Insight campus SeSAC

통계기반 자연어 처리 과정

실무형 인공지능 자연어처리

Insight campus SeSAC

통계기반 자연어 처리 절차

데이터 수집 단계

- 데이터 수집
  Data Collection
- 데이터 정제
  Data Cleansing

텍스트 전처리 단계

- 토큰화
  Tokenizing
- 품사 부착
  POS Tagging
- 원형 복원
  Stemming,
  Lemmatizing
- 불용어 처리
  Stopword

텍스트 분석 단계

- 주제어 찾기
  Topic modeling
- 문서 요약
  Summarize
- 문서 분류
  Category classification
- 감성 분석
  Sentiment analysis

시각화 단계

Insight campus SeSAC

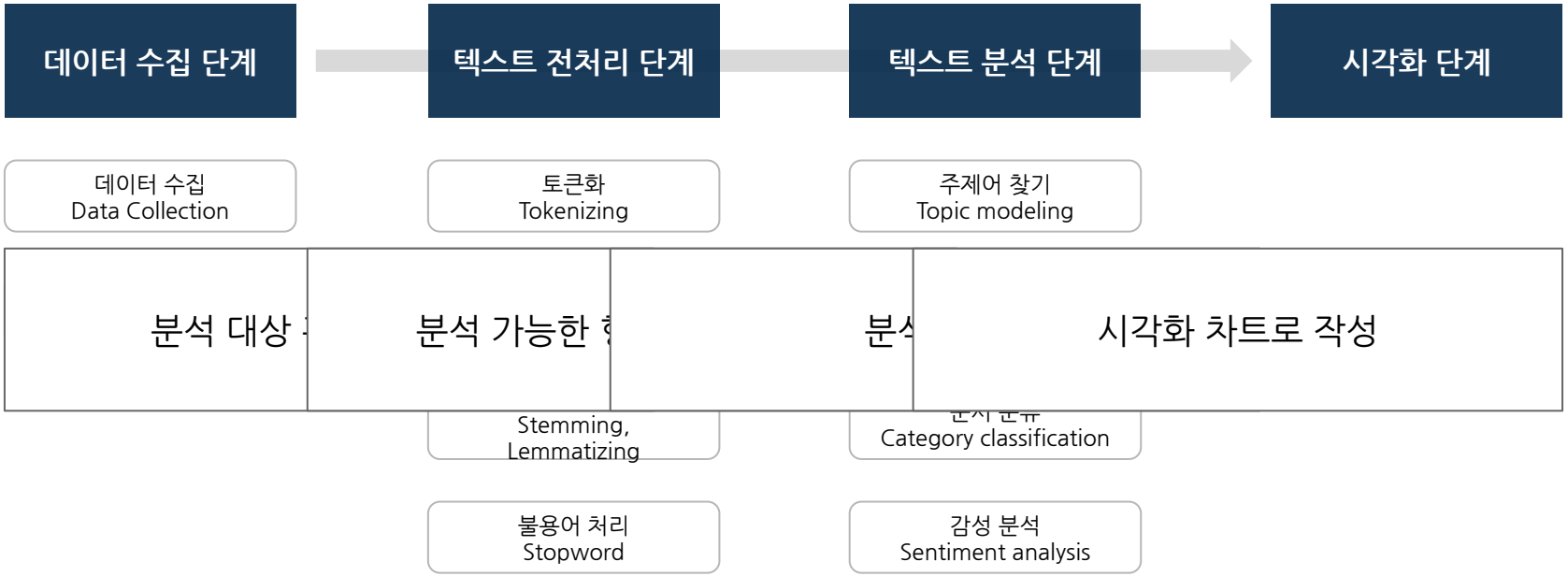
자연어 처리 텍스트 분석 절차

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 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

Insight campus SeSAC

자연어 처리 텍스트 분석 절차



Insight campus SeSAC

자연어 처리 텍스트 분석 절차

데이터 수집 단계

- 데이터 수집
  Data Collection
- 데이터 정제
  Data Cleansing

텍스트 전처리 단계

- 토큰화
  Tokenizing
- 품사 부착
  POS Tagging
- 원형 복원
  Stemming, Lemmatizing
- 불용어 처리
  Stopword

텍스트 분석 단계

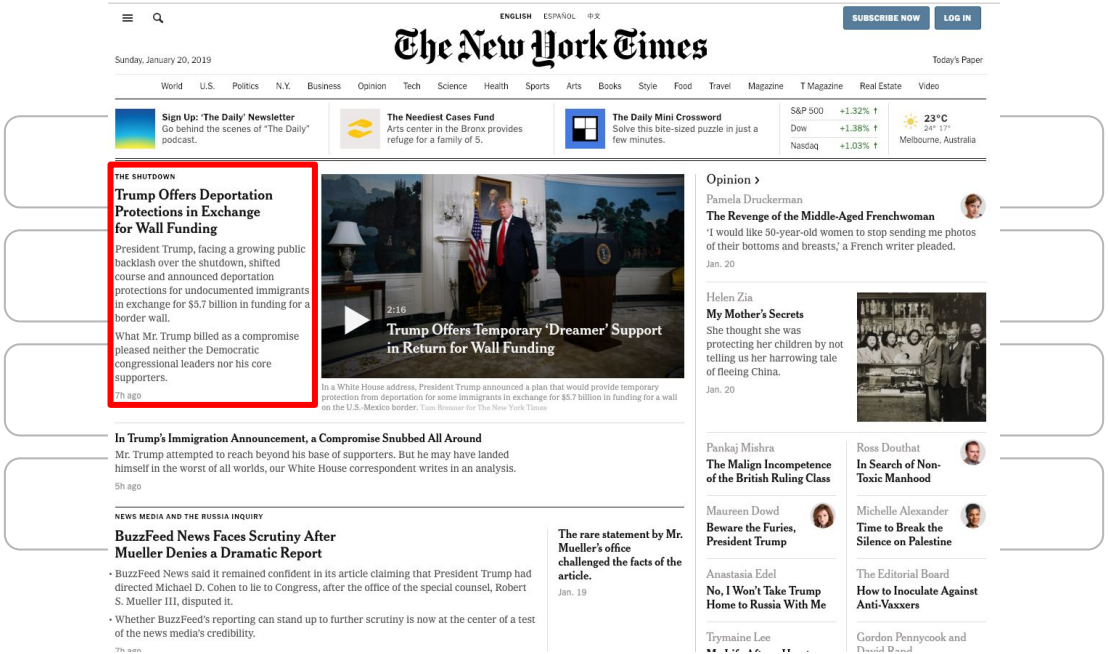
- 주제어 찾기
  Topic modeling
- 문서 요약
  Summarize
- 문서 분류
  Category classification
- 감성 분석
  Sentiment analysis

시각화 단계

Insight campus SeSAC

데이터 수집 (Data Collection)

필요한 데이터를 선별하고 수집하여 저장하는 것



Insight campus SeSAC

데이터 정제 (Data Cleansing)

데이터를 쉽게 사용할 수 있도록 불필요한 부분을 제거

Insight campus SeSAC

자연어 처리 텍스트 분석 절차

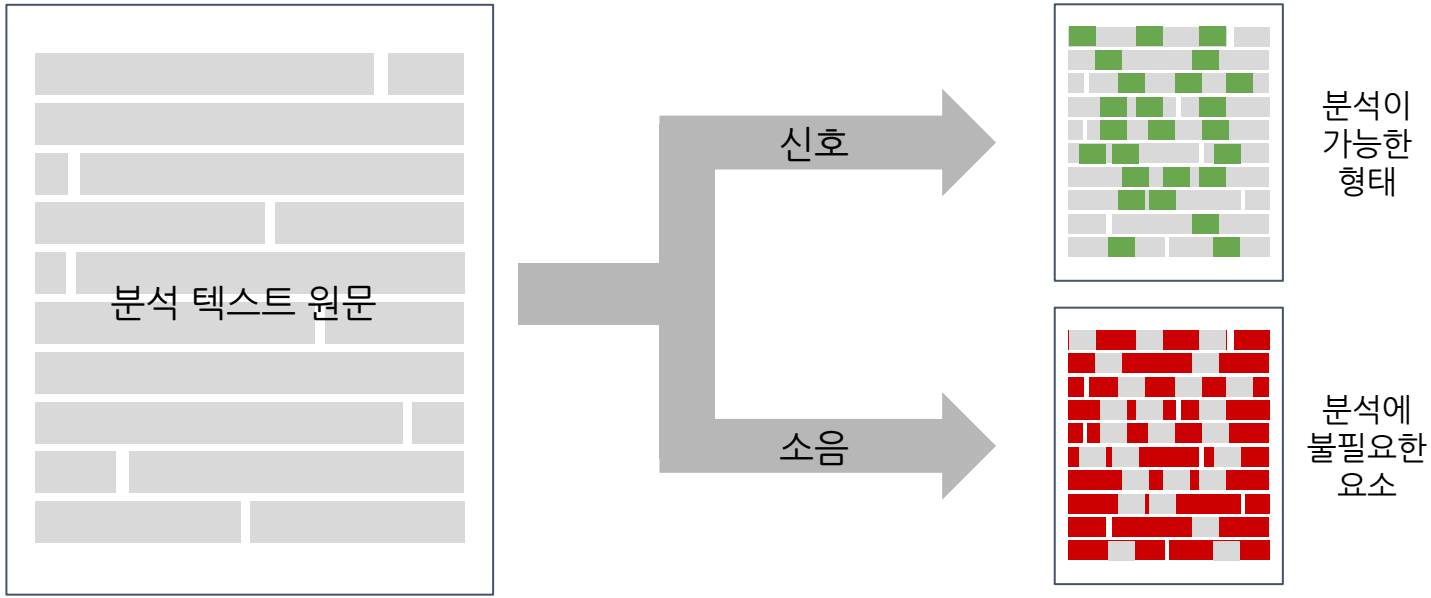
Insight campus SeSAC

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

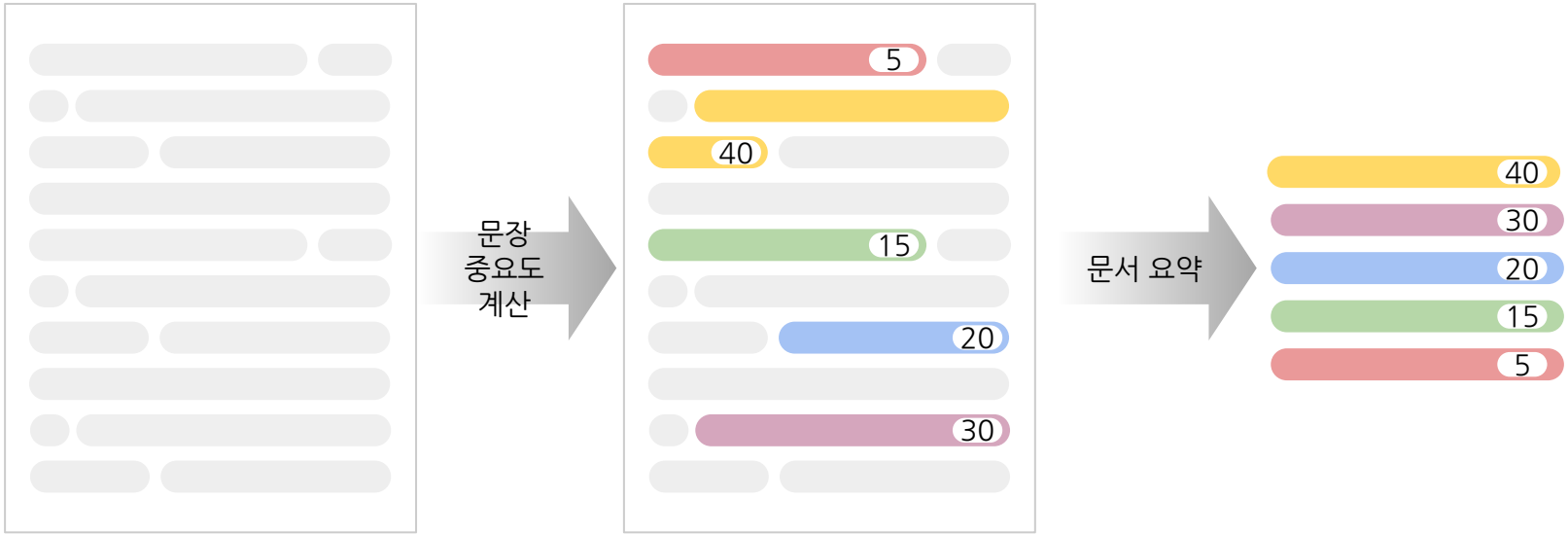
Insight campus SeSAC

텍스트 분석 단계

데이터 수집 Data Collection	텍스트 전처리 Pre-processing	텍스트 분석 Text Analytics	시각화 Visualization
데이터 수집 : Data Collection	형태소 분석 : Tokenizing	주제어 찾기 : Topic modeling	Word Cloud
데이터 정제 : Data Cleansing	어간 추출 : Stemming	문서 요약 : Summarize	Sentiment Pie Chart
	원형복원 : Lemmatizing	문서 분류 : Category classification	
	품사 태깅 : POS Tagging	감성 분석 : Sentiment analysis	

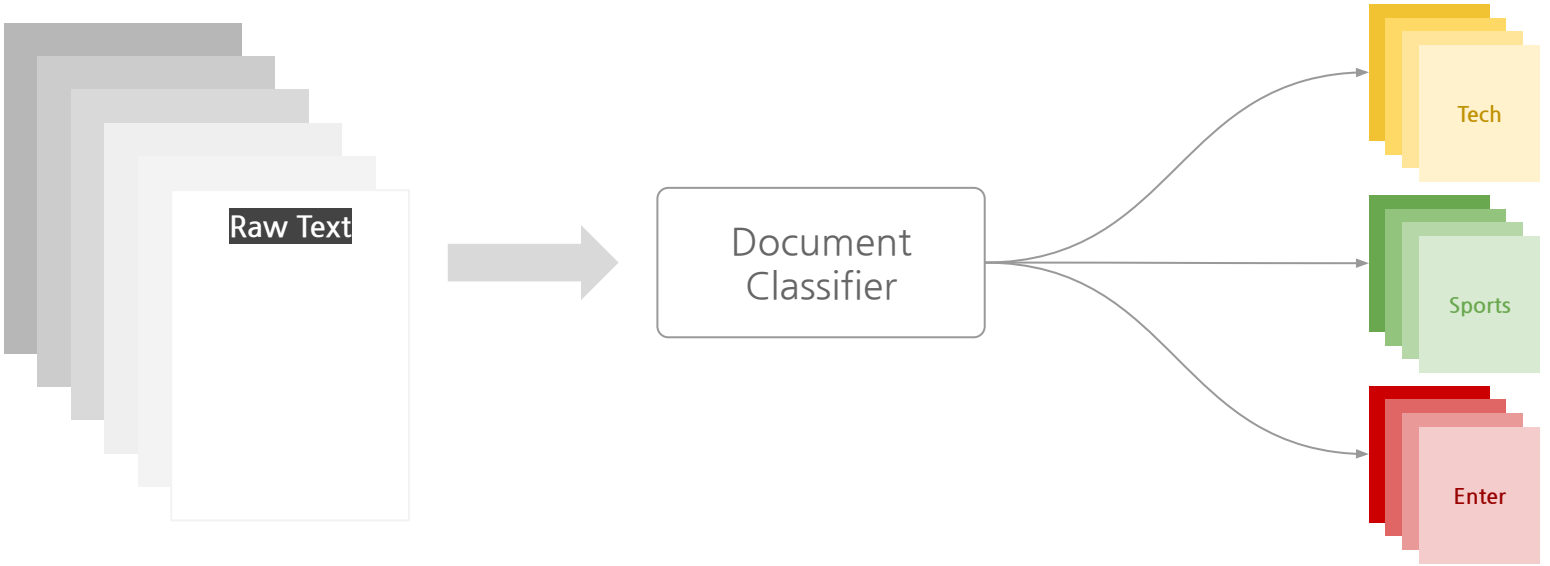
문서 요약 (Text Summarize)

문서 내에서 주요 문장을 찾아 요약



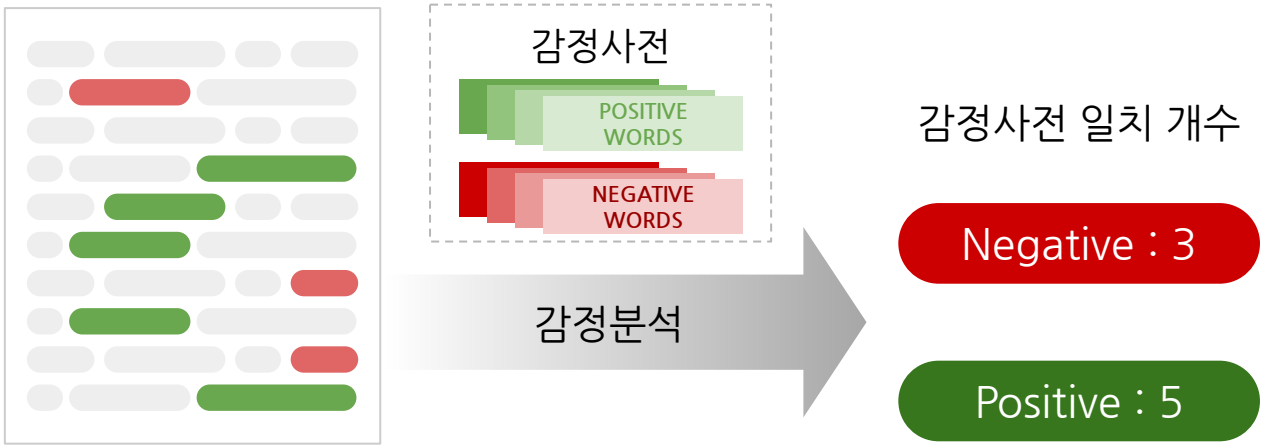
문서 분류 (Category Classification)

문서 내 단어 혹은 문장을 분석하여 문서를 분류



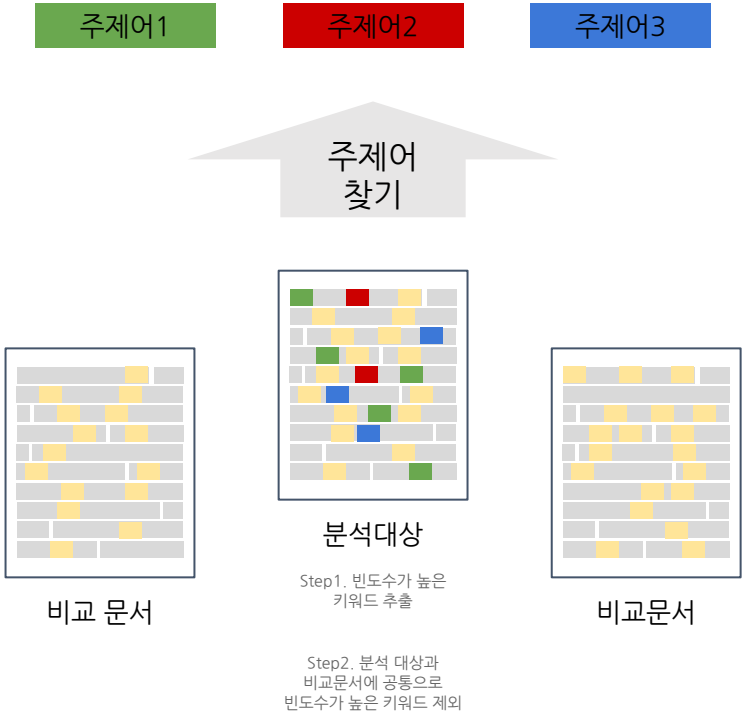
감성 분석 (Sentiment Analysis)

문서 내 나타난 사람들의 태도, 의견, 성향 같은 주관성을 분석



주제어 찾기 (Topic Modeling)

문서 내에서 주제를 발견하기 위한 모델



Insight campus SeSAC

시각화 단계

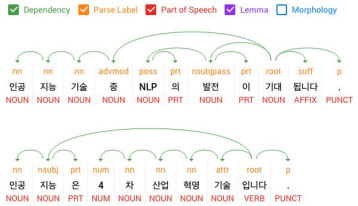
데이터 수집 Data Collection	텍스트 전처리 Pre-processing	텍스트 분석 Text Analytics	시각화 Visualization
데이터 수집 : Data Collection	형태소 분석 : Tokenizing	주제어 찾기 : Topic modeling	Word Cloud
데이터 정제 : Data Cleansing	어간 추출 : Stemming	문서 요약 : Summarize	Sentiment Pie Chart
	원형복원 : Lemmatizing	문서 분류 : Category classification	
	품사 태깅 : POS Tagging	감성 분석 : Sentiment analysis	

시각화

데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정

데이터 수집
Data Collection

텍스트 전처리
Pre-processing



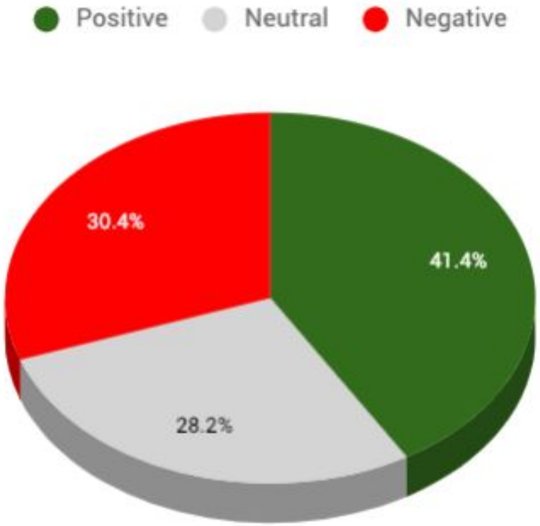
텍스트 분석
Text Analytics

- 주제어 찾기
- 문서 요약
- 문서 분류
- 감성 분석

시각화
Visualization

시각화 예시


send
features
binary
simple
audience
messages
tasks
communication
hardware
development
Wide
browsers
information
search
applications
ethical
responsibilities
engines
components
protocols
files
networking
professionals
transmit
illicit
preferences
security
opportunities
computer
careers
viruses
essential
Web
access
legal
configure
job
functions
proprietary
concepts
Internet
industry
servers
certificates
PIM
malware
resources
e-mail
system
skills
name
workplace
models
caching
software
users
issues
DNS
email
project
principles
instant
clients
popular
database
services
infrastructure
World
social
open-source
technology
roles
common
management
customization
personal
fundamentals
basic
webcast
cookies
domain


Positive
Neutral
Negative
30.4%
41.4%
28.2%

감사합니다.

Insight campus SeSAC

daniel@fins.ai