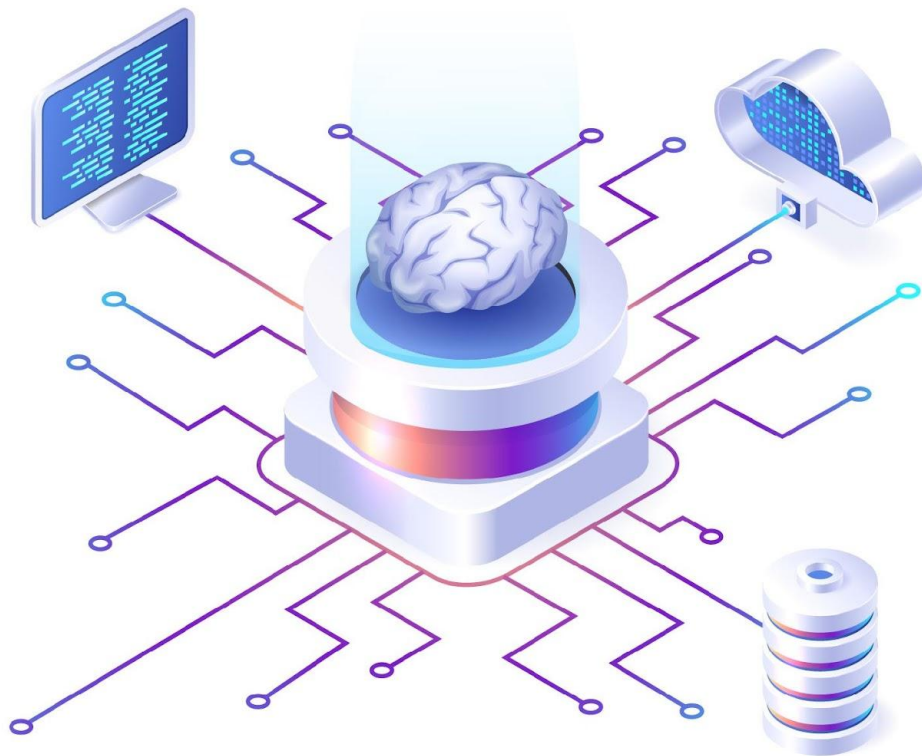
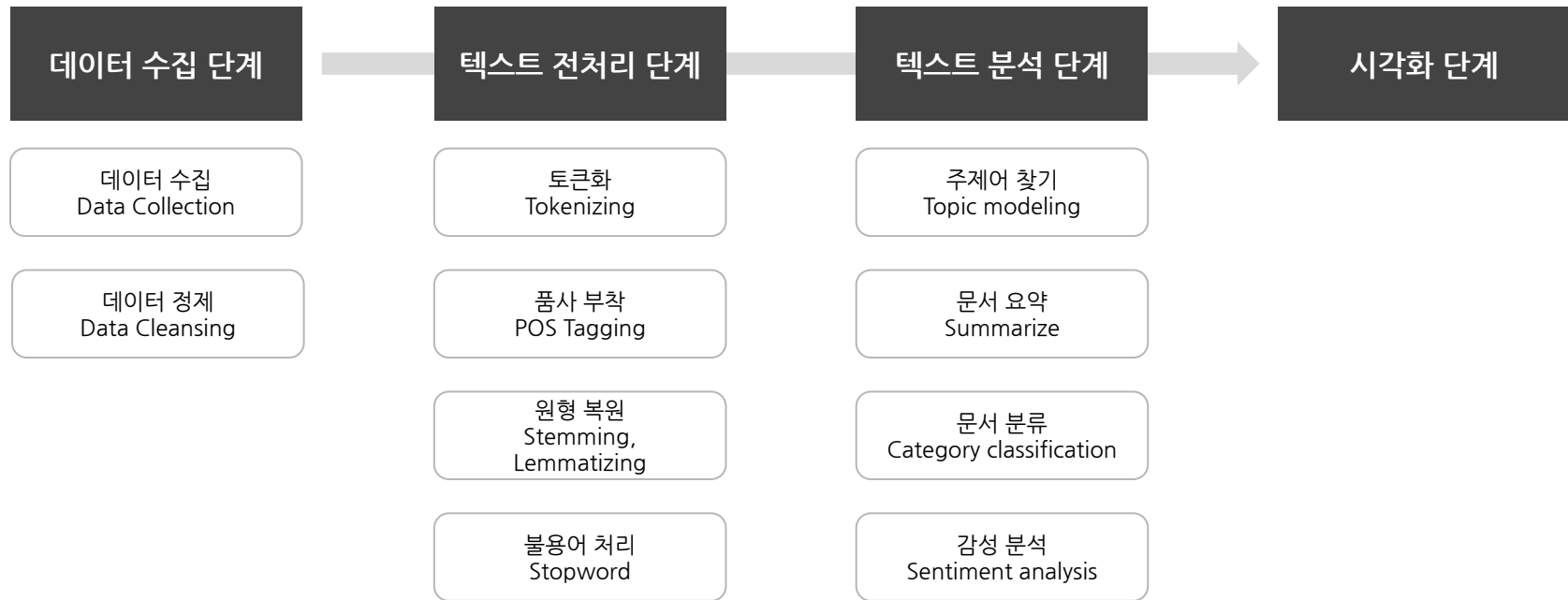


통계기반 자연어 처리 과정

실무형 인공지능 자연어처리



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

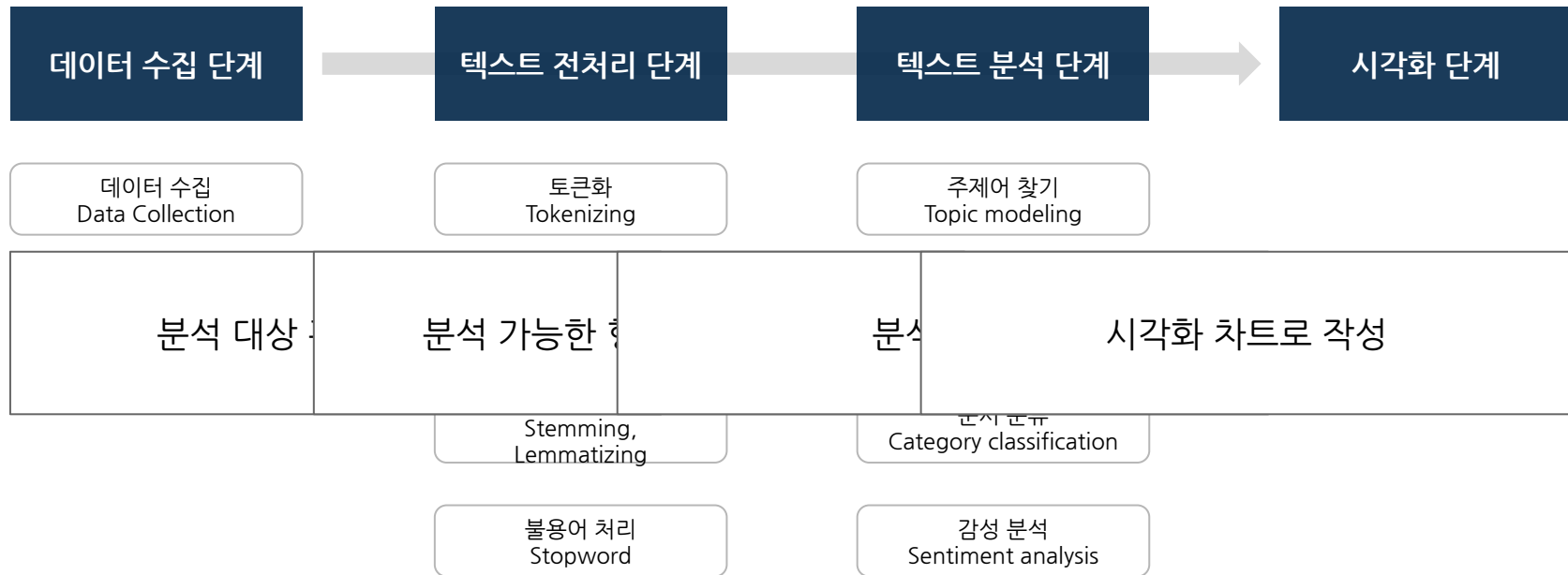
각 포스트 내용을 읽어보고 핵심 키워드, 긍정/부정/중립을 판단

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 subscription/login buttons. Below the masthead, there are several featured articles and sections. Annotations with white boxes and lines point to specific elements:

- Left Column (Main Article):**
 - Section Header:** "THE SHUTDOWN" and "Trump Offers Deportation Protections in Exchange for Wall Funding".
 - Text:** "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." and "What Mr. Trump billed as a compromise pleased neither the Democratic congressional leaders nor his core supporters."
 - Image:** A photo of President Trump at a podium with the caption "Trump Offers Temporary 'Dreamer' Support in Return for Wall Funding".
 - Caption:** "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 Brumer for The New York Times."
 - Section Header:** "In Trump's Immigration Announcement, a Compromise Snubbed All Around".
 - Text:** "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."
 - Section Header:** "BuzzFeed News Faces Scrutiny After Mueller Denies a Dramatic Report".
 - Text:** "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." and "Whether BuzzFeed's reporting can stand up to further scrutiny is now at the center of a test of the news media's credibility."
- Right Column (Opinion & Other Articles):**
 - Opinion >**
 - Section Header:** "The Revenge of the Middle-Aged Frenchwoman".
 - Text:** "'I would like 50-year-old women to stop sending me photos of their bottoms and breasts,' a French writer pleaded."
 - Section Header:** "My Mother's Secrets".
 - Text:** "She thought she was protecting her children by not telling us her harrowing tale of fleeing China."
 - Section Header:** "The Malign Incompetence of the British Ruling Class".
 - Section Header:** "In Search of Non-Toxic Manhood".
 - Section Header:** "Time to Break the Silence on Palestine".
 - Section Header:** "How to Inoculate Against Anti-Vaxxers".
- Bottom Section:**
 - Section Header:** "The rare statement by Mr. Mueller's office challenged the facts of the article."
 - Text:** "Anastasia Edel" and "Trymaine Lee".

데이터 정제 (Data Cleansing)

데이터를 쉽게 사용할 수 있도록 불필요한 부분을 제거

2019.05.21 20:35

메뉴

불필요
문구

사람은 흔적을 남기고...흔적은 기회를 낳는다

당신의 흔적에 기회가 있다

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



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

직관은 영감을 가져다주지만 이에 기반한 결정이 항상 옳은 것은 아니다.

당신의 흔적에

정시간 인기기사

인기 기사
링크

1. [미국] 기수 일제히 '낙동강 상류'로 남동해로...
2. [미국] 기수 일제히 '낙동강 상류'로 남동해로...
3. [미국] 기수 일제히 '낙동강 상류'로 남동해로...
4. [미국] 기수 일제히 '낙동강 상류'로 남동해로...
5. [미국] 기수 일제히 '낙동강 상류'로 남동해로...

광고

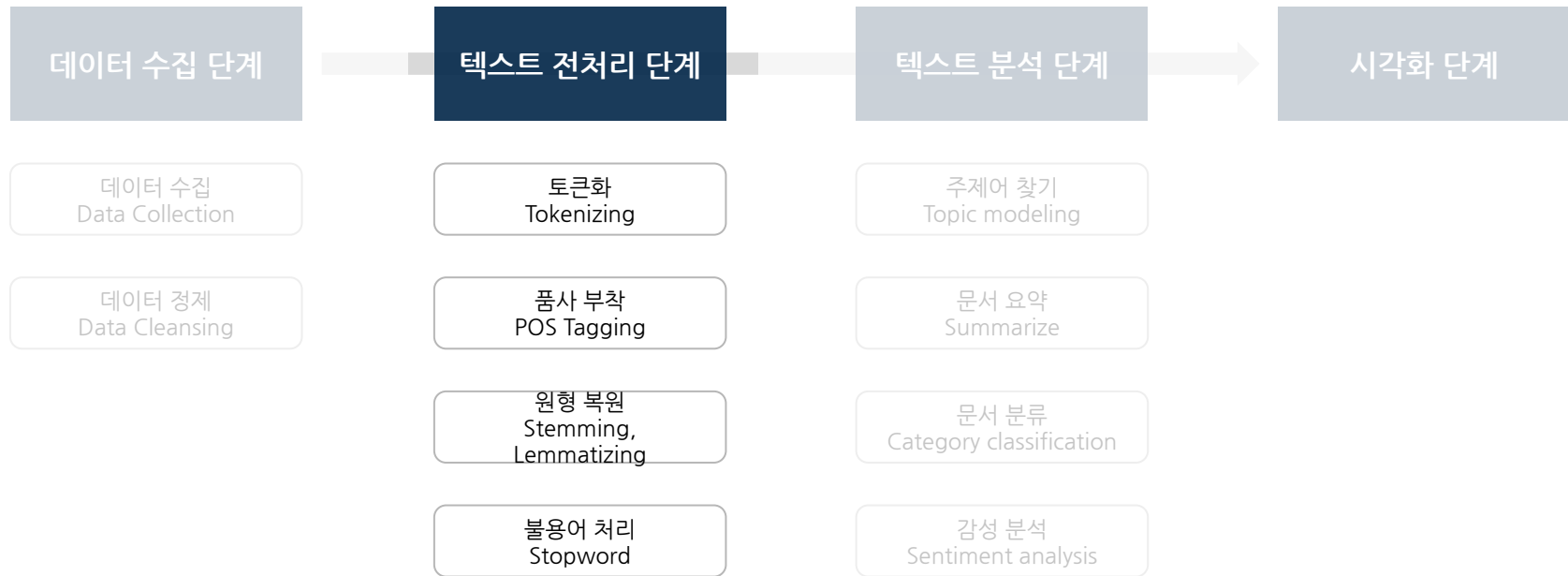
Look forward to a secure investment with DHA.

Find out more

DHA

Look forward

자연어 처리 텍스트 분석 절차

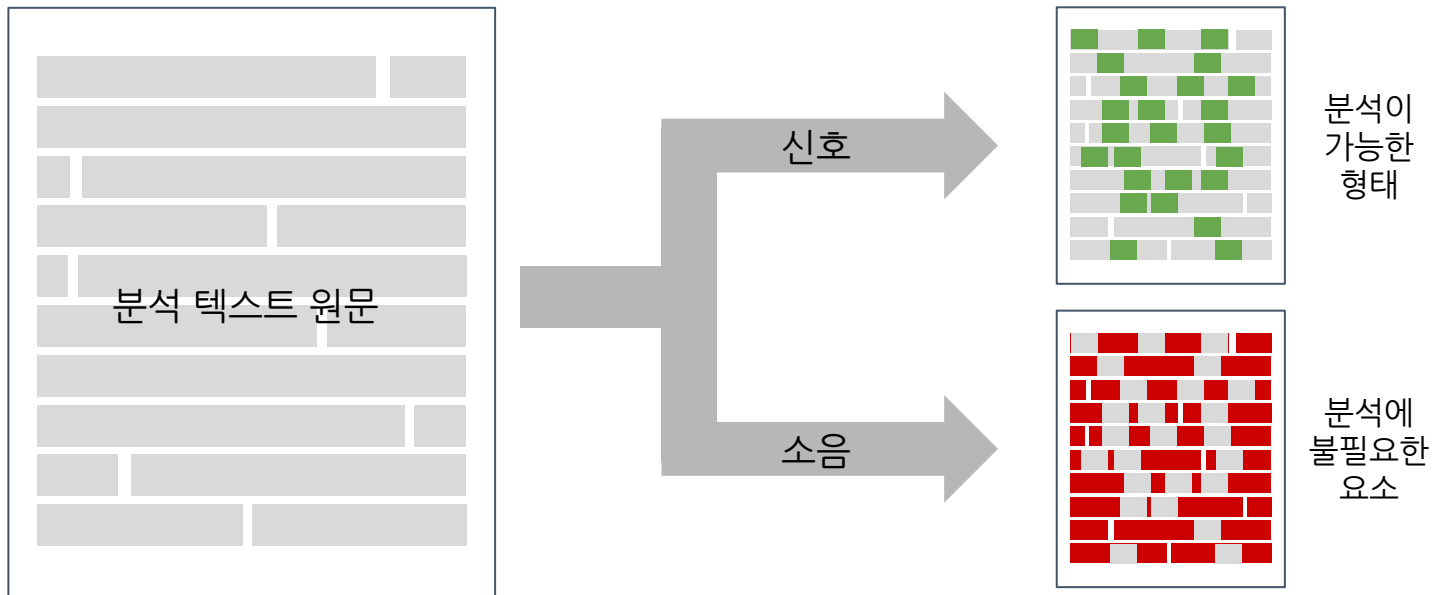


텍스트 전처리 단계

“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

형태-소 形態素

+ 단어장 저장

표준국어대사전

고려대한국어대사전

우리말샘

<

>

예문 열기 ~

명사

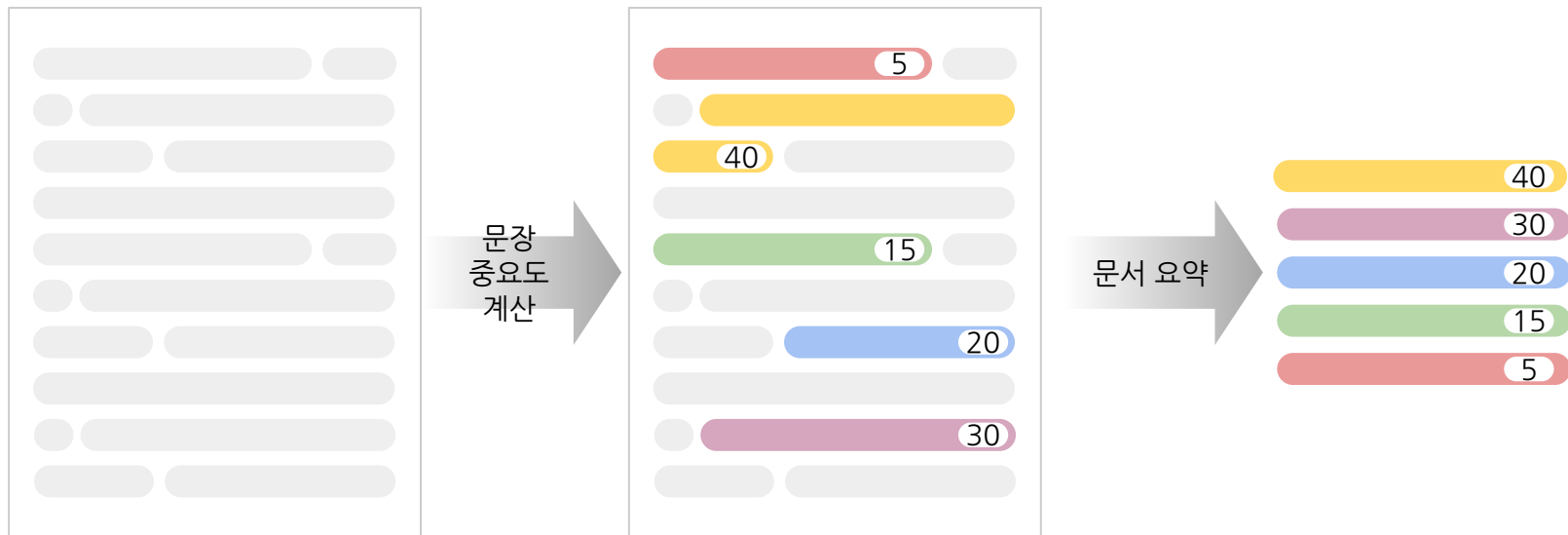
- 언어 뜻을 가진 가장 작은 말의 단위. '이야기책'의 '이야기', '책¹' 따위이다.
- 언어 문법적 또는 관계적인 뜻만을 나타내는 단어나 단어 성분. 프랑스의 언어학자 마르티네(Martinet, A.)가 제시하였다.
ㄴ형태질.

텍스트 분석 단계



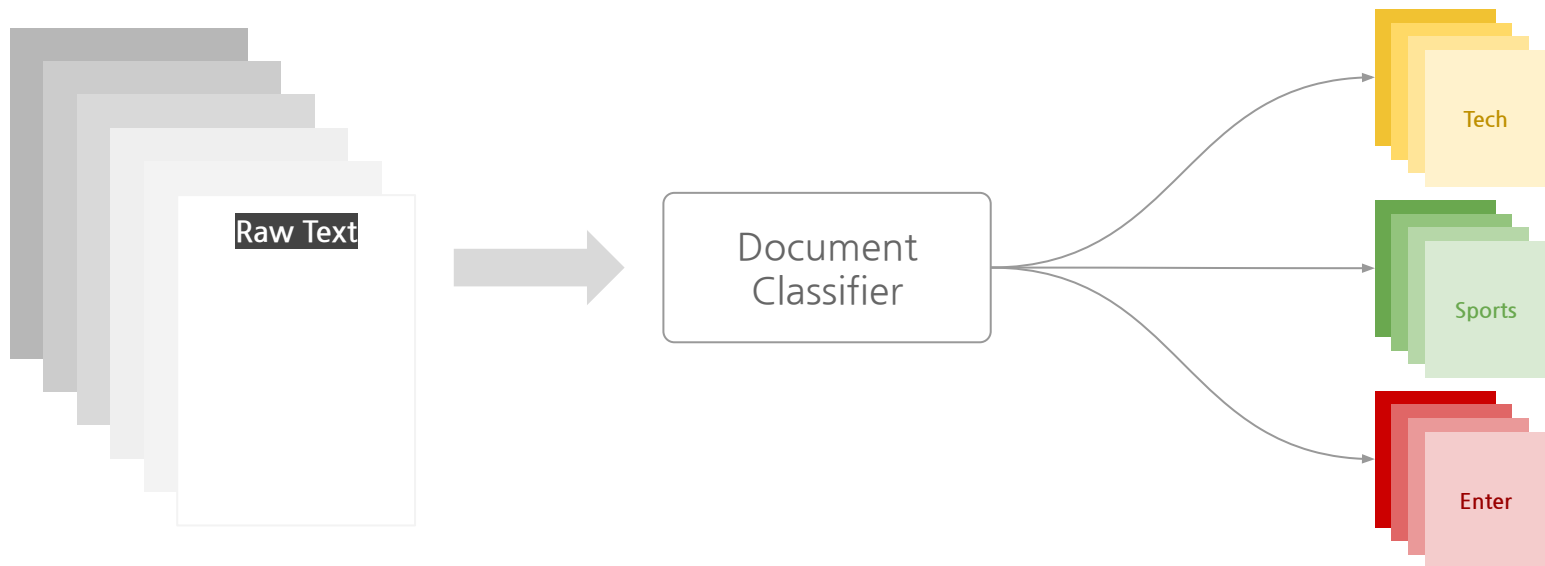
문서 요약 (Text Summarize)

문서 내에서 주요 문장을 찾아 요약



문서 분류 (Category Classification)

문서 내 단어 혹은 문장을 분석하여 문서를 분류



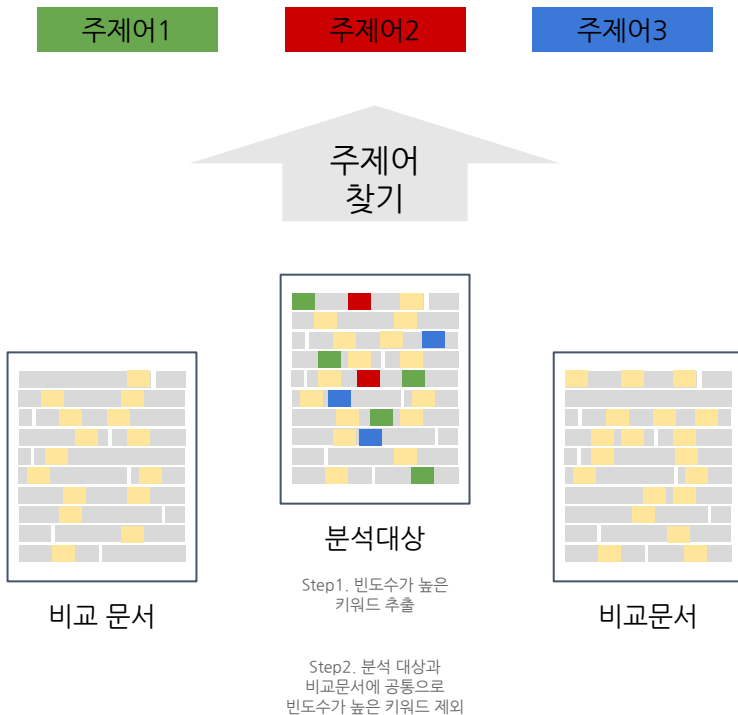
감성 분석 (Sentiment Analysis)

문서 내 나타난 사람들의 태도, 의견, 성향 같은 주관성을 분석



주제어 찾기 (Topic Modeling)

문서 내에서 주제를 발견하기 위한 모델

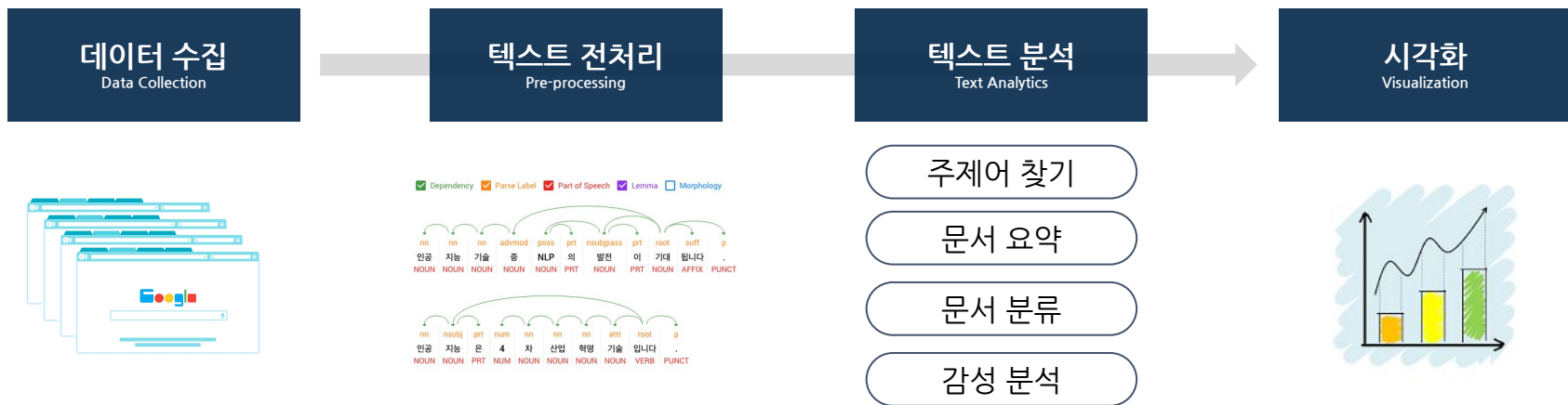


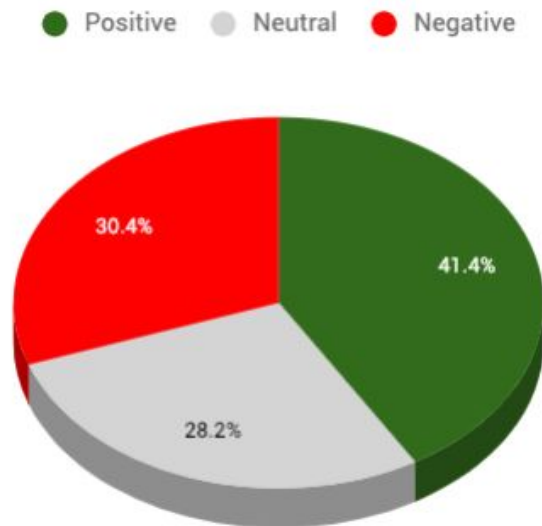
시각화 단계



시각화

데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정



[illegible]

감사합니다.

Insight⁺campus

