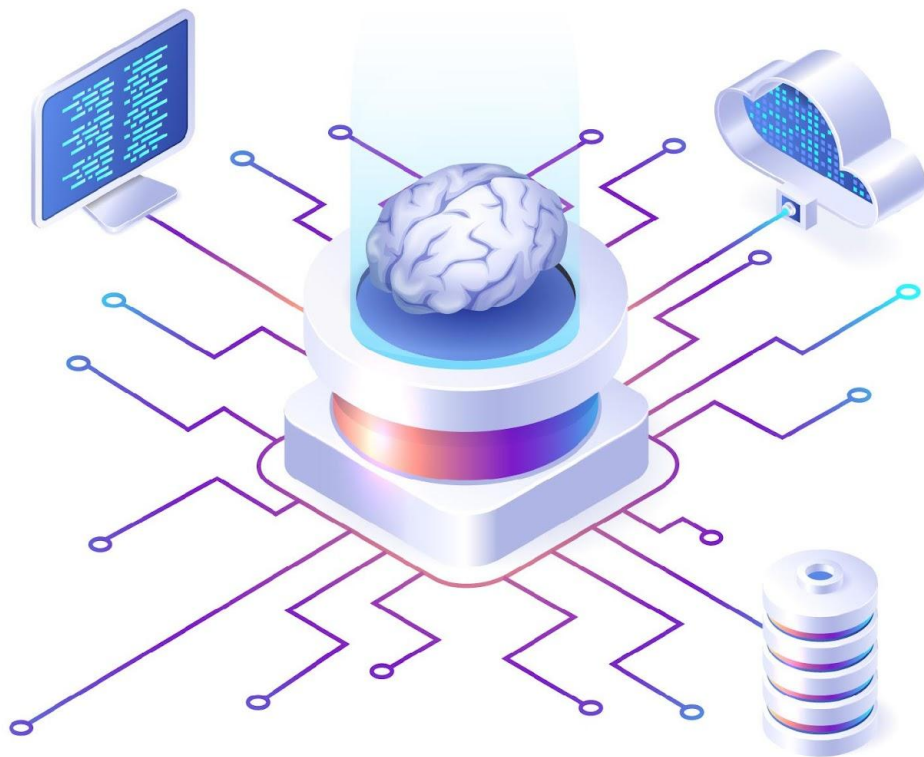


자연어 처리란?

실무형 인공지능 자연어처리



자연어 처리란?

자연어 처리 텍스트 마이닝 Overview

1

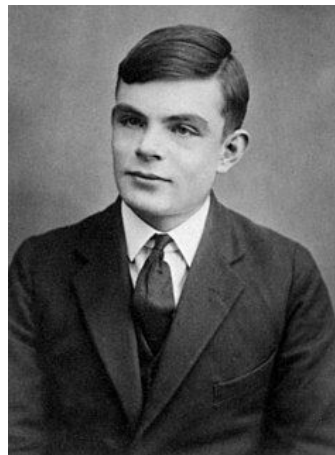
자연어 처리란?



인공지능(AI)의 시작

자연어 처리에 대한 관심은 1950년 앨런 튜링(Alan Turing)이 이른바 튜링 테스트(Turing Test)가 등장한 "Computing Machinery and Intelligence"라는 논문을 발표하면서 본격적으로 시작되었다.

“ 인간이 컴퓨터와 대화하고 있다는 것을 깨닫지 못하고
인간과 대화를 계속할 수 있다면
컴퓨터는 지능적(Intelligence)인 것으로 간주될 수 있다.”
- 앨런 튜링 -



앨런 매티슨 튜링은 영국의 수학자, 암호학자, 논리학자이자 컴퓨터 과학의 선구적 인물이다. 알고리즘과 계산 개념을 튜링 기계라는 추상 모델을 통해 형식화함으로써 컴퓨터 과학의 발전에 지대한 공헌을 했다. 튜링 테스트의 고안으로도 유명하다.

자연어 처리(NLP)란?

자연어 처리란?

자연어 처리(自然語處理) 또는 자연 언어 처리(自然言語處理)는 인간의 언어 현상을 컴퓨터와 같은 기계를 이용해서 모사할 수 있도록 연구하고 이를 구현하는 인공지능의 주요 분야 중 하나다.

정보 검색, QA 시스템, 문서 자동 분류, 신문기사 클러스터링, 대화형 Agent 등 다양한 응용이 이루어지고 있다.



WIKIPEDIA
The Free Encyclopedia

https://ko.wikipedia.org/wiki/%EC%9E%90%EC%97%B0%EC%96%B4_%EC%B2%98%EB%A6%AC

자연어 처리(NLP)란?

전통적인 프로그래밍 언어

: 기계(혹은 컴퓨터)를 실행하기 위해서 기계가 이해할 수 있는 프로그래밍 언어로 명령을 내리고, 그 결과를 사용자에게 전달

프로그래밍 언어
(Programming Language)

기계
(Computing Machine)

사용자
(User)



```

10 def __init__(self):
11     self.folder_path = folder_path
12     self.folder_path = folder_path
13     self.folder_path = folder_path
14     self.folder_path = folder_path
15     self.folder_path = folder_path
16     self.folder_path = folder_path
17     self.folder_path = folder_path
18     self.folder_path = folder_path
19     self.folder_path = folder_path
20     self.folder_path = folder_path
21     self.folder_path = folder_path
22     self.folder_path = folder_path
23     self.folder_path = folder_path
24     self.folder_path = folder_path
25     self.folder_path = folder_path
26     self.folder_path = folder_path
27     self.folder_path = folder_path
28     self.folder_path = folder_path
29     self.folder_path = folder_path
30     self.folder_path = folder_path
31     self.folder_path = folder_path
32     self.folder_path = folder_path
33     self.folder_path = folder_path
34     self.folder_path = folder_path
35     self.folder_path = folder_path
36     self.folder_path = folder_path
37     self.folder_path = folder_path
38     self.folder_path = folder_path
39     self.folder_path = folder_path
40     self.folder_path = folder_path
41     self.folder_path = folder_path
42     self.folder_path = folder_path
43     self.folder_path = folder_path
44     self.folder_path = folder_path
45     self.folder_path = folder_path
46     self.folder_path = folder_path
47     self.folder_path = folder_path
48     self.folder_path = folder_path
49     self.folder_path = folder_path
50     self.folder_path = folder_path
51     self.folder_path = folder_path
52     self.folder_path = folder_path
53     self.folder_path = folder_path
54     self.folder_path = folder_path
55     self.folder_path = folder_path
56     self.folder_path = folder_path
57     self.folder_path = folder_path
58     self.folder_path = folder_path
59     self.folder_path = folder_path
60     self.folder_path = folder_path
61     self.folder_path = folder_path
62     self.folder_path = folder_path
63     self.folder_path = folder_path
64     self.folder_path = folder_path
65     self.folder_path = folder_path
66     self.folder_path = folder_path
67     self.folder_path = folder_path
68     self.folder_path = folder_path
69     self.folder_path = folder_path
70     self.folder_path = folder_path
71     self.folder_path = folder_path
72     self.folder_path = folder_path
73     self.folder_path = folder_path
74     self.folder_path = folder_path
75     self.folder_path = folder_path
76     self.folder_path = folder_path
77     self.folder_path = folder_path
78     self.folder_path = folder_path
79     self.folder_path = folder_path
80     self.folder_path = folder_path
81     self.folder_path = folder_path
82     self.folder_path = folder_path
83     self.folder_path = folder_path
84     self.folder_path = folder_path
85     self.folder_path = folder_path
86     self.folder_path = folder_path
87     self.folder_path = folder_path
88     self.folder_path = folder_path
89     self.folder_path = folder_path
90     self.folder_path = folder_path
91     self.folder_path = folder_path
92     self.folder_path = folder_path
93     self.folder_path = folder_path
94     self.folder_path = folder_path
95     self.folder_path = folder_path
96     self.folder_path = folder_path
97     self.folder_path = folder_path
98     self.folder_path = folder_path
99     self.folder_path = folder_path
100    self.folder_path = folder_path

```

0100 1111 0101 0100 1111 0101 0100 1111 0101
 0000 1111 0101 0000 0100 0101 0000 0100 0000
 0100 1111 0100 1111 0101 0100 1111 0101 0100
 1111 0101 0000 1111 0101 0000 0100 0101 0000
 0100 0000 0100 1111 0100 0000 0100 1111 1111



자연어 처리(NLP)란?

자연어 처리

: 인간의 언어(=자연 언어)로 명령을 내리면 기계가 자연어 처리(NLP)를 통해 이해하여 처리하고, 그 결과를 사용자에게 전달



자연어 처리란?

자연어 처리(NLP)란?

자연어 처리란?

전통적인 프로그래밍 언어가 인간이 기계 언어로 기계(=컴퓨터)를 이해시키는 것이었다면,

자연어 처리는 기계가 인간의 언어(=자연 언어)를 이해하여 소통하는 것을 말한다.

2

자연어 처리, 왜 관심 가져야 하나



비정형 데이터의 중요성

- 인터넷과 모바일의 발달로 온라인 매체에 대한 데이터가 급격하게 증가
- 전 세계에서 생성되는 데이터 70~80%가 비정형 데이터(뉴스, SNS, 블로그, 기타 문서 등)
- 의사 결정을 내림에 있어 비정형 데이터 분석은 필수적

정형 데이터
(Structured Data)
사전 정의된 모델을 통해
구조화된 데이터
예시 : 엑셀, RDMS



비정형 데이터
(Unstructured Data)
내부 구조를 갖지만 미리
정의된 데이터 모델을 통해
구조화되지 않음.
예시 : 텍스트파일,
전자메일, 소셜미디어,
웹사이트

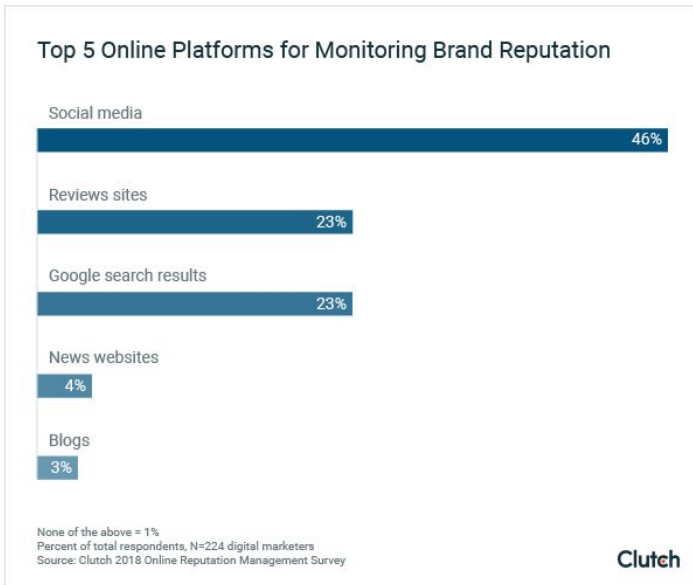


Company Info		KEY FINANCIALS			
Company Name	Country	Revenue (\$billions)	Profit (\$billions)	Assets (\$billions)	Market Value As of 10/10/18 (\$b)
1. ICBC	China	\$185,300	\$45,700.0	\$4,210,000	\$111,000
2. China Construction Bank	China	\$143,200	\$27,200.0	\$1,831,000	\$281,200
3. JPMorgan Chase	United States	\$118,200	\$28,000.0	\$2,409,000	\$287,700
4. Bank of China	China	\$105,000	\$24,700.0	\$1,702,700	\$491,000
5. Agricultural Bank of China	China	\$103,000	\$24,000.0	\$1,436,000	\$184,100
6. Bank of America	United States	\$100,000	\$20,000.0	\$2,528,000	\$173,000
7. Wells Fargo	United States	\$100,100	\$21,700.0	\$1,915,400	\$280,000
8. Apple	United States	\$247,000	\$65,000.0	\$387,000	\$620,000
9. Bank of China	China	\$118,200	\$28,000.0	\$1,204,200	\$108,000
10. Ping An Insurance Group	China	\$141,000	\$15,000.0	\$1,000,400	\$108,000
11. Royal Dutch Shell	Netherlands	\$121,000	\$15,200.0	\$470,700	\$100,000
12. Toyota Motor	Japan	\$280,200	\$22,000.0	\$475,000	\$201,700
13. ExxonMobil	United States	\$270,100	\$20,400.0	\$248,000	\$204,100
14. Samsung Electronics	South Korea	\$224,000	\$41,000.0	\$201,200	\$120,000
15. AIG	United States	\$100,200	\$10,000.0	\$440,000	\$100,000
16. Volkswagen Group	Germany	\$270,000	\$15,100.0	\$501,000	\$101,000

온라인 데이터의 중요성

- 포브스(Forbes)지에 따르면 “97%의 기업이 온라인 평판 관리(ORM, Online Reputation Management)가 매우 중요하다”
- 온라인 평판은 비정형 데이터(뉴스, SNS, 블로그 등)를 분석하여 평가 가능
- 분석 대상과 관련된 비정형 데이터를 수집하고 자연어 처리를 통해서 문서 내 인사이트 도출 가능

예) 제품에 대한 시장의 반응 (긍정, 부정, 중립)



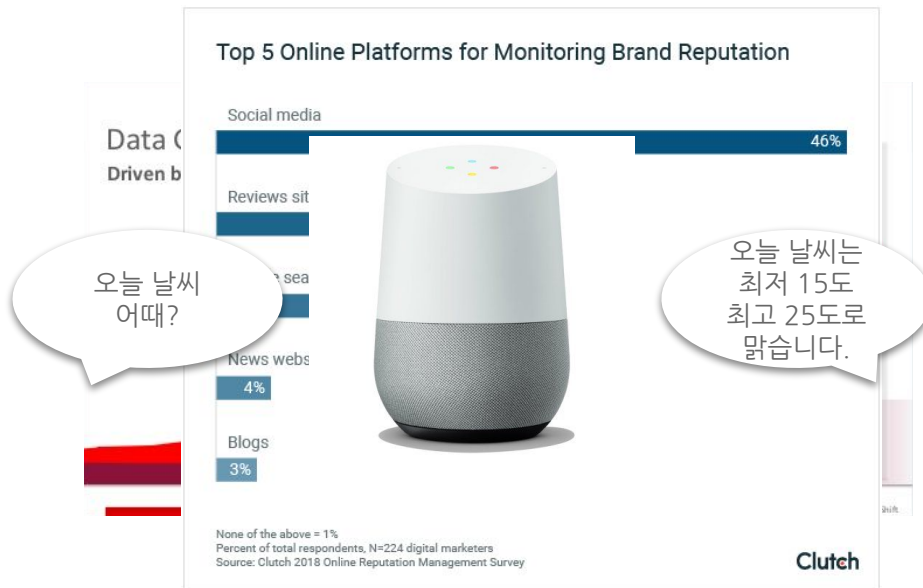
소통 패러다임의 변화

- 인간과 기계의 소통 패러다임 변화. 대화형 인터페이스로 변화
- 인터페이스가 점차 인간처럼 자연스러운 방법으로 개선되어 감
- 예) 인공지능 스피커, 인공지능 챗봇 등



자연어 처리, 왜 관심 가져야 하나

1. 비정형 데이터의 중요성
2. 온라인 데이터의 중요성
3. 소통 패러다임의 변화



3

자연어 처리가 어려운 이유



자연어 처리에 필요한 것





**자연어 처리가 좋은 성능을 내기
위해서는 명확해야 합니다.**

언어의 모호성 - 동음이의어

	동형이의어	동음이형어
의미	철자와 발음이 모두 같은 동음이의어	철자는 다르나 발음이 같은 동음이의어
예시	Turn right (부사, 오른쪽) That's right (형용사, 옳은)	I went to the sea(바다) to see(보다) my friend.
어려움	품사 및 의미파악 어려움	음성인식 어려움

언어의 모호성 - 다의어

하나의 단어가 여러개의 의미를 가질 수 있음

Bolt		
Apple		

개체명 인식의 어려움

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

1

구어와 문어의 차이

2

띄어쓰기에 어려움

3

청자와 화자의 관계에 따른 높임법

4

동음이의어, 운율적 요소에 따른 의미 변화

5

주어·서술어·목적어 등의 빈번한 생략

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

1

구어와 문어의 차이

문어 : 정돈된 문법을 사용하고 있어 애매모호함이 적음

구어 : 완벽한 문법이나 형식적인 의미에 구애받지 않고 사용

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

2

띄어쓰기에 어려움

아버지 가방에 들어가신다

아버지가 방에 들어가신다

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

3

청자와 화자의 관계에 따른 높임법

김 교수님한테 나 먼저 간다고 문자 보내줘.”

“네 알겠습니다. ‘나 먼저 간다’고 문자를 보냅니다.”

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

4

동음이의어, 운율적 요소에 따른 의미 변화

	만날때	헤어질때
영어	Hi	Bye
한국어	안녕!	안녕~~

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

5

주어·서술어·목적어 등의 빈번한 생략

문장의 필수 요소(주어, 서술어, 목적어 등)가 생략되면서 겪는 분석의 어려움

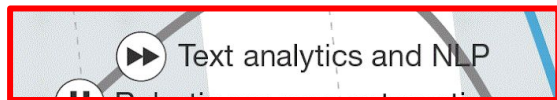
4

자연어 처리 전망



자연어 처리 기술 전망

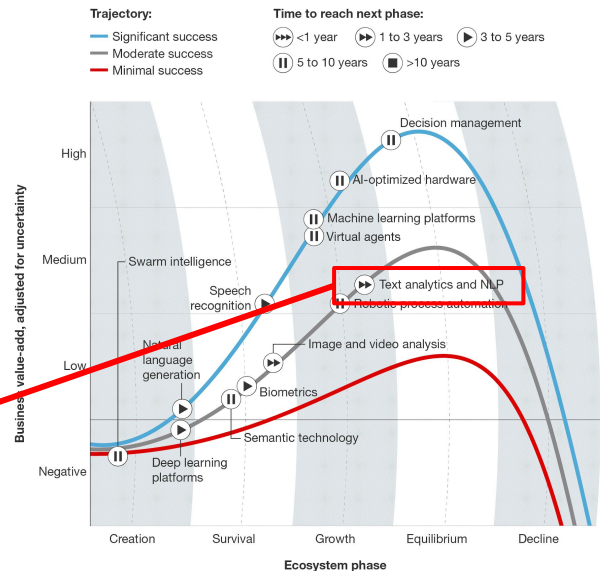
- 인공지능 기술 내에서도 자연어 처리는 빠르게 성장
- 기술전문 매체 테그레이더(TechRadar) 자료를 보면 인공지능 기술 중에서도 가능, 빠르게 성장하는 기술



FORRESTER RESEARCH

TechRadar™: Artificial Intelligence Technologies, Q1 '17

TechRadar™: Artificial Intelligence Technologies, Q1 2017

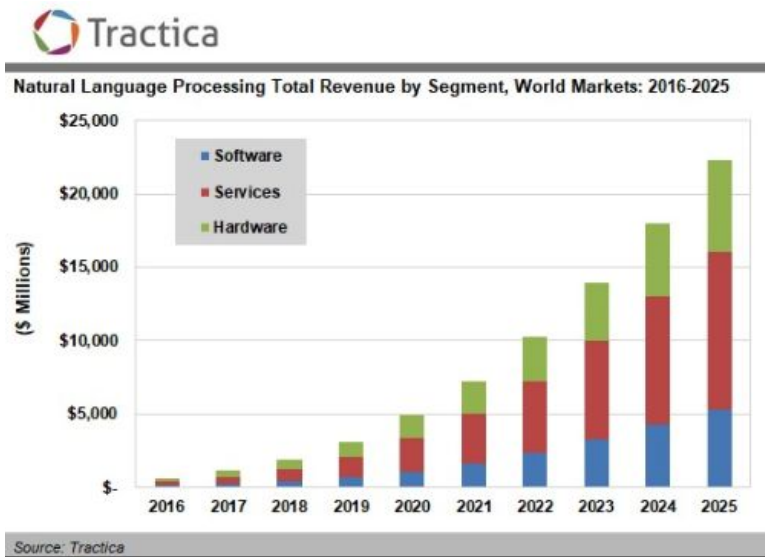


129161

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

자연어 처리 시장 전망

- 2016년 \$500M(한화 5,600억원)에서
- 2025년 \$22.3B(한화 24.9조)로 증가 (10년 내 44.6배 성장)
- 자연어 처리 시장 성장 동력은 “수요 증가”
 - 인공지능 스피커와 같은 스마트 장치 사용 증가
 - 웹 및 클라우드 기반 비즈니스 응용프로그램 증가
 - 비정형 데이터(Unstructured data)로부터 인사이트를 도출 Needs 증가



자연어 처리 현재 한계점

첫째, 도메인(산업)에 독립적인 범용 자연어 처리 솔루션이 없음


둘째, 자연어 처리 교육이 얼마나 오래 걸릴지, 결과가 얼마나 정확하며, 비즈니스 이점을 제공하기 위해 얼마나 정확해야 하는지를 예측하기 어려움

“인간이 컴퓨터와 대화하고 있다는 것을 깨닫지 못하고
인간과 대화를 계속할 수 있다면
컴퓨터는 지능적(Intelligence)인 것으로 간주될 수 있습니다.”

- 앨런 튜링 -



자연어 처리 전망

A woman in a yellow and blue climbing harness is ascending a white rock wall with various colored holds (yellow, orange, red, black). A man in a grey t-shirt is standing below her, spotting her. The wall has a grid of small holes. The text is overlaid on a semi-transparent white rectangle in the center of the image.

긍정적 시장 전망과
기술의 한계는
앞으로 충분히 성장할 수 있는 기회

5

일상 속 자연어 처리



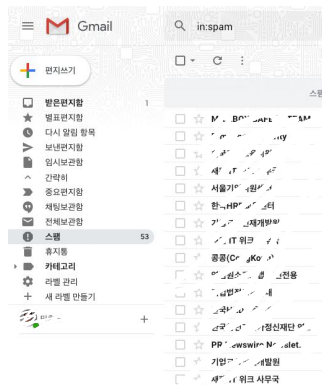
검색 엔진

- 과거 검색 엔진은 연산자(and, or 등)를 통한 검색이 가능
- 최근 검색 엔진은 검색창에 자연어 질의를 입력하면 적합한 답변



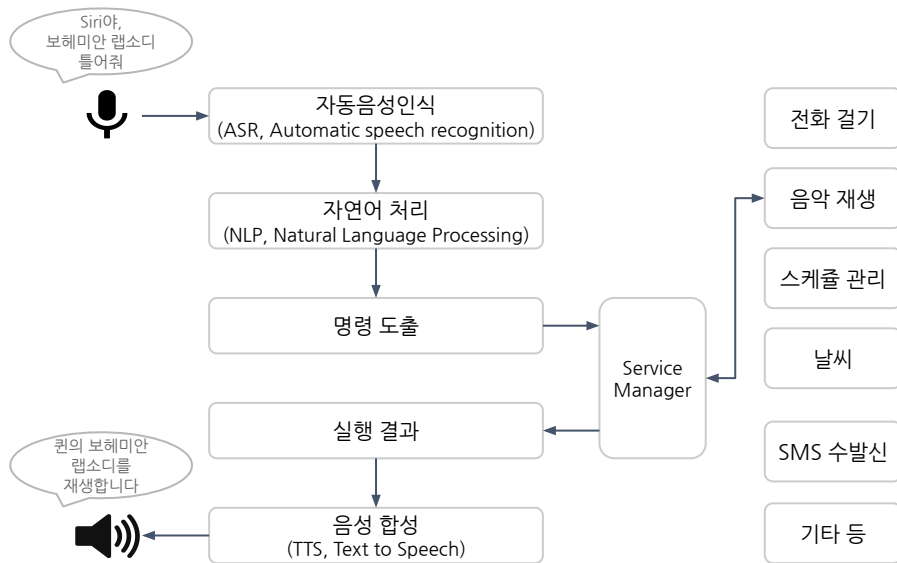
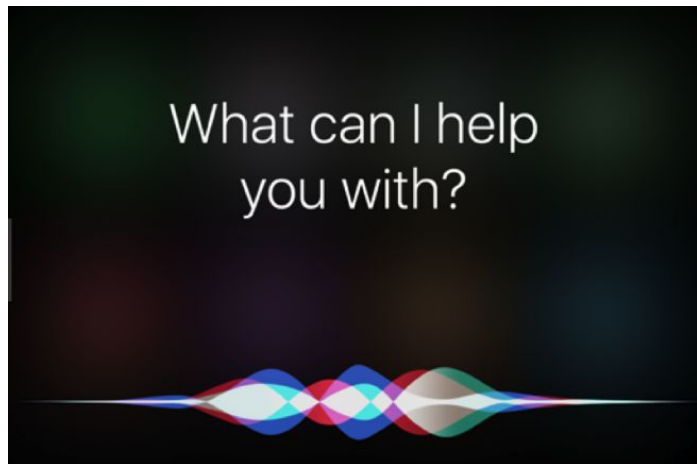
스팸 메일 분류

- 온라인 메일서비스를 사용하면서 따로 스팸메일 설정을 하지 않음
- 설정을 하지 않음



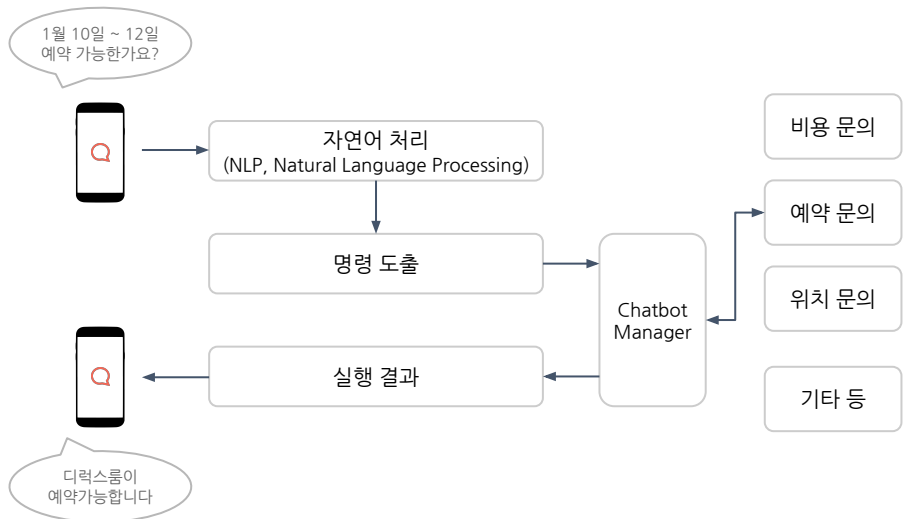
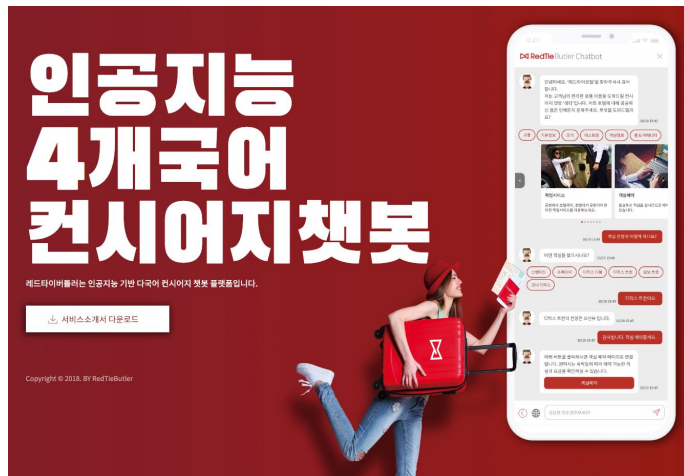
인공지능 비서

- 시리(Siri), 알렉사(Alexa) 등 음성기반의 인공지능 비서
- 음성으로 요청을 하면 문자로 변환하여 자연어 처리 엔진이 질의를 이해하여 처리하고 답변



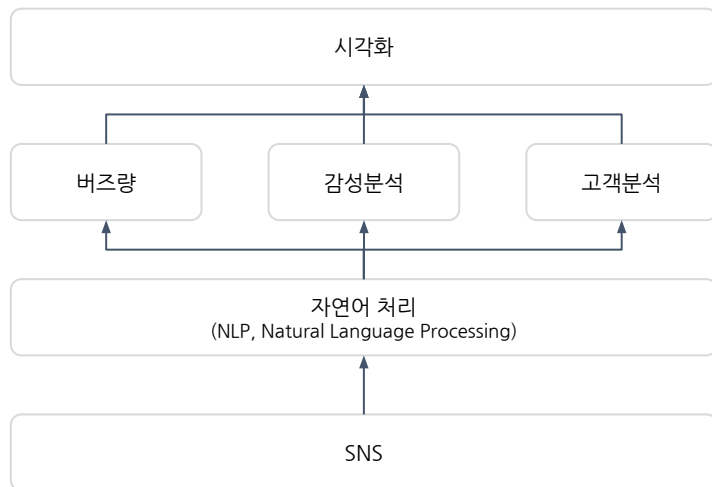
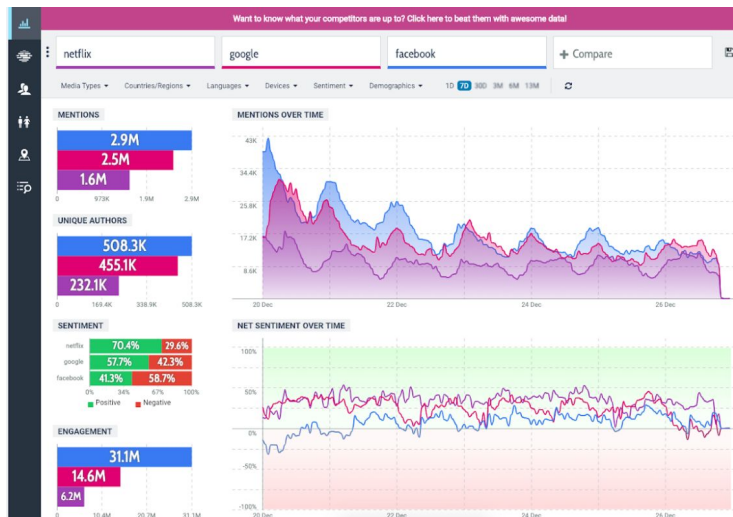
인공지능 챗봇

- 음성기반 인공지능 비서와 다르게 질의를 텍스트로 입력
- 텍스트를 입력하면 자연어 처리를 통해서 질의를 검색하여 텍스트 형태로 응답

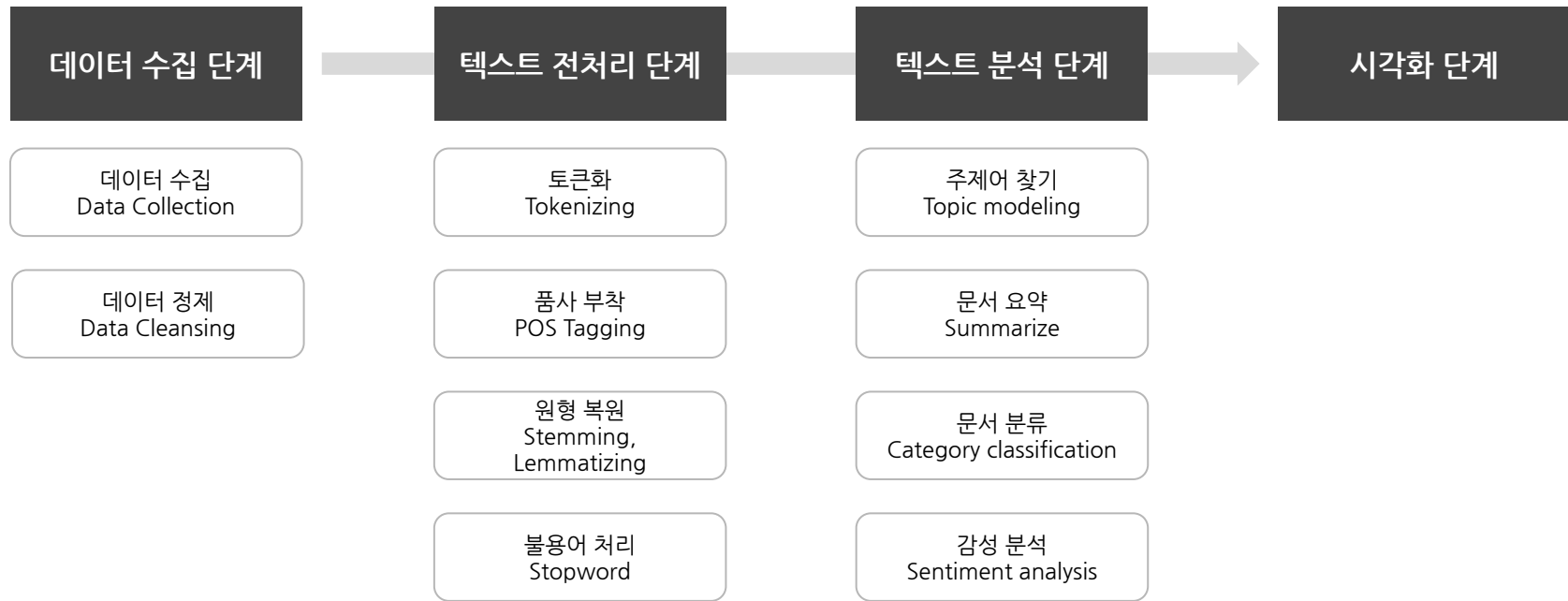


SNS 내 인지도 분석

- 온라인 미디어(뉴스, 블로그, SNS, 리뷰 등) 데이터를 수집하여 버즈량 및 감성분석
- 분석하고자 하는 대상의 시장 반응(긍정, 부정, 중립) 여부를 판단하여 전략수립



통계기반 자연어 처리 절차



자연어 처리 텍스트 분석 절차

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

인스타그램에서 “#보헤미안랩소디” 해시태그를 입력하고 포스트 검색결과를 수집

데이터 수집 단계

포스트 내용을 일괄된 포맷으로 정리

텍스트 전처리 단계

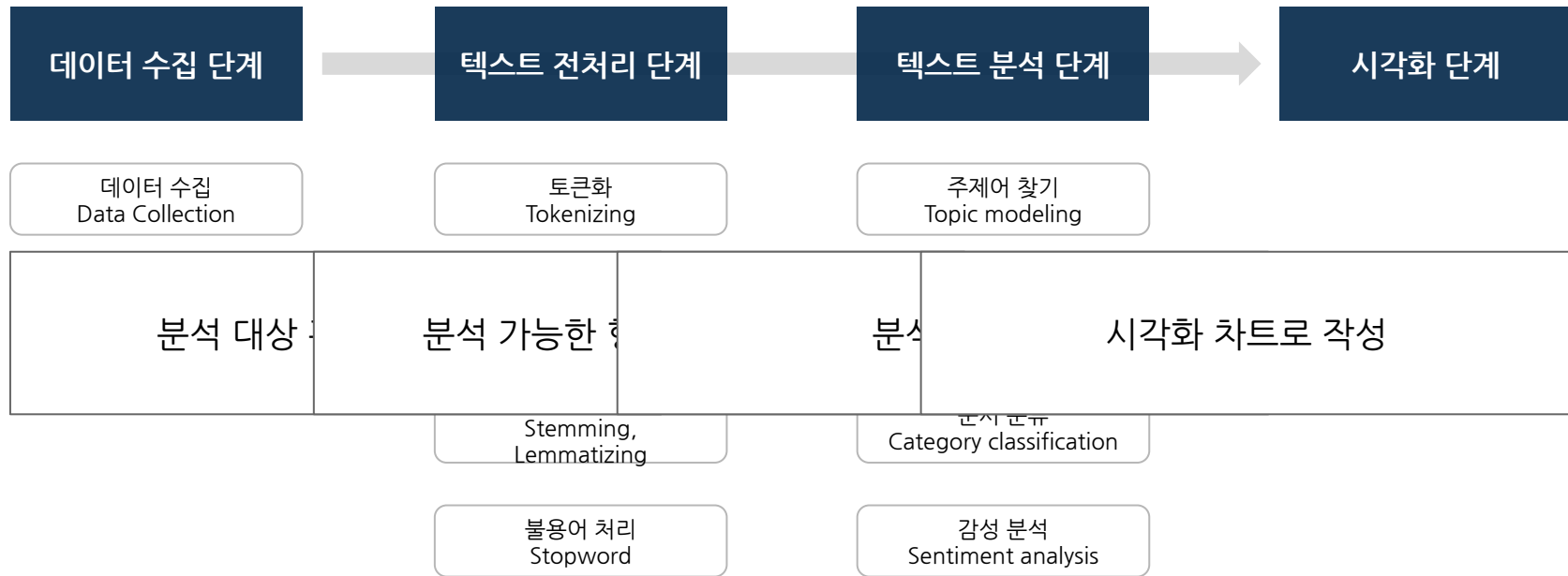
각 포스트 내용을 읽어보고 핵심 키워드, 긍정/부정/중립을 판단

텍스트 분석 단계

정리한 내용을 보고서로 정리

시각화 단계

자연어 처리 텍스트 분석 절차



자연어 처리 텍스트 분석 절차



데이터 수집 (Data Collection)

필요한 데이터를 선별하고 수집하여 저장하는 것

The screenshot shows the homepage of The New York Times dated Sunday, January 20, 2019. The page layout includes a top navigation bar with language options (English, Español, 中文), a search bar, and a 'SUBSCRIBE NOW' button. Below the masthead, there are several featured articles and sections. Annotations with brackets on the left side of the page highlight the following content for data collection:

- Sign Up: 'The Daily' Newsletter**: A promotional banner for the 'The Daily' podcast.
- The Neediest Cases Fund**: A small article snippet about arts center in the Bronx providing refuge for a family of 5.
- The Daily Mini Crossword**: A small article snippet about solving a bite-sized puzzle in just a few minutes.
- THE SHUTDOWN**: A section header for the main article.
- Trump Offers Deportation Protections in Exchange for Wall Funding**: The main headline of the featured article.
- Trump Offers Temporary 'Dreamer' Support in Return for Wall Funding**: A sub-headline for a video player.
- In Trump's Immigration Announcement, a Compromise Snubbed All Around**: A sub-headline for an analysis piece.
- BuzzFeed News Faces Scrutiny After Mueller Denies a Dramatic Report**: A headline for a news article.
- The rare statement by Mr. Mueller's office challenged the facts of the article.**: A headline for a news article.
- Opinion >**: A section header for opinion pieces.
- The Revenge of the Middle-Aged Frenchwoman**: A headline for an opinion piece.
- My Mother's Secrets**: A headline for an opinion piece.
- The Malign Incompetence of the British Ruling Class**: A headline for an opinion piece.
- In Search of Non-Toxic Manhood**: A headline for an opinion piece.
- Beware the Furies, President Trump**: A headline for an opinion piece.
- Time to Break the Silence on Palestine**: A headline for an opinion piece.
- How to Inoculate Against Anti-Vaxxers**: A headline for an opinion piece.

데이터를 쉽게 사용할 수 있도록 불필요한 부분을 제거



자연어 처리 텍스트 분석 절차

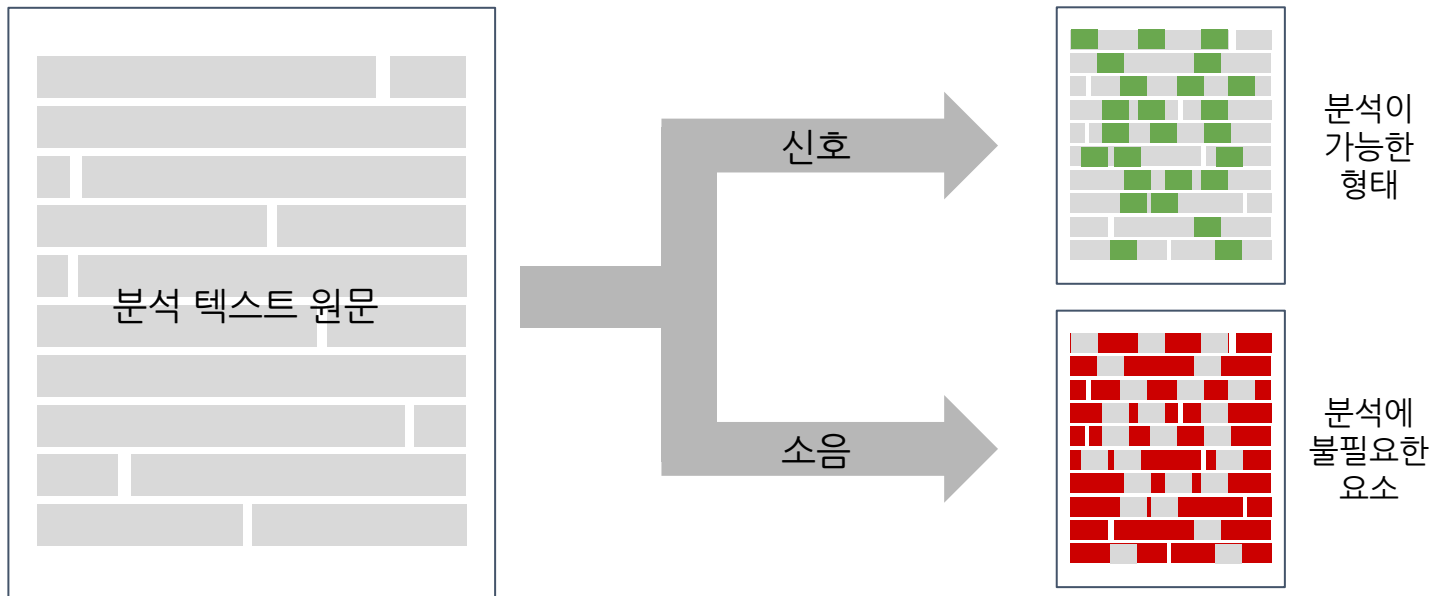


텍스트 전처리 단계

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

텍스트 전처리 단계

텍스트 분석을 위해서 기계가 텍스트를 이해할 수 있도록 표준화하는 단계



토큰화 (Tokenizing)

문장을 형태소로 분리하는 작업

형태소 形態素

+ 단어장 저장

표준국어대사전

고려대한국어대사전

우리말샘

<

>

예문 열기 ~

명사

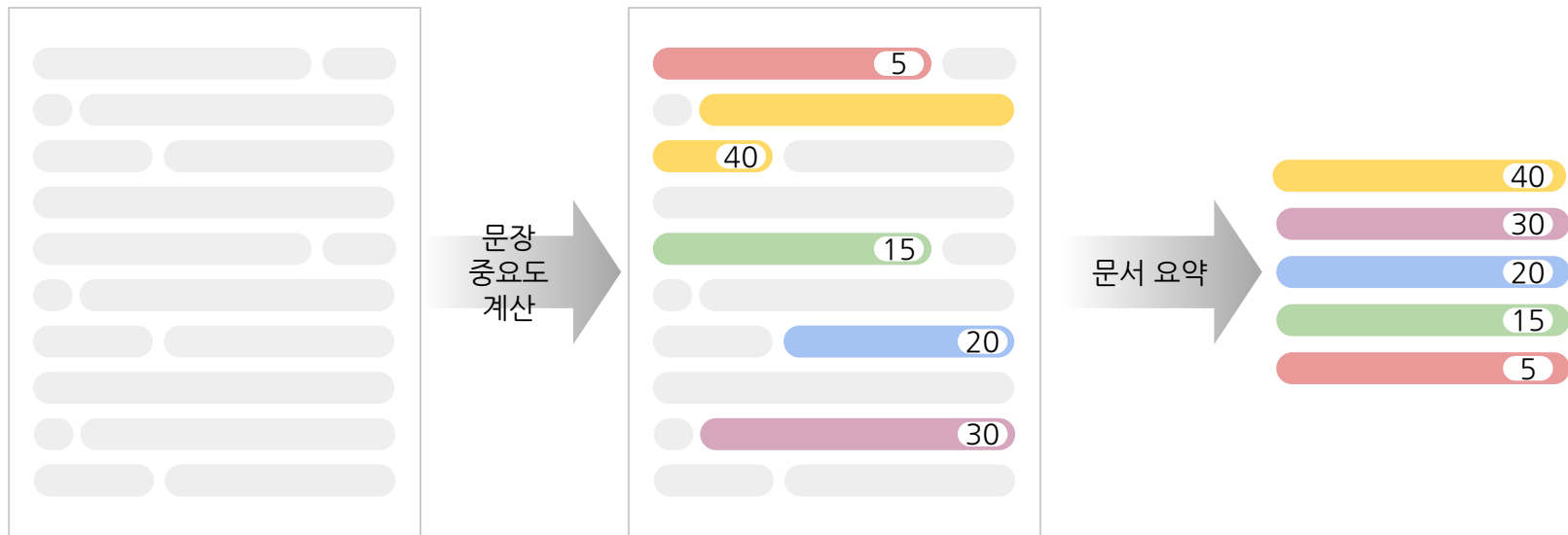
- 언어 뜻을 가진 가장 작은 말의 단위. '이야기책'의 '이야기', '책¹' 따위이다.
- 언어 문법적 또는 관계적인 뜻만을 나타내는 단어나 단어 성분. 프랑스의 언어학자 마르티네(Martinet, A.)가 제시하였다.
ㄴ형태질.

텍스트 분석 단계



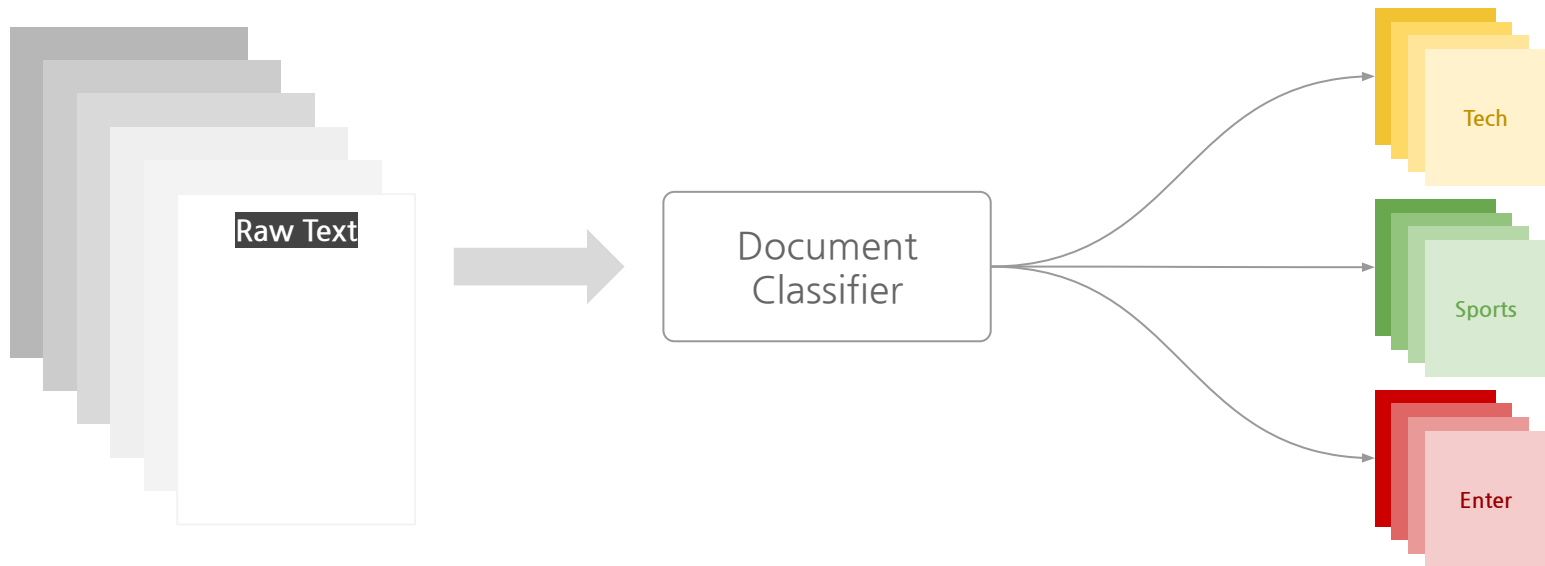
문서 요약 (Text Summarize)

문서 내에서 주요 문장을 찾아 요약



문서 분류 (Category Classification)

문서 내 단어 혹은 문장을 분석하여 문서를 분류



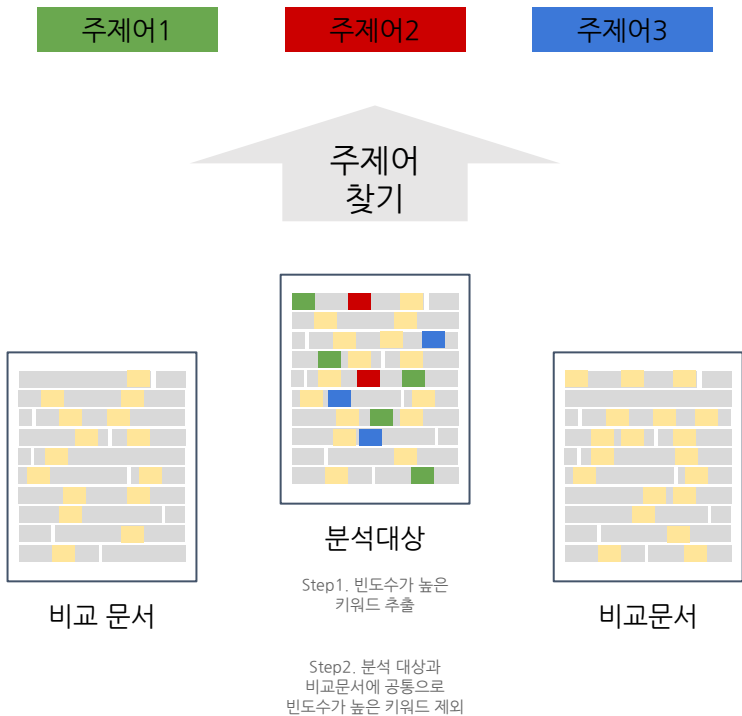
감성 분석 (Sentiment Analysis)

문서 내 나타난 사람들의 태도, 의견, 성향 같은 주관성을 분석



주제어 찾기 (Topic Modeling)

문서 내에서 주제를 발견하기 위한 모델

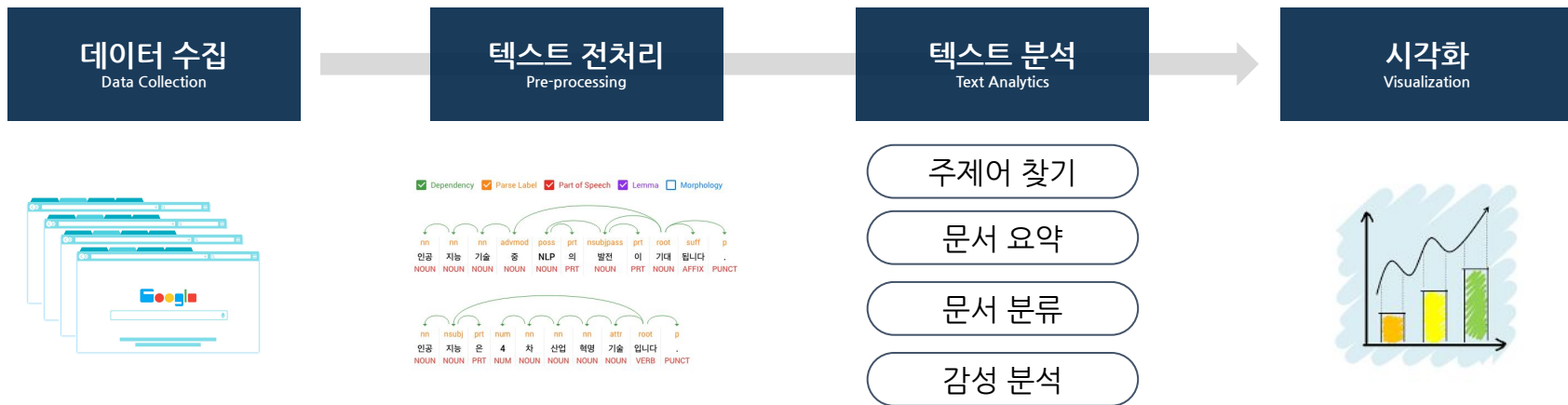


시각화 단계



시각화

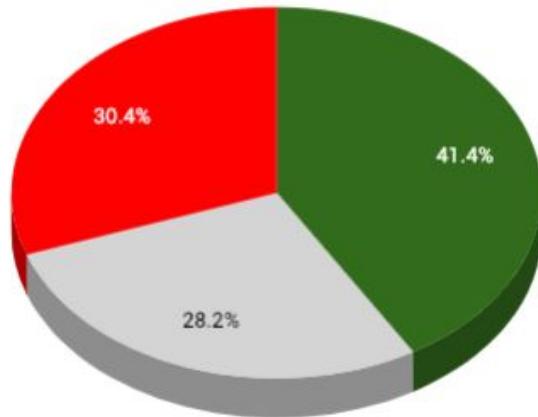
데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정



시각화 예시



● Positive ● Neutral ● Negative



감사합니다.

Insight⁺campus