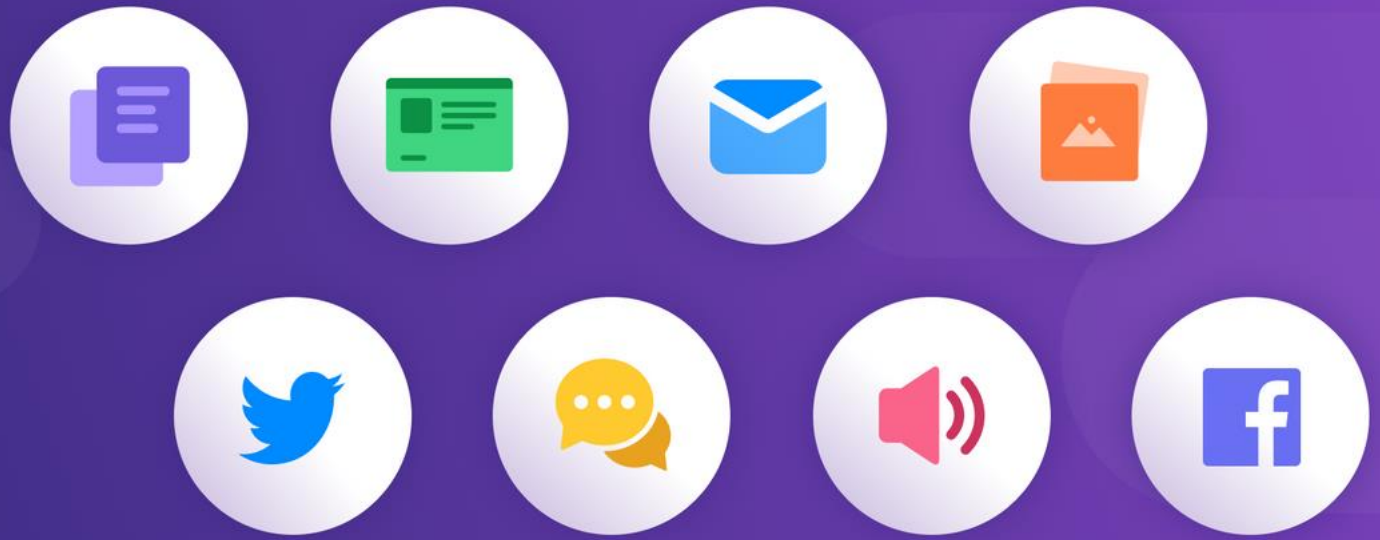
비정형 데이터 분석



박경규

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

관계형 데이터베이스(RDBMS)

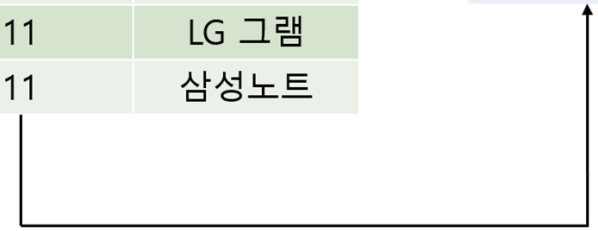
데이터는 행(row)에 저장되고, 속성은 열(column)로 표현하며, 테이블의 행과 행이 연결되는 관계를 맺을 수 있습니다.

< 주문 테이블 >

주문 번호	고객 번호	주문 상품
1	11	Hp 노트북
2	22	맥북
3	11	LG 그램
4	11	삼성노트

< 고객 테이블 >

고객 번호	고객 이름	고객 지역
11	노아	부산
22	두루미	서울



■ SQL(Structured Query Language)

SQL을 통해 RDBMS에서 데이터를 검색하고, 추가하고, 업데이트하고, 삭제하는 작업 등 데이터를 관리
Ex) SELECT customer_no, order_product WHERE order_no = 1

■ 트랜잭션(transaction

데이터베이스 관리시스템(DBMS)에서 하나의 작업의 단위
원자성 (Atomicity), 일관성 (Consistency), 격리성 (Isolation), 지속성 (Durability)

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 토끼정에 다녀왔습니다.
회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요
다만, 강남 토끼정이 강남 쉑쉑버거 골목길로 쭉 올라가야 하는데
다들 쉑쉑버거의 유혹에 넘어갈 뻔 했답니다

NoSQL

전형적인 데이터베이스 시스템에서 찾을 수 있는 행과 열로 이루어진 테이블 형식 스키마를 사용하지 않는 비관계형 데이터베이스는 저장되는 데이터 형식의 특정 요구 사항에 맞게 최적화된 스토리지 모델을 사용합니다.

■ 문서 데이터 저장소

Key	Document
1001	{ "CustomerID": 99, "OrderItems": [{ "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" }

■ 칼럼 형식 데이터 저장소

CustomerID	Column Family: Identity
001	**First name:** Mu Bae **Last name:** Min
002	**First name:** Francisco **Last name:** Vila Nova **Suffix:** Jr.

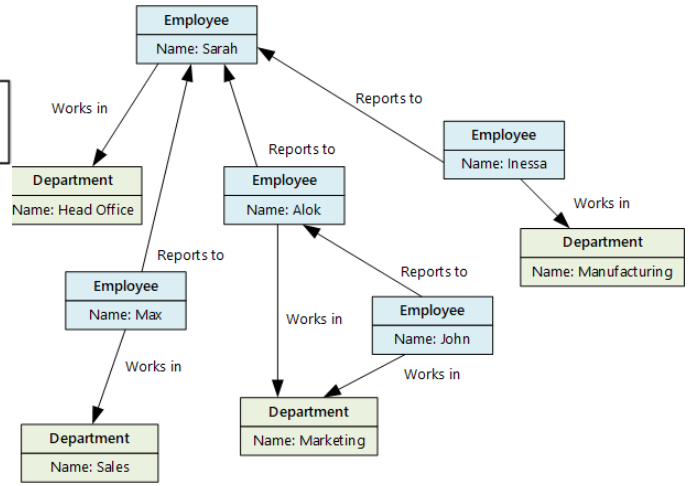
CustomerID	Column Family: Contact Info
001	**Phone number:** 555-0100 **Email:** someone@example.com
002	**Email:** vilanova@contoso.com
003	**Phone number:** 555-0120

■ 키/값 데이터 저장소

Key	Value
AAAAA	11010011110101001101011 11...
AABAB	10011000010110011010111 10...
DFA766	00000000001010101101010 10...
FABCC4	11101101101010101001011 01...

Opaque to data store

■ 그래프 데이터 저장소



■ 시계열 데이터 저장소

timestamp	deviceid	value
2017-01-05T08:00:00.123	1	90.0
2017-01-05T08:00:01.225	2	75.0
2017-01-05T08:01:01.525	2	78.0

■ 개체 데이터 저장소

path	blob	metadata
/delays/2017/06/01/flights.cscv	0XAABBCCDDEEF...	{created: 2017-06-02}
/delays/2017/06/02/flights.cscv	0XAADDCCDDEEF...	{created: 2017-06-03}
/delays/2017/06/03/flights.cscv	0XAEBBDEDDEEF...	{created: 2017-06-03}

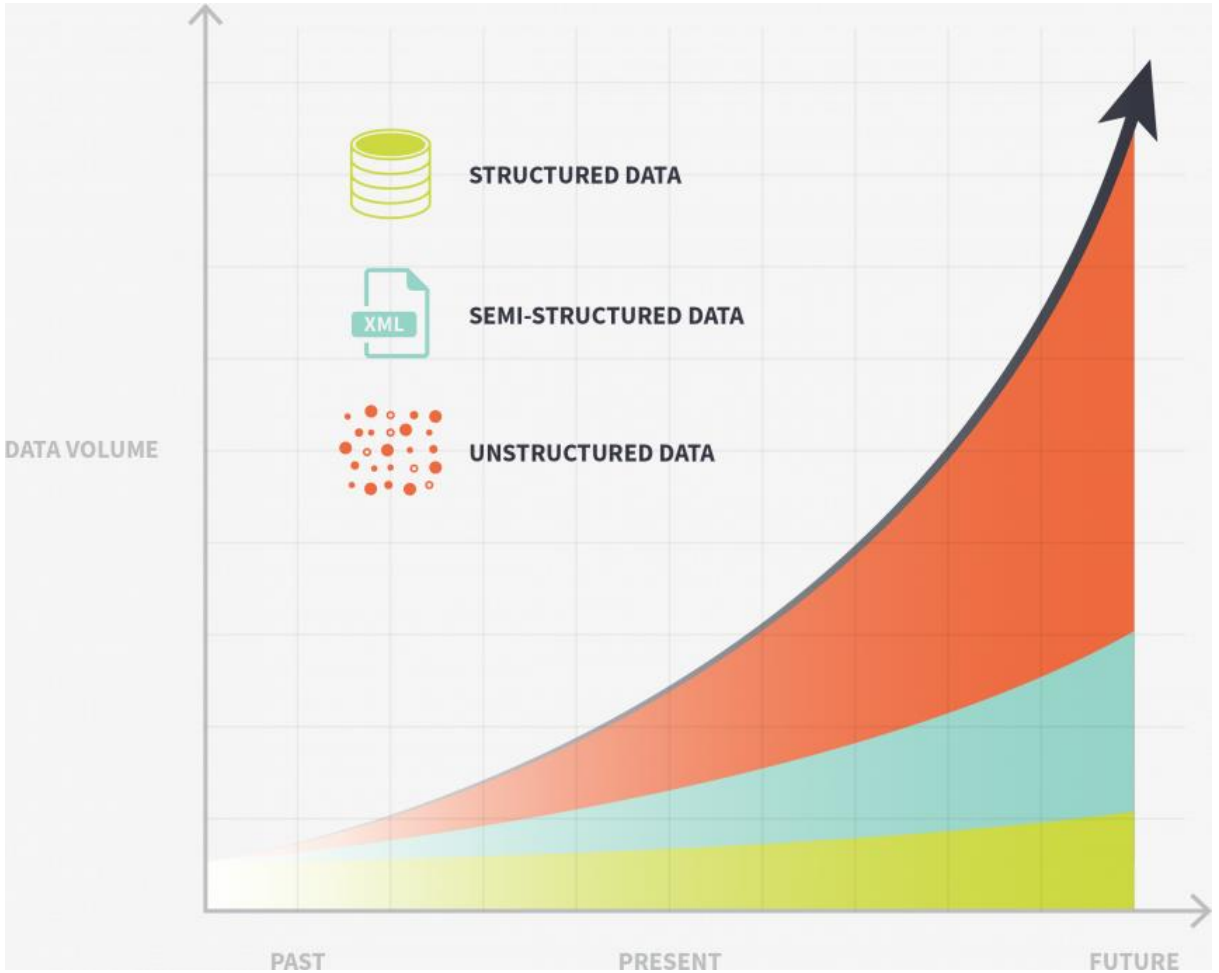
https://docs.microsoft.com/ko-kr/azure/architecture/data-guide/big-data/non-relational-data

NoSQL

https://www.youtube.com/watch?v=0buKQHokLK8

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**

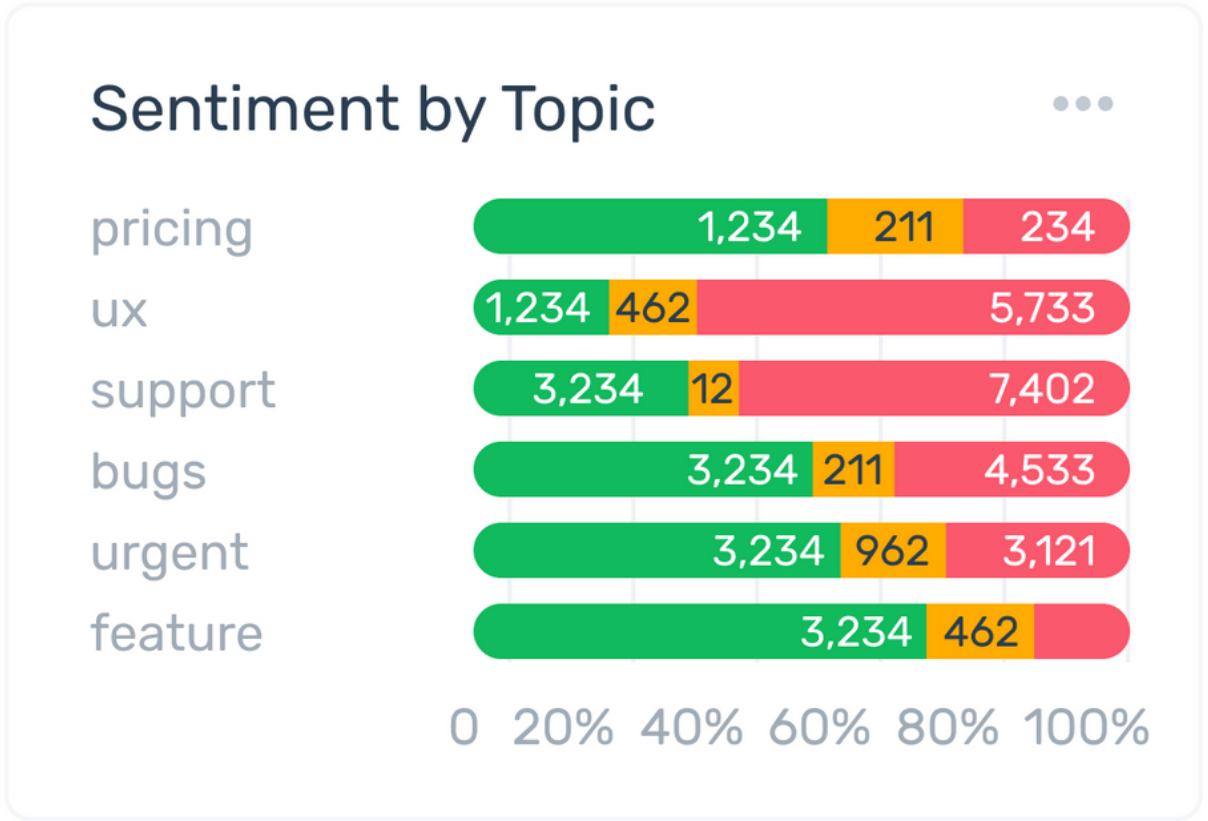
- 어떤 주제(제품, 이슈, 인물 등)에 대해 단어/구문 단위로 어떤 변화가 있는지 분석하는 방법
- 실시간으로 데이터가 게재되는 소셜네트워크서비스(SNS)로부터 데이터를 수집하여 분석

■ **군집화**

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

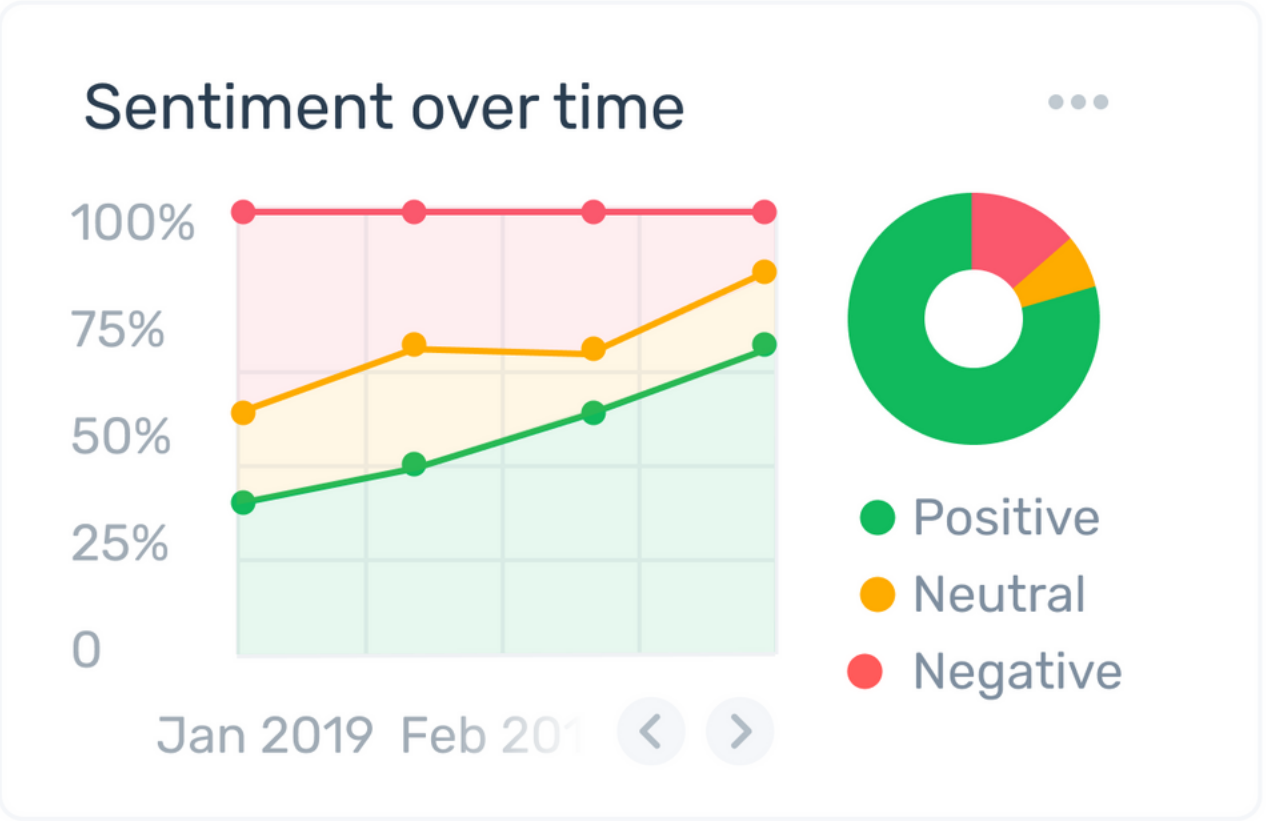
■ **연관어 분석** 같이 출현하는 단어 간의 관계를 분석

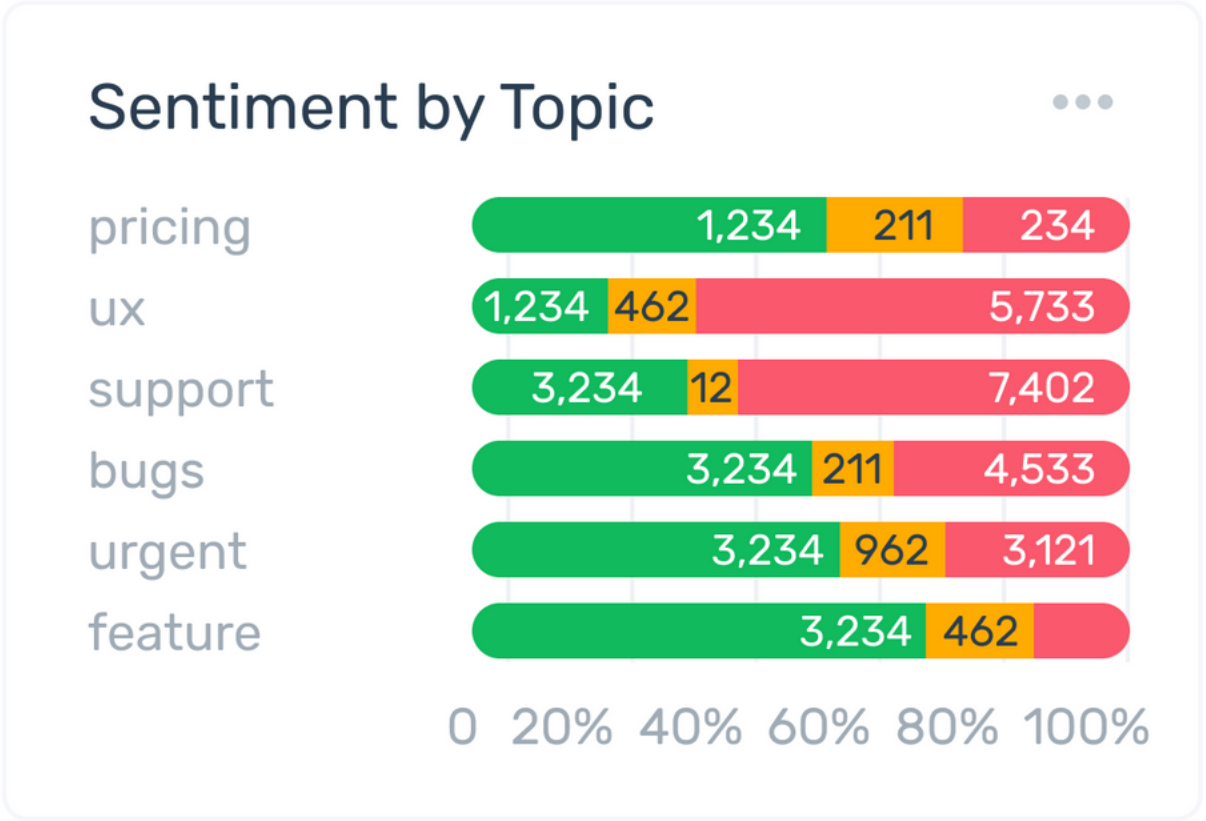
키워드 추출(Keyword Extractor)



https://monkeylearn.com/keyword-extractor-online/

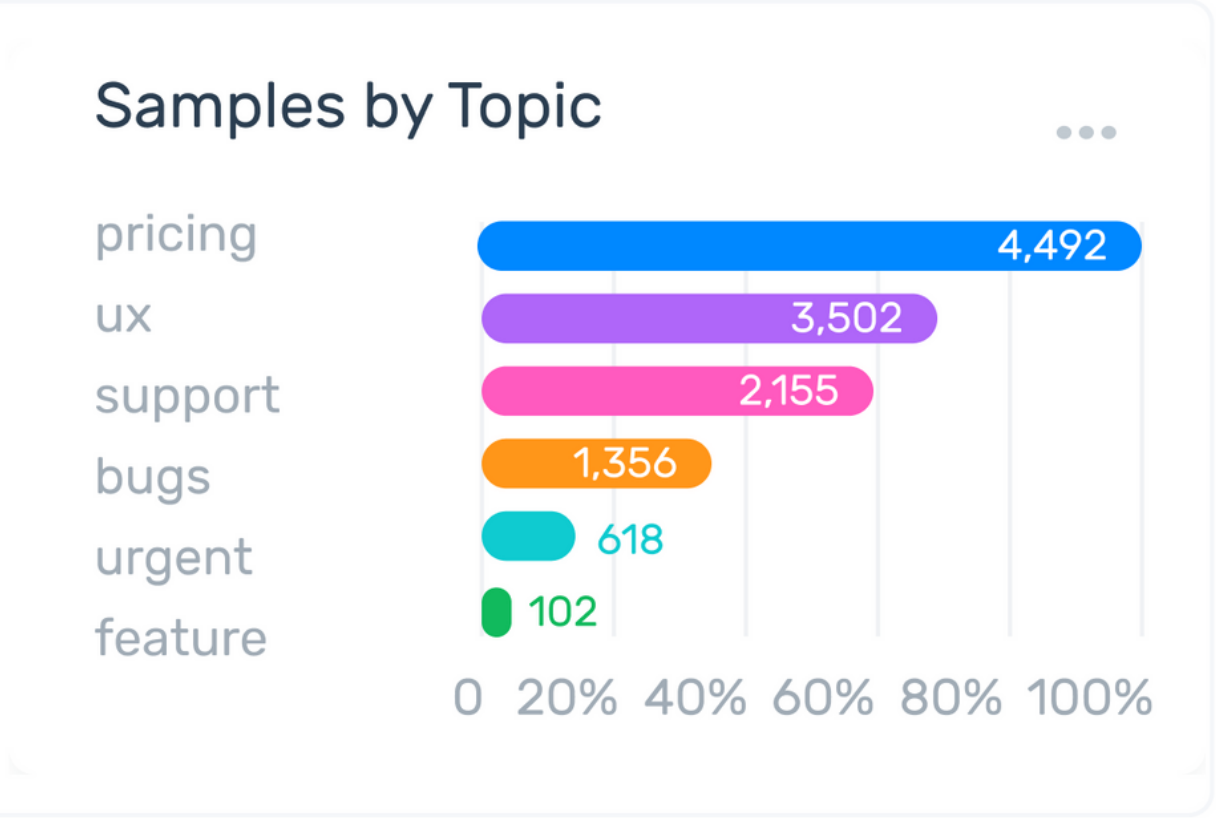
감성분석(Sentiment Analyzer)

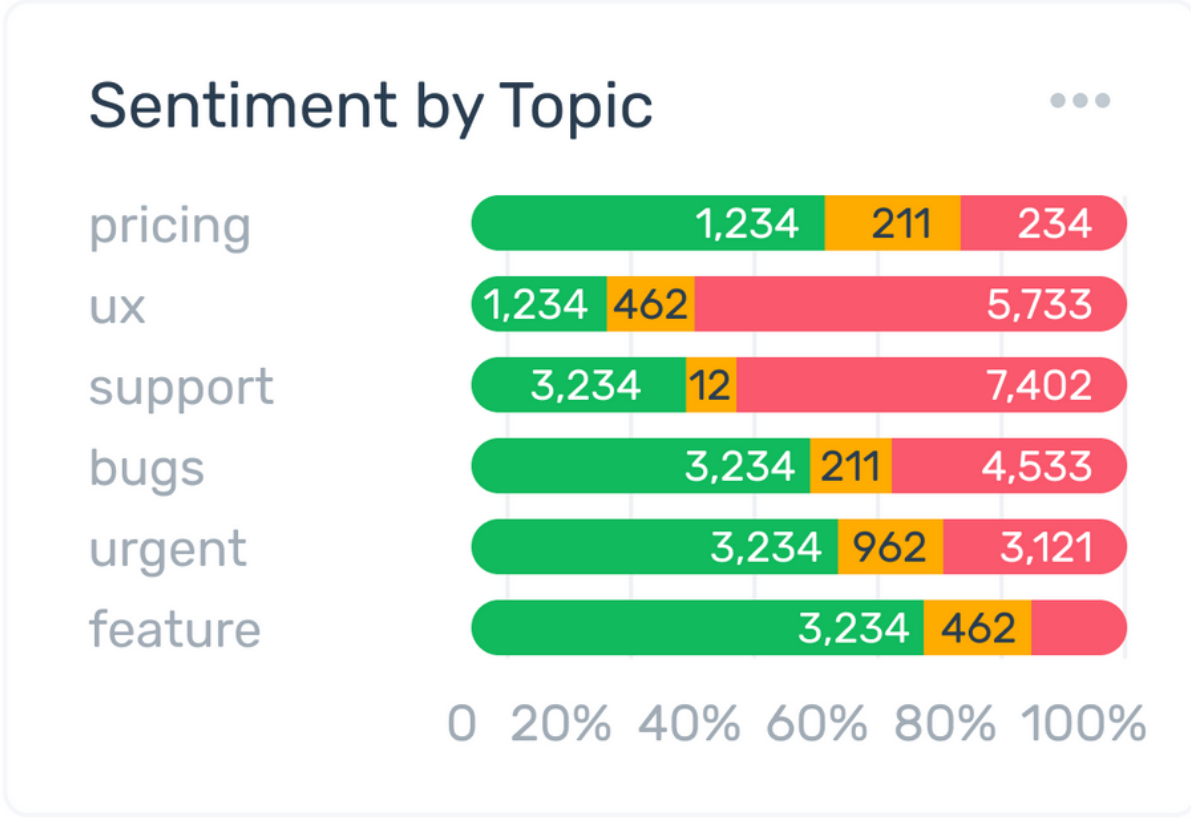




https://monkeylearn.com/sentiment-analysis-online/

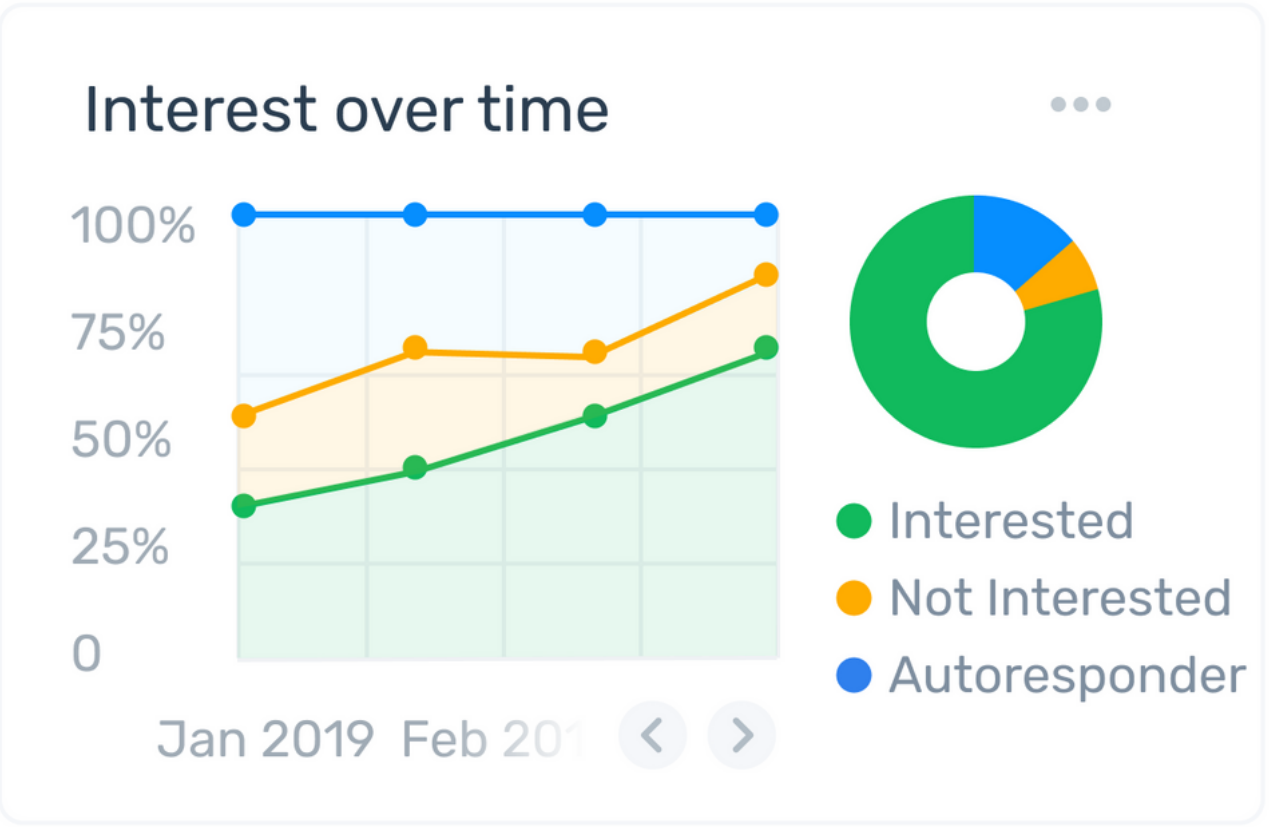
설문응답분류기(Survey Feedback Classifier)

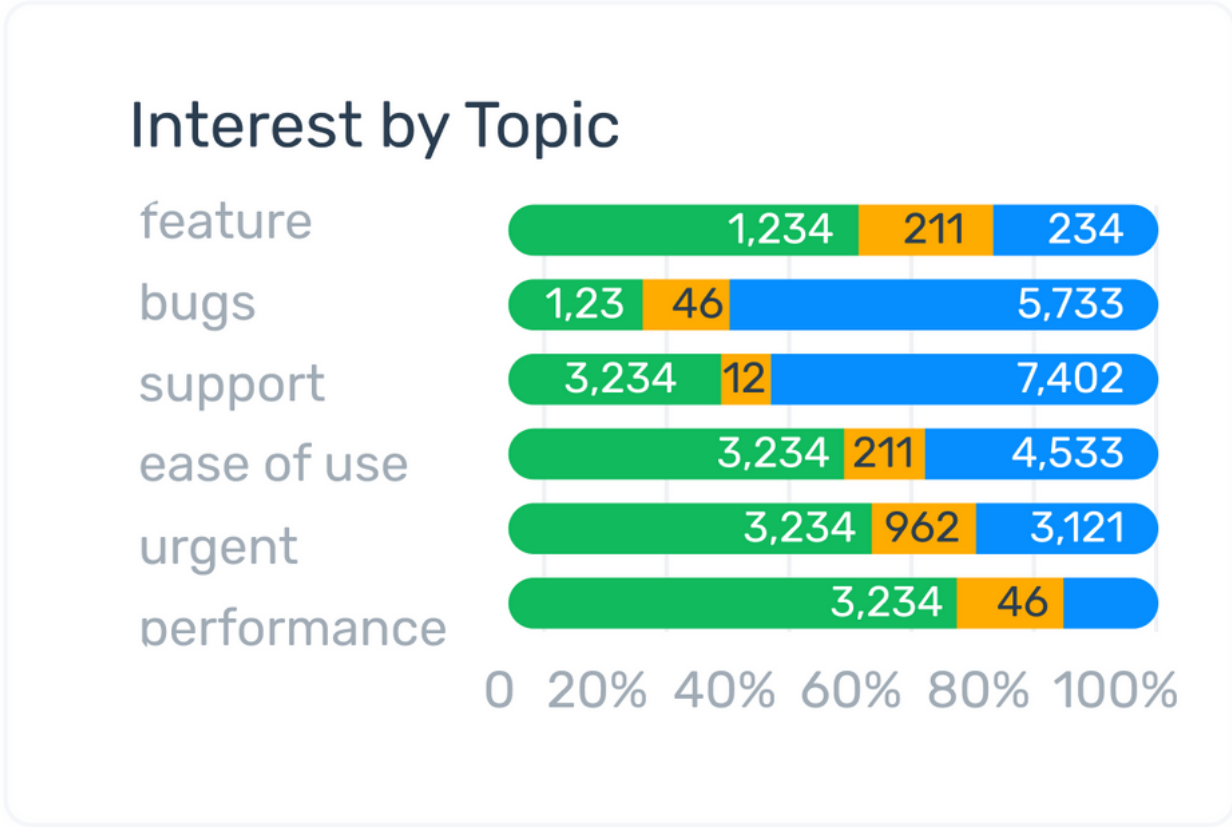




https://monkeylearn.com/nps-feedback-analysis-online/

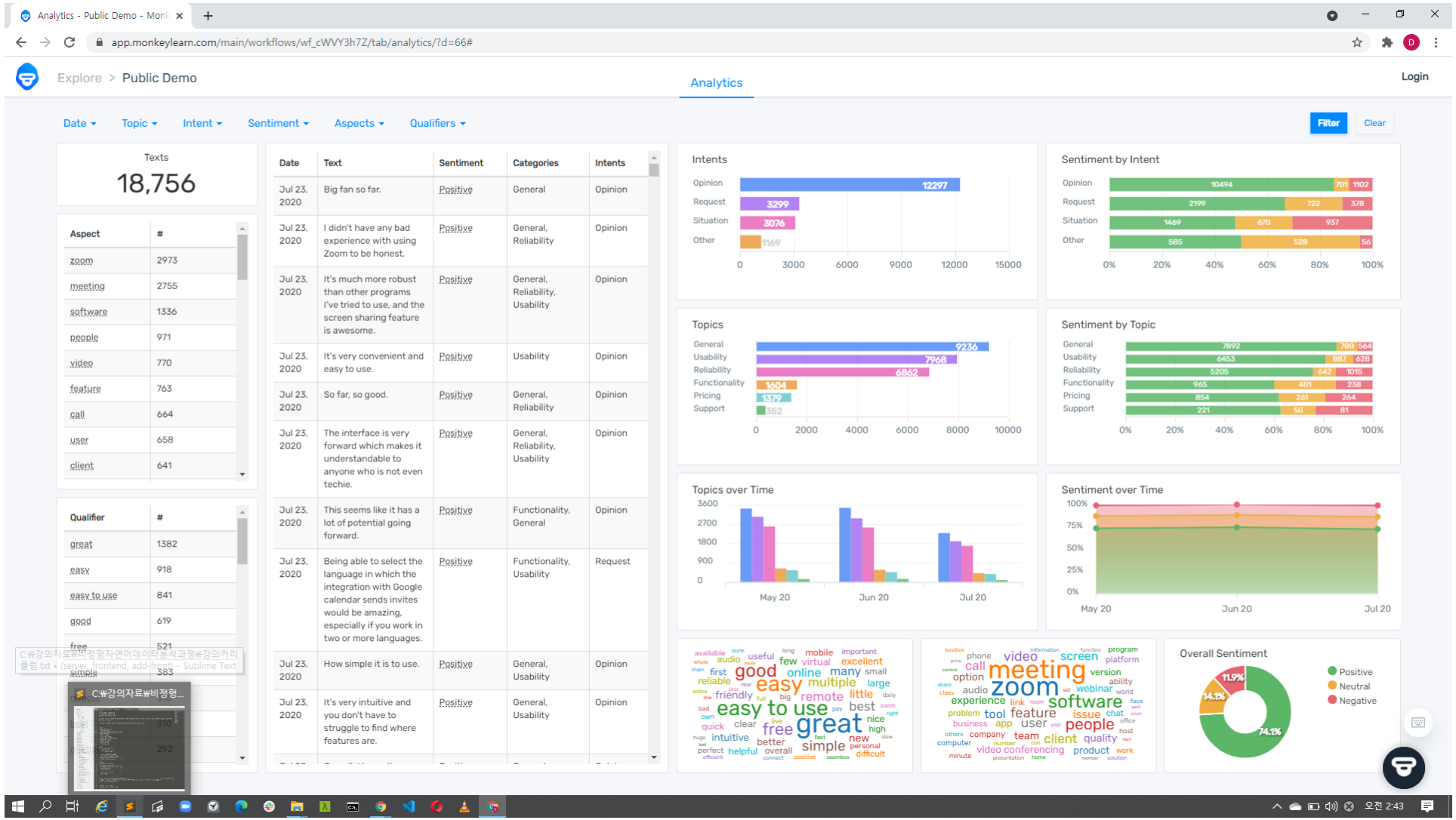
이메일 응답 분류(Email Response Classifier)





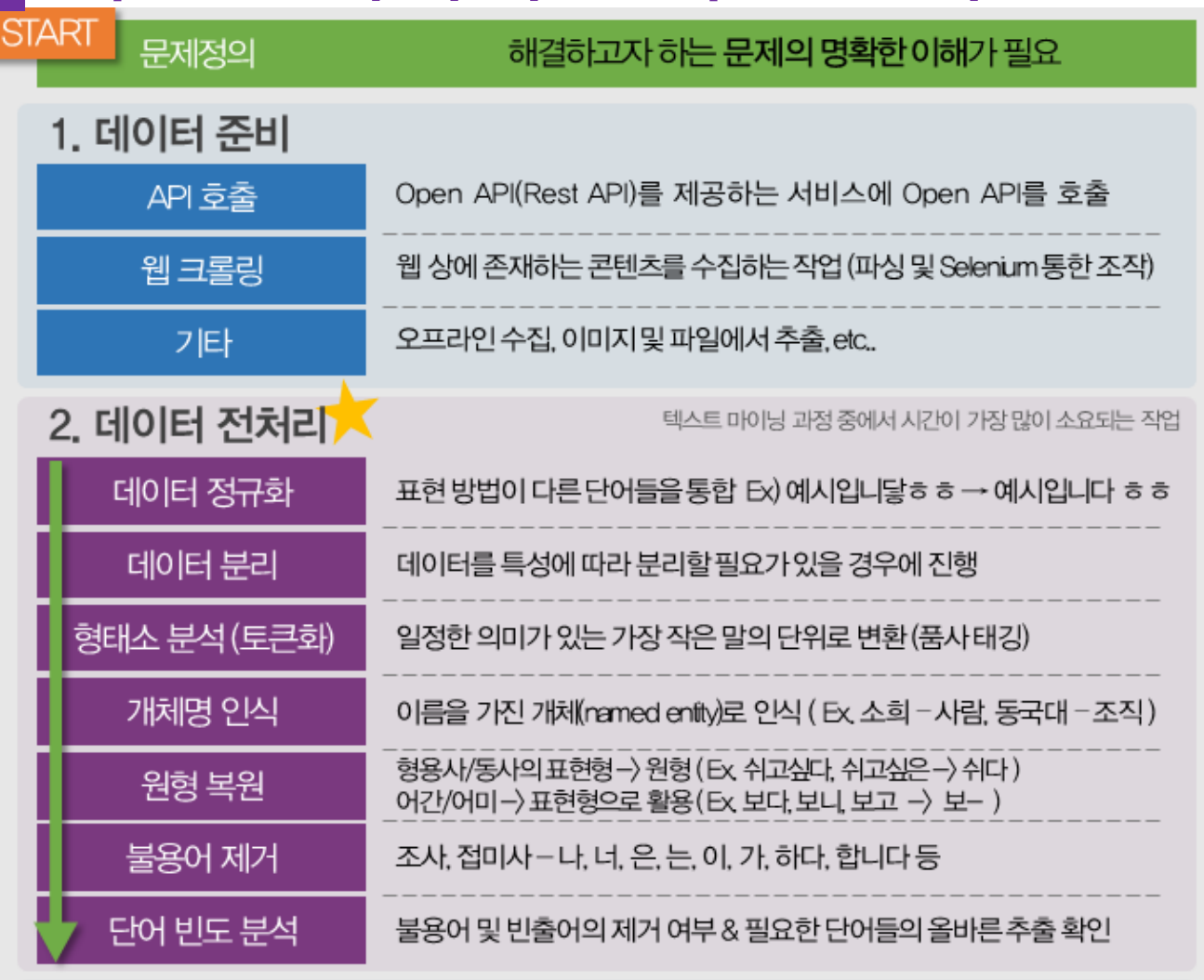
https://monkeylearn.com/email-response-classifier-online/

텍스트 데이터 분석



https://app.monkeylearn.com/main/workflows/wf_cWVY3h7Z/tab/analytics/?d=66#

텍스트 데이터 분석 프로세스



Big Data & Data Mining Lab.

https://bit.ly/3uqvSkA

텍스트 데이터 분석 활용 사례

SR1. 감성 분석 활용 사례

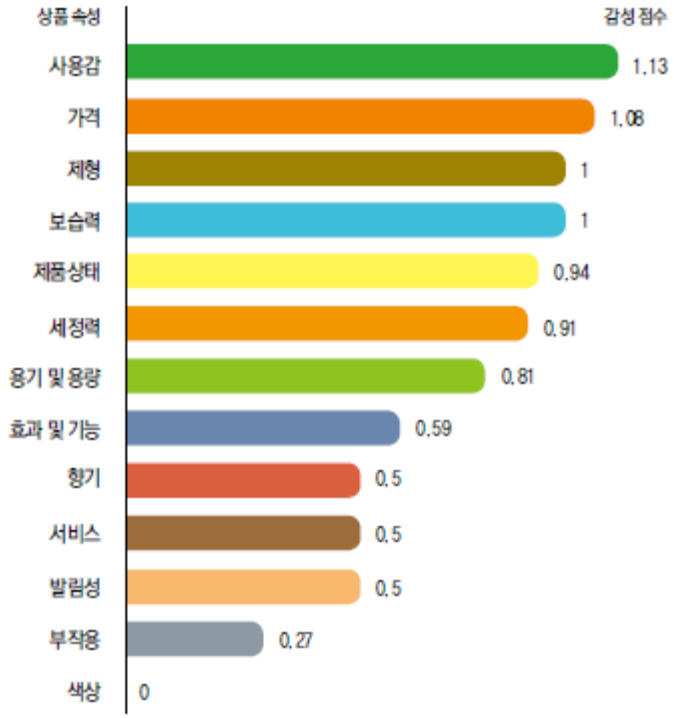
“구매후기 한 줄에 고객의 이런 속마음이” 마케팅 난제, 속 시원히 풀어주는 분석

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 주관이 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



https://dbr.donga.com/article/view/1202/article_no/8891

텍스트 데이터 분석 활용 사례

https://www.bigkinds.or.kr/v2/community/cases.do

Thank you

kgpark88@gmail.com