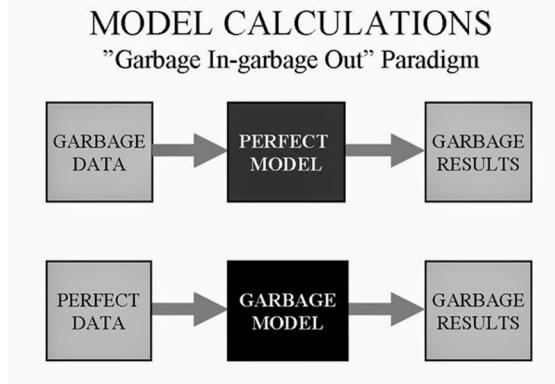
자연어 처리 분석 절차

# 자연어 처리 텍스트 분석 절차



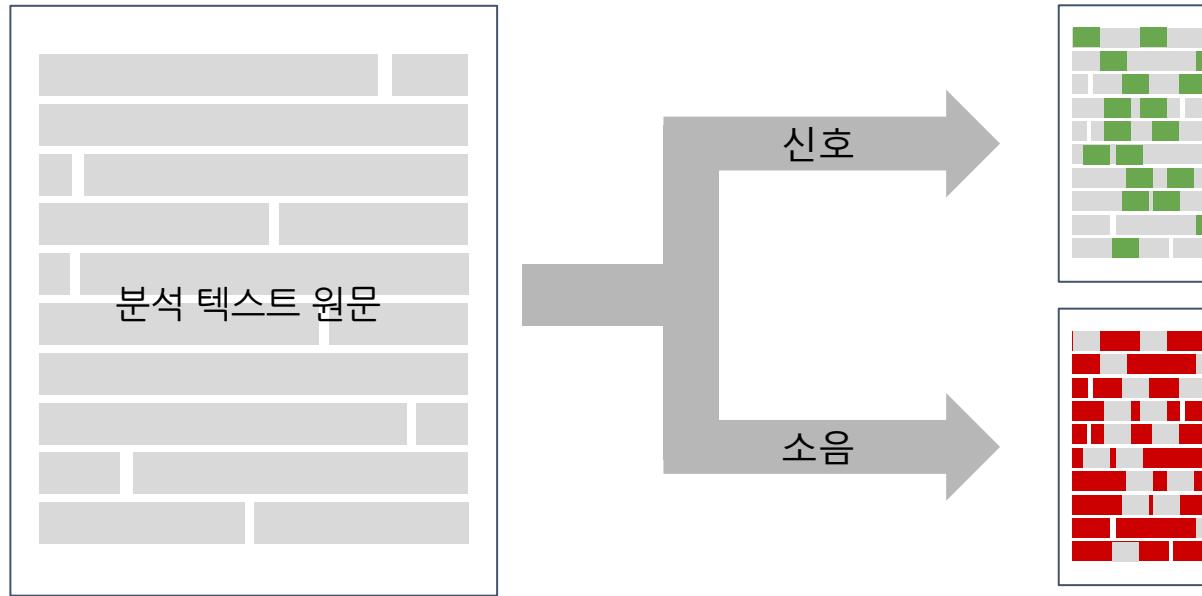
# 텍스트 전처리 단계

“쓰레기를 넣으면 쓰레기가 나온다  
(garbage in, garbage out)”

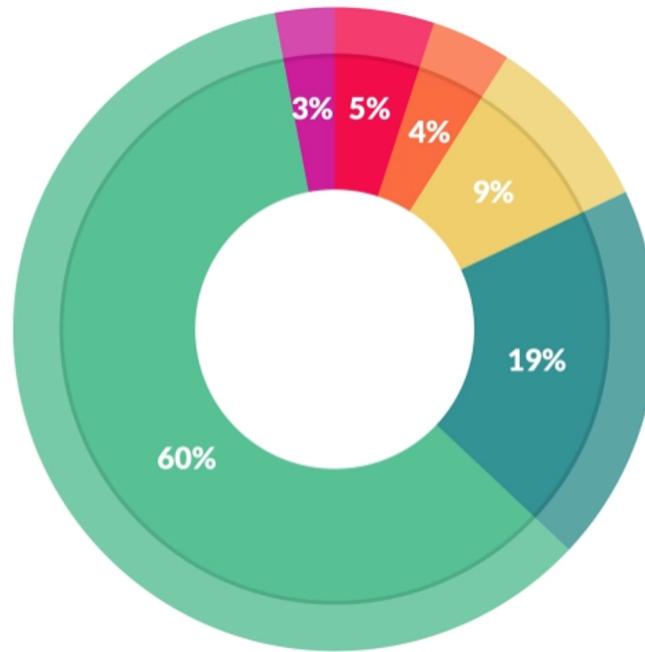


# 텍스트 전처리 단계

텍스트 분석을 위해서 기계가 텍스트를 이해할 수 있도록 표준화하는 단계



# 전처리 중요성



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

**79%**

# 토큰화 (Tokenizing)

---

자연어 처리가 가능한 단위(토큰)로 분리하는 방법

- 문장 토큰화 (sentence tokenization)
- 단어 토큰화 (word tokenization)
- 형태소 분석

# 문장 토큰화 (Sentence Tokenization)

---

- 문장(Sentence)를 기준으로 토큰화
- 마침표(.), 느낌표(!), 물음표(?) 등으로 분류하면 해결 될 것으로 생각됨
- 하지만 단순하게 분리할 경우 정확한 분리가 어려움

My name is Kyungsoo Hong. Just call me Mr.Hong

My name is Kyungsoo Hong.

Just call me Mr.Hong

# 단어 토큰화(Word Tokenization)

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

영문 토큰화

Barack Obama likes fried chicken very much.

Barack    Obama    likes    fried    chicken    very    much    .

한글 토큰화

버락 오바마는 후라이드 치킨을 너무 좋아한다

버락    오바마    는    후라이드    치킨    을    너무    좋아한다    .

# 형태소 분석 (Morphological Analysis)

문장을 형태소로 분리하는 작업

**형태-소** 形態素

+ 단어장 저장

표준국어대사전 고려대한국어대사전 우리말샘 < >

예문 열기 ▾

명사

- 언어 뜻을 가진 가장 작은 말의 단위. '이야기책'의 '이야기', '책<sup>1</sup>' 따위이다.
- 언어 문법적 또는 관계적인 뜻만을 나타내는 단어나 단어 성분. 프랑스의 언어학자 마르티네(Martinet, A.)가 제시하였다.  
=형태질.

# 품사 태깅 (POS Tagging)

명사

Noun

```
>>> from konlpy.tag import Mecab
>>> mecab = Mecab()
>>> print(mecab.morphs(u'영등포구청역에 있는 맛집 좀 알려주세요.'))
['영등포구', '청역', '에', '있', '는', '맛집', '좀', '알려', '주', '세요', '.']
>>> print(mecab.nouns(u'우리나라에는 무릎 치료를 잘하는 정형외과가 없는가!'))
['우리', '나라', '무릎', '치료', '정형외과']
>>> print(mecab.pos(u'자연주의 쇼핑몰은 어떤 곳인가?'))
```

[('자연', 'NNG'), ('주', 'NNG'), ('의', 'JKG'), ('쇼핑몰', 'NNG'), ('은', 'JX'), ('어떤', 'MM'), ('곳', 'NNG'), ('인가', 'VCP+EF'), ('?', 'SF')]

동사

Interjection

형용사

# 개체명 인식 (NER)

사람, 조직, 지역, 날짜, 숫자 등의 개체 유형을 식별하는 것

- 텍스트가 무엇과 관련되어 있는지 구분하기 위해 사용됨

ex) apple vs Apple



- 청킹 (Chunking) : 정보를 의미 있는 단위로 묶어주는 기술

ex) President Barack Obama, ‘금융 통화 위원회’

# 원형 복원 (Stemming, Lemmatization)

---

분리한 토큰을 표준화 하는 작업

- 어간 추출 (stemming) - 어간 : 활용시 변하지 않는 부분

ex) ‘먹고’, ‘먹는’, ‘먹지’, ‘먹을’, ‘먹은’, ‘먹어’, ‘먹었’ ⇒ ‘먹’

copy ⇒ copi

- 표제어 추출 (Lemmatization) : - 표제어 : 사전에 등재된 단어

ex) ‘나무들’ ⇒ 나무

playing ⇒ play

# 원형복원

원형 복원 예시)

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

# 불용어 처리 (Stopwords)

---

의미가 없는 단어 토큰을 제거하는 작업

- `from nltk.corpus import stopwords`  
ex) 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your'
- <https://www.ranks.nl/stopwords/korean>  
ex) '아', '휴', '아이구', '아이쿠', '아이고', '어', '나', '우리', '저희'
- 목적에 맞게 설정

# 텍스트 분석

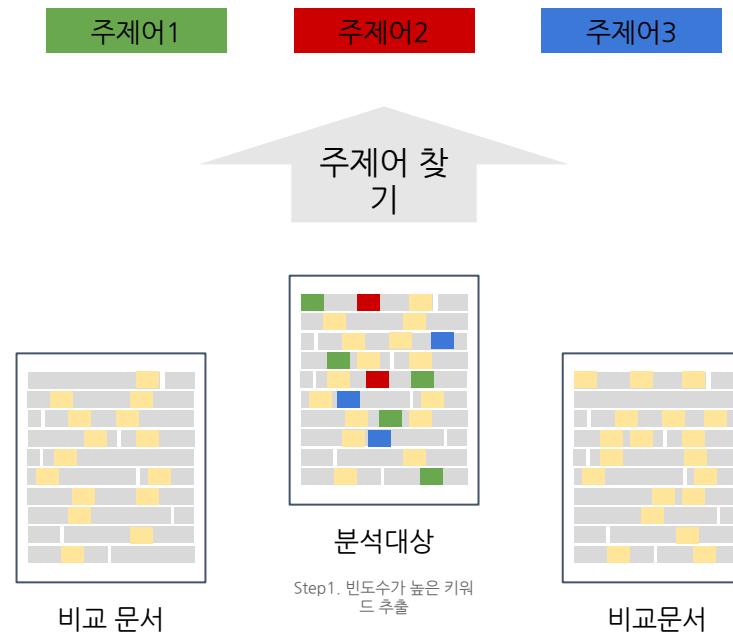
자연어 처리 분석 절차

# 텍스트 분석 단계



# 주제어 찾기 (Topic Modeling)

문서 내에서 주제를 발견하기 위한 모델



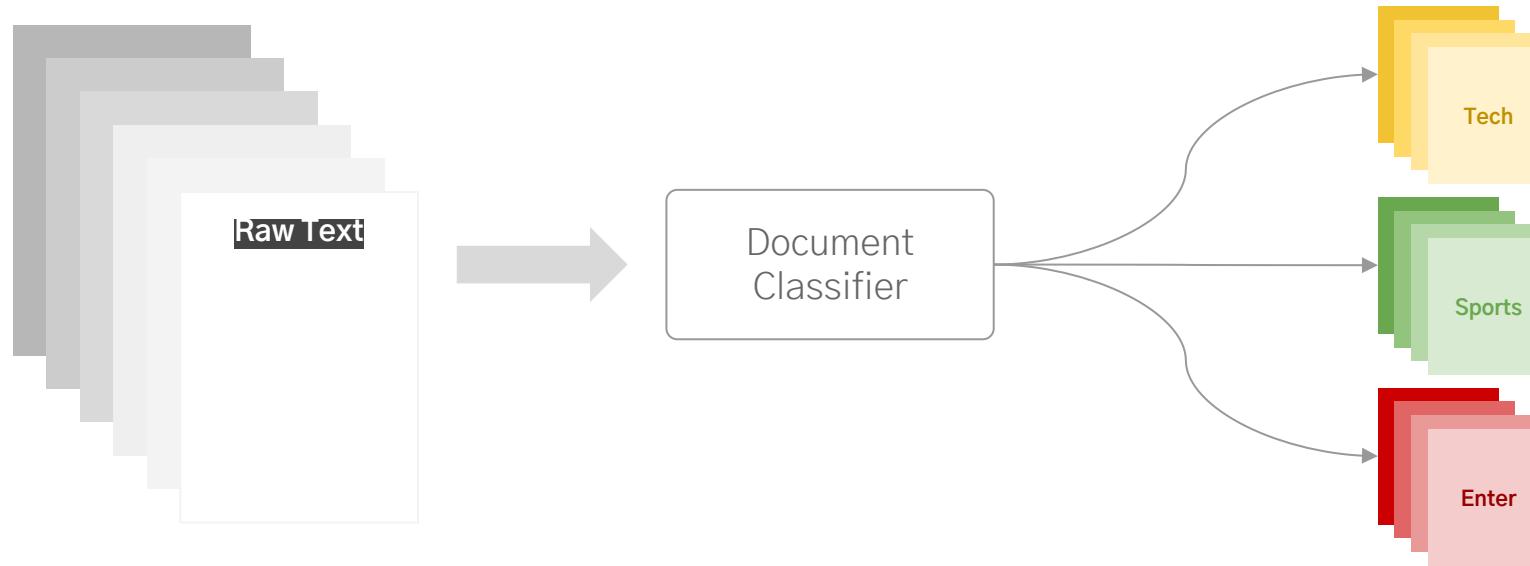
# 문서 요약 (Text Summarize)

문서 내에서 주요 문장을 찾아 요약



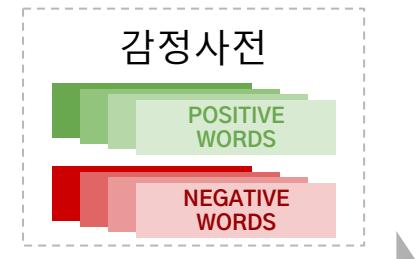
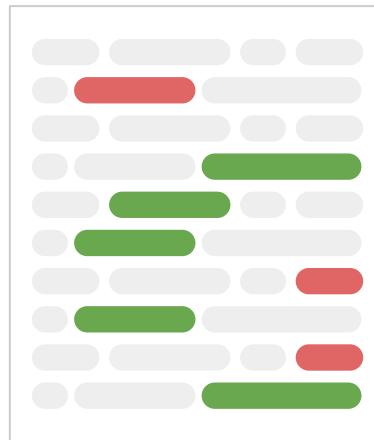
# 문서 분류 (Category Classification)

문서 내 단어 혹은 문장을 분석하여 문서를 분류



# 감성 분석 (Sentiment Analysis)

문서 내 나타난 사람들의 태도, 의견, 성향 같은 주관성을 분석



감정분석

감정사전 일치 개수

Negative : 3

Positive : 5

# 시각화

자연어 처리 분석 절차

# 텍스트 전처리 단계



# 시각화

데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정



## 시각화 예시

