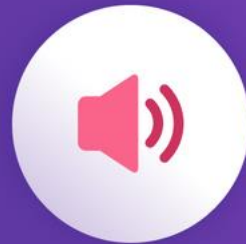
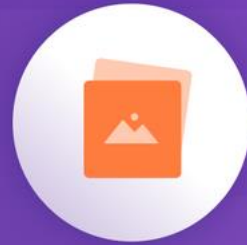
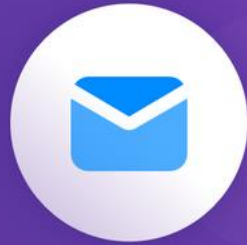


비정형 데이터 분석방법



정형 데이터(Structured Data)

미리 만들어진 형식 틀에 저장되는 데이터로 행과 열로 이루어진 표에 저장할 수 있는 데이터입니다.

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

반정형 데이터 (Semi-Structured Data)

관계형 데이터베이스나 다른 형태의 데이터 테이블과 연결된 정형 구조의 데이터 모델을 준수하지 않는 정형 데이터의 한 형태입니다. 태그나 기타 마커가 포함되어 있어서 시맨틱 요소를 구분하고 데이터 내의 레코드와 필드 계층을 강제합니다.

```
<!DOCTYPE html>
<html>
<!-- created 2010-01-01 -->
<head>
  <title>sample</title>
</head>
<body>
  <p>Voluptatem accusantium
  totam rem aperiam.</p>
</body>
</html>
```

HTML

```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

```
{
  "회사": [
    {
      "이름": "Apple",
      "운영체제": [
        "macOS",
        "iOS"
      ]
    },
    {
      "이름": "Microsoft",
      "운영체제": [
        "MS-DOS",
        "Windows"
      ]
    }
  ]
}
```

JSON

비정형 데이터(Unstructured Data)

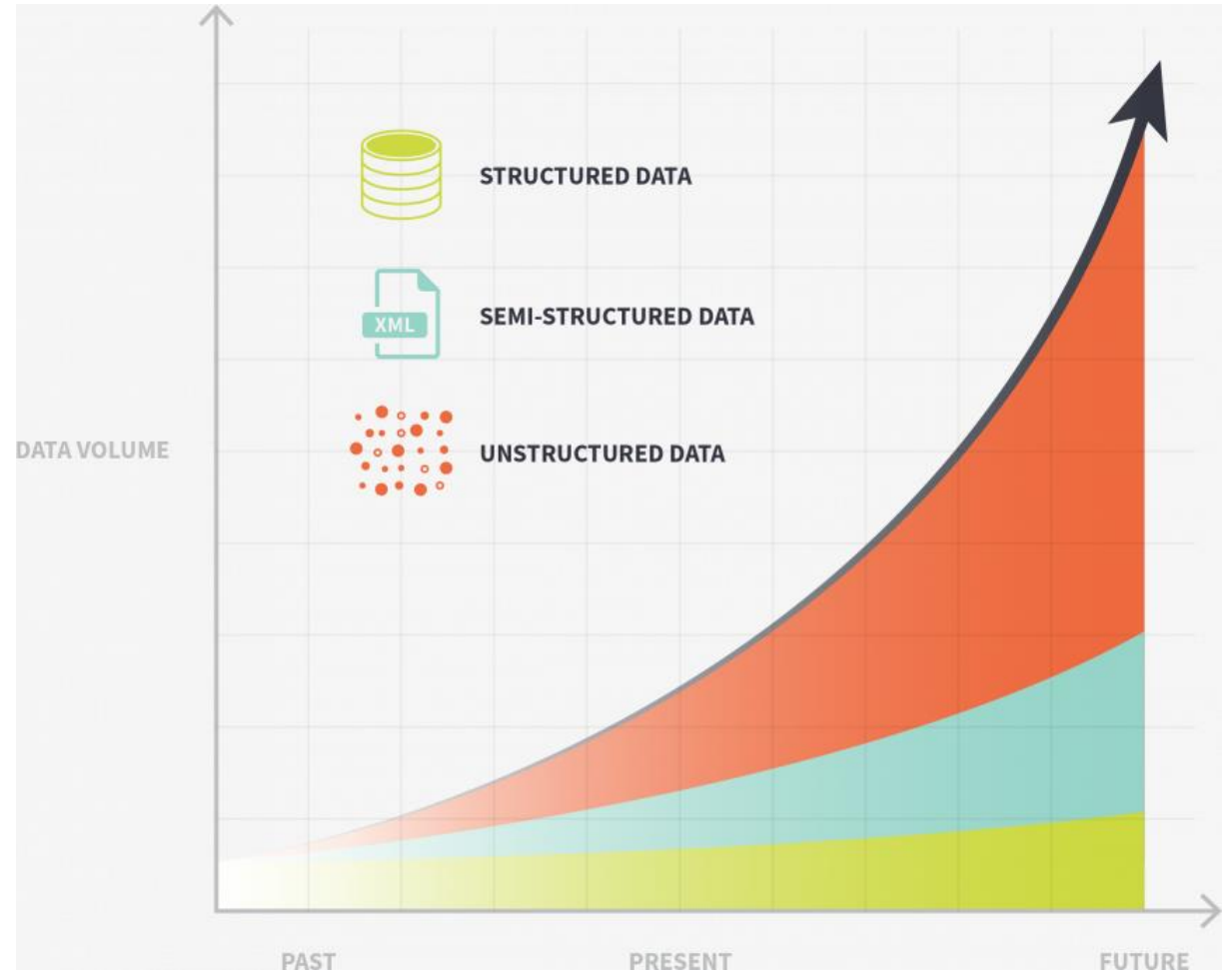
정의된 데이터 모델이 없거나, 미리 정의된 방식으로 정리되지 않은 데이터로 텍스트, 오디오, 이미지, 비디오가 대표적입니다.

강남역 맛집으로 소문난 **강남 토끼정**에 다녀왔습니다.
회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 🍷
다만, 강남 토끼정이 강남 썬썬버거 골목길로 쪽 올라가야 하는데
다들 썬썬버거의 유혹에 넘어갈 뻔 했답니다 🐰



텍스트 데이터의 중요성

- SNS 등장 이후 텍스트 데이터 양이 기하급수적으로 증가
- 머신러닝, 딥러닝 기술의 발전으로
텍스트 데이터 처리의 정확성이 높아짐
- STT(Speech to Text) 로 여러 형태의
비정형 데이터(오디오, 비디오)가
텍스트 형태로 변환되어 분석에 활용
- 다양한 활용 분야 : 소비자 트렌드, 감성분석,
문서요약, 분류



텍스트 데이터 분석 방법

■ 분류

- 문서의 제목/내용을 기반으로 미리 정해진 카테고리 값으로 분류
- 지도 학습(supervised learning) 사용 예) 뉴스 : 스포츠, 경제, 정치

■ 감성 분석

- 문서/텍스트로부터 '긍정', '부정', '중립'의 감성을 추출하는 방법

■ 의도 분류, 이메일 분류(Intent and Email Classifier)

- 질문의 의도 파악(분류)
- 제목 및 텍스트 기반으로 이메일 자동 분류

■ 설문 응답 분류(Survey Feedback Classifier)

- 설문 조사 응답을 고객 지원, 사용 용이성, 기능 및 가격과 같은 범주로 자동 분류

텍스트 데이터 분석 방법

■ **키워드 추출(Keyword Extractor)** 텍스트에서 가장 많이 사용되고 가장 중요한 키워드를 추출

■ **NER(Named Entity Recognition)**

- 문장에서 엔티티 이름을 식별하는 방법
- 예:철수[인명]는 서울역[지명]에서 영희[인명]와 10시[시간]에 만나기로 약속하였다.

■ **트렌드 분석**

- 어떤 주제(제품, 이슈, 인물 등)에 대해 단어/구문 단위로 어떤 변화가 있는지 분석하는 방법
- 실시간으로 데이터가 게재되는 소셜네트워크서비스(SNS)로부터 데이터를 수집하여 분석

■ **군집화**

- 비슷한 주제의 문서끼리 그룹핑을 하는 방법
- 비지도 학습 (unsupervised learning) 사용

■ **연관어 분석** 같이 출현하는 단어 간의 관계를 분석

키워드 추출(Keyword Extractor)

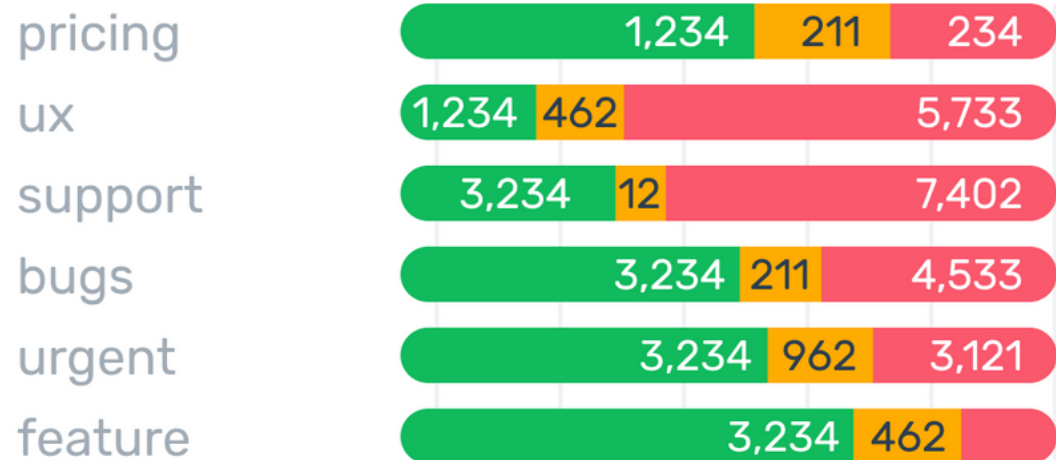
Keyword Cloud

qualifier



apple macbook pro
apple macbook air
solid state drive
intel core i5
refurbished dell xps
refurbished apple macbook
core processor laptop pc
touch screen windows gb ram
gb gb memory gb ssd laptop
intel core i7
dell inspiron

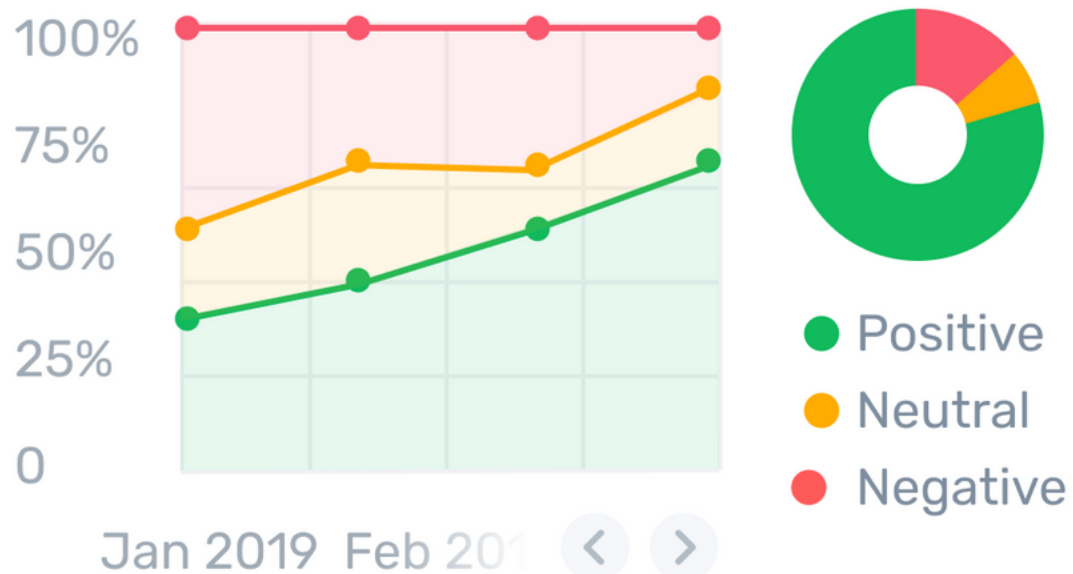
Sentiment by Topic



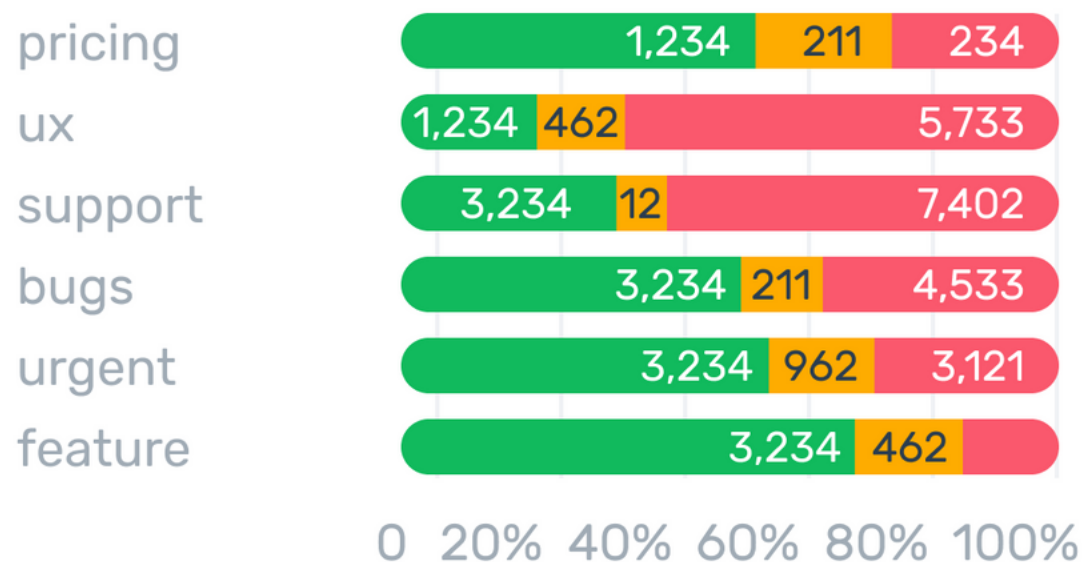
0 20% 40% 60% 80% 100%

감성분석 (Sentiment Analyzer)

Sentiment over time

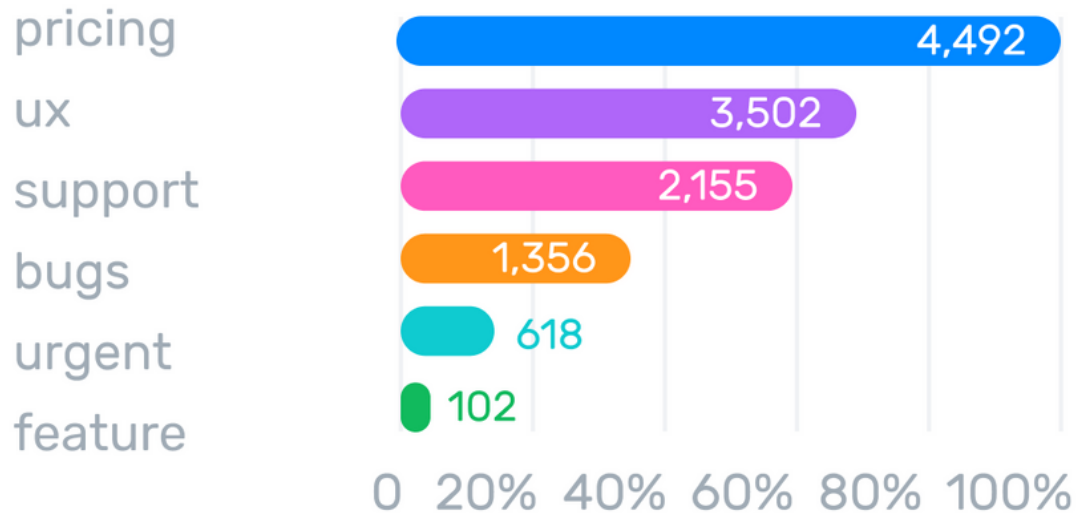


Sentiment by Topic

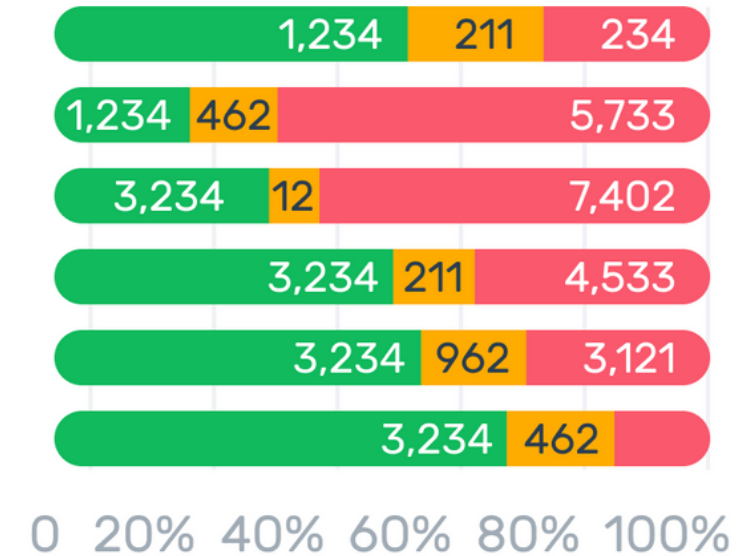


설문응답분류기 (Survey Feedback Classifier)

Samples by Topic

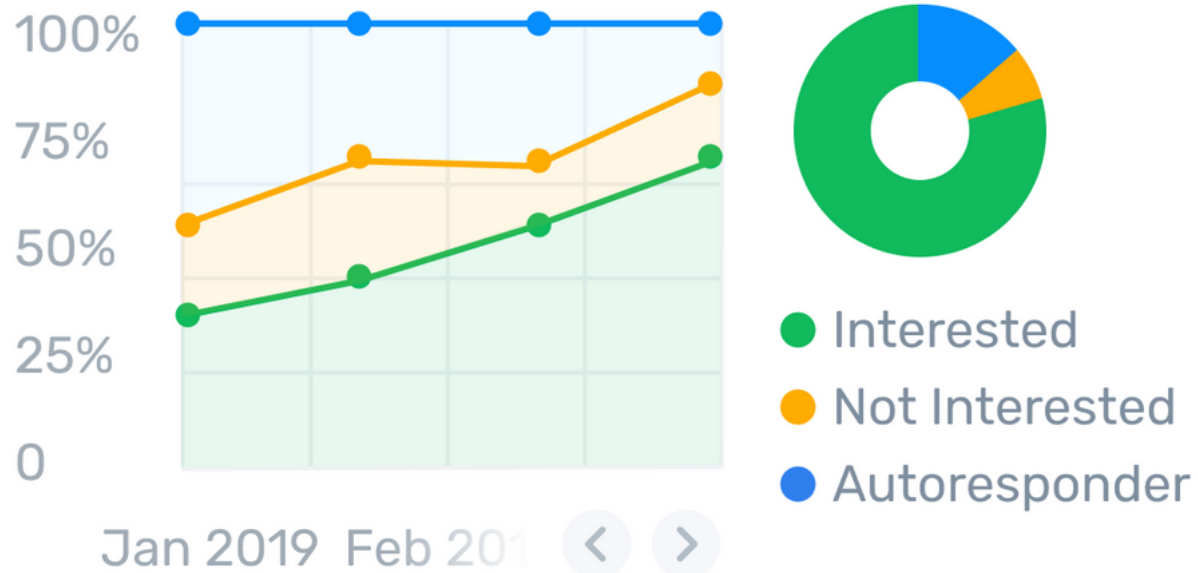


Sentiment by Topic

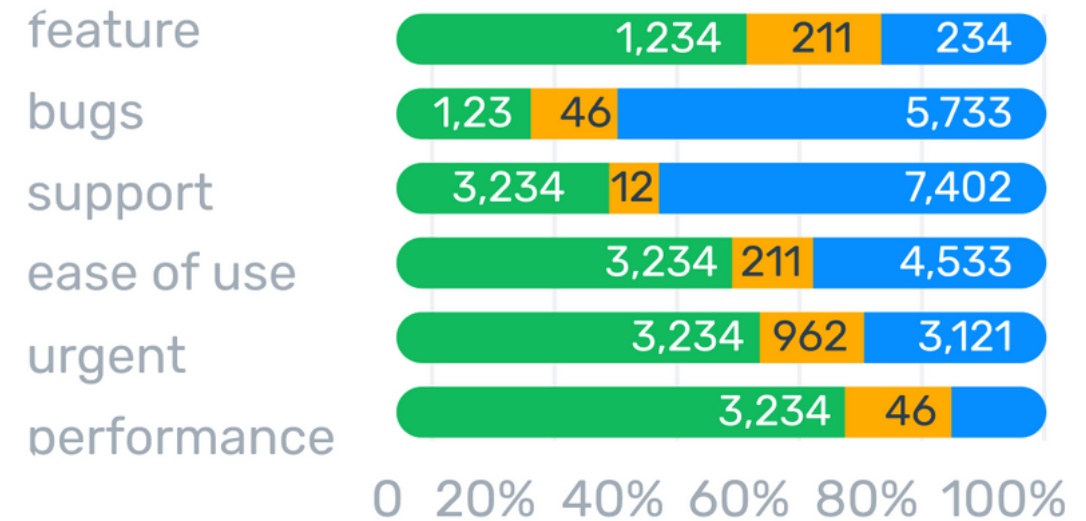


이메일 응답 분류 (Email Response Classifier)

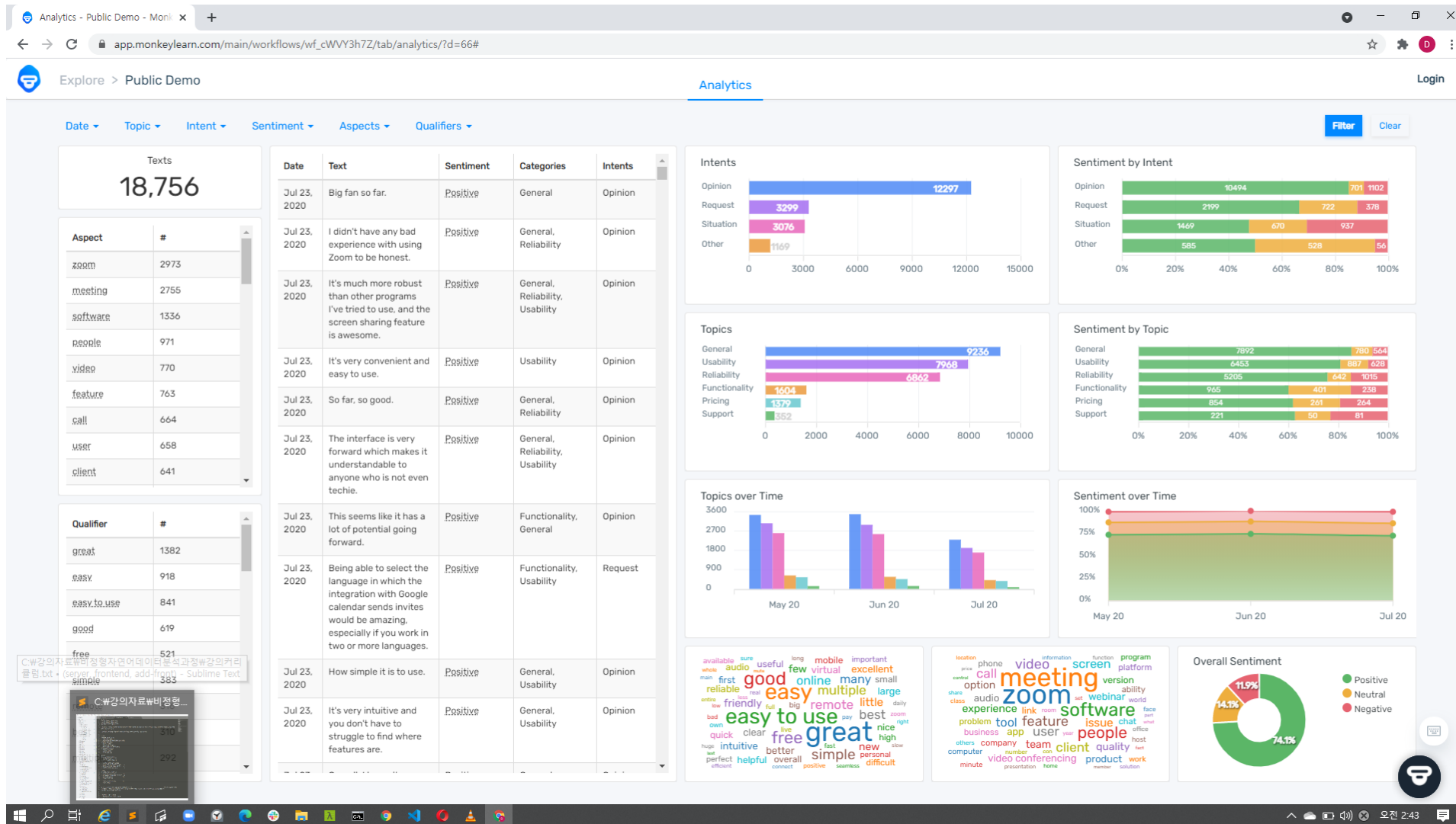
Interest over time



Interest by Topic



텍스트 데이터 분석



텍스트 데이터 분석 프로세스

START 문제정의 해결하고자 하는 문제의 명확한 이해가 필요

1. 데이터 준비

API 호출	Open API(Rest API)를 제공하는 서비스에 Open API를 호출
웹 크롤링	웹 상에 존재하는 콘텐츠를 수집하는 작업 (파싱 및 Selenium 통한 조작)
기타	오프라인 수집, 이미지 및 파일에서 추출, etc..

2. 데이터 전처리

텍스트 마이닝 과정에서 시간이 가장 많이 소요되는 작업

데이터 정규화	표현 방법이 다른 단어들을 통합 (Ex) 예시입니달 ㅎㅎ → 예시입니다 ㅎㅎ
데이터 분리	데이터를 특성에 따라 분리할 필요가 있을 경우에 진행
형태소 분석(토큰화)	일정한 의미가 있는 가장 작은 말의 단위로 변환 (품사태깅)
개체명 인식	이름을 가진 개체(named entity)로 인식 (Ex. 소희 - 사람, 동국대 - 조직)
원형 복원	형용사/동사의 표현형 → 원형 (Ex. 쉬고싶다, 쉬고싶은 → 쉬다) 어간/어미 → 표현형으로 활용 (Ex. 보다, 보니, 보고 → 보-)
불용어 제거	조사, 접미사 - 나, 너, 은, 는, 이, 가, 하다, 합니다 등
단어 빈도 분석	불용어 및 빈출어의 제거 여부 & 필요한 단어들의 올바른 추출 확인

3. 데이터 분석

동시 출현 분석	문서 요약	텍스트 생성
키워드 추출	군집화	네트워크 분석
단어 임베딩	감성 분석	토픽 분석

4. 분석 결과 시각화

히스토그램	네트워크 다이어그램	덴드로그램
테이블	워드 클라우드	히트맵

5. 보고

분석 결과 비교/해석
분석 결과 보고
COMPLETE

Big Data &
Data Mining Lab.

텍스트 데이터 분석 활용 사례

SR1. 감성 분석 활용 사례

“구매후기 한 줄에 고객의 이런 속마음이” 마케팅 난제, 속 시원히 풀어주는 분석

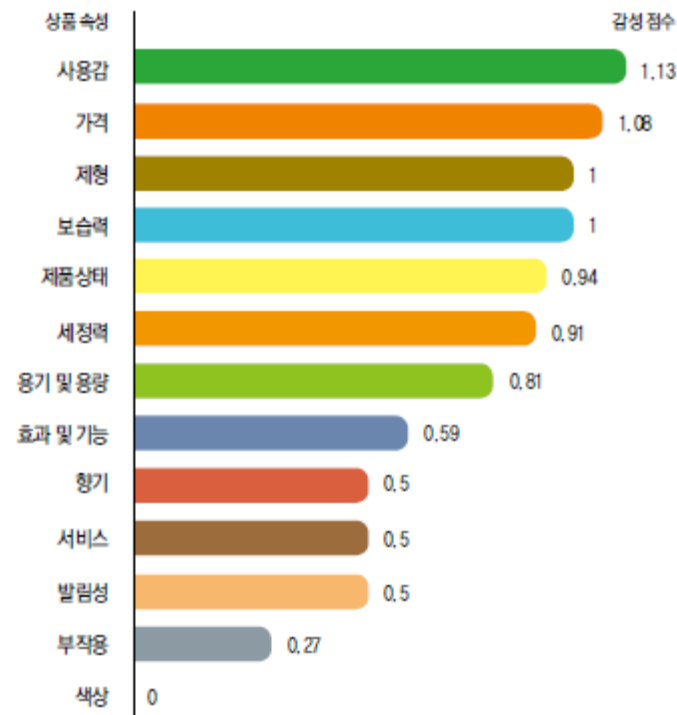
261호 (2018년 11월 Issue 2)



Article at a Glance

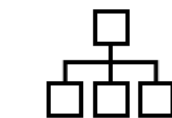
최근 ‘글에 내재해 있는 사람들의 주관적 태도나 감성을 추출해 내는 분석 기법’인 ‘감성 분석’에 대한 관심이 높아지고 있다. 감성 분석은 소셜미디어와 같은 웹사이트/매체에서 정보를 수집하는 ‘데이터 수집’ 단계, 수집된 정보에서 텍스트 작성자의 주관에 드러난 부분만을 걸러내는 ‘주관성 탐지’ 과정, 마지막으로 ‘주관성의 극성’이나 ‘정도’를 측정하고 분류하는 과정으로 나뉠 수 있다. 대표적 성공 사례로 초코바 스니커즈의 소비자 감성 변화에 따른 가격변동 마케팅, 국내 유명 화장품 브랜드 에뛰드하우스의 감성 분석 등을 꼽을 수 있다. 감성 분석에 성공하기 위해서는 적용 분야별 특성을 살린 사전을 구축하고, 데이터 수집 전략을 세우며, 다른 데이터와 연계해 다양한 분석을 수행할 수 있어야 한다.

그림7 A상품에 대한 속성별 감성 분석 결과



텍스트 데이터 분석 활용 사례

뉴스빅데이터 분석 서비스, BIGKinds



정형화된 데이터

비정형 텍스트를 분석이 가능한 정형화된 데이터로 바꾸어, 사회현상을 분석할 수 있는 기초 자료 제공



빅데이터화

1990년부터 현재까지 54개 매체의 약 7천만건 뉴스 콘텐츠를 빅데이터화



가치 있는 정보

한번 읽고 버려지는 하루살이 정보인 뉴스 콘텐츠를 축적해 분석할 수 있는 정보로

Thank you