빅데이터 분석 솔루션

TEXTOM 소개서

CONTENTS

1 TEXTOM 히스토리

2013. 04 — The SCRM 서비스 시작(텍스톰의 전신)
- 1-mode, 2-mode 매트릭스 생성
- 사전 학습을 통한 불용어 처리 / 가정제 데이터 편집 및 저장
- 수집 데이터 및 보유 데이터 정제 / 데이터 수집
- 한국어 기반 텍스트 빅데이터 분석

2014. 06 — 텍스톰 1.0 버전
- 다국어 수집(영어, 중국어)
- TF-IDF
- N-gram

2015. 01 — 텍스톰 1.5 버전
- 사용자 정의 사전
- 전체 정제 및 타이틀 정제
- 가정제 데이터 편집

2015. 08 — 텍스톰 2.0 버전
- 에고네트워크 시각화 / 워드트리 시각화
- 서베이 / 정제모듈 MeCab-ko
- 개체명 인식

2018. 04 — 텍스톰 2.5 버전
- Customized Data / Visualization_ Upload Data
- News Data 채널 확장 / User Data 키워드 지정
- Sentiment Analysis / 수집, 분석 처리 속도 향상

2018. 12 — 텍스톰 3.0 버전
- UI, UX 변경 / 정제단계에서 용량 삭감
- 복수의 수집데이터 통합
- 키워드 필터링 / 분석품사 외국어, 숫자 추가
- 중복제거 / Window size

2019. 09 — 텍스톰 3.5 버전
- 수집데이터 관리 기능개선 / 수집량 시각화 제공
- 정제 단어 추천 / 엣지리스트 제공
- 상세품사 형태소 분석 / 형태소 분석결과 미리보기
- 감성분석 기본 학습데이터 제공 / 키워드 기반 감성분석 추가

2020. 03 — 텍스톰 4.0 버전
- 토픽모델링 기능 개선 / 감성분석 고도화
- 시각화 UI 디자인 개선 / 수집 처리 속도 향상
- 자동정제 기능 추가 / 요금제 개편
- 바로편집하기 기능 개선 / 교육용 수집 속도 향상

웹 환경에서 데이터를 수집하고 정제하며 다양한 분석을 처리할 수 있습니다

별도의 설치 없이, 회원가입 승인과 동시에 바로 사용할 수 있는 웹 환경의 솔루션입니다.
다양한 사이트에서 원하는 기간의 데이터를 수집한 후 텍스트마이닝 작업을 거쳐
매트릭스, 감성분석, 토픽모델링, 시각화 등 원하는 결과물을 산출할 수 있습니다.

- **실시간 대용량 데이터 수집**
  - + Web
  - + Portal
  - + 사용자 요청 사이트

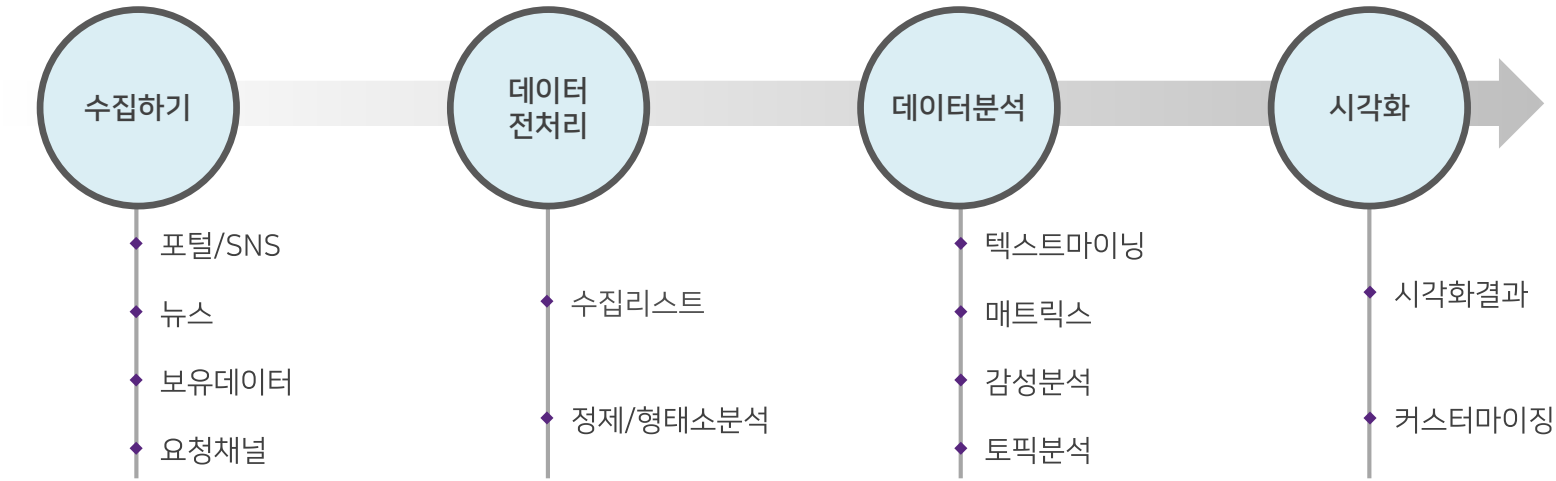
- **데이터 저장 및 정제**
  - + 분산저장기술
  - + 분산병렬처리
  - + Hadoop을 이용한 데이터 저장
  - + 국문, 영문, 중문 형태소 분석기

- **결과물 산출**
  - + 토픽모델링
  - + 매트릭스
  - + 엣지리스트
  - + 감성분석
  - + 단어빈도
  - + N-gram
  - + TF-IDF
  - + 연결중심성
  - + 개체명인식

- **시각화**
  - + 수집량 라인차트
  - + 워드클라우드
  - + 에고네트워크
  - + N-gram 네트워크
  - + 토픽모델링
  - + 트리맵
  - + 매트릭스차트
  - + 감성분석

사용이 쉽고 편리합니다

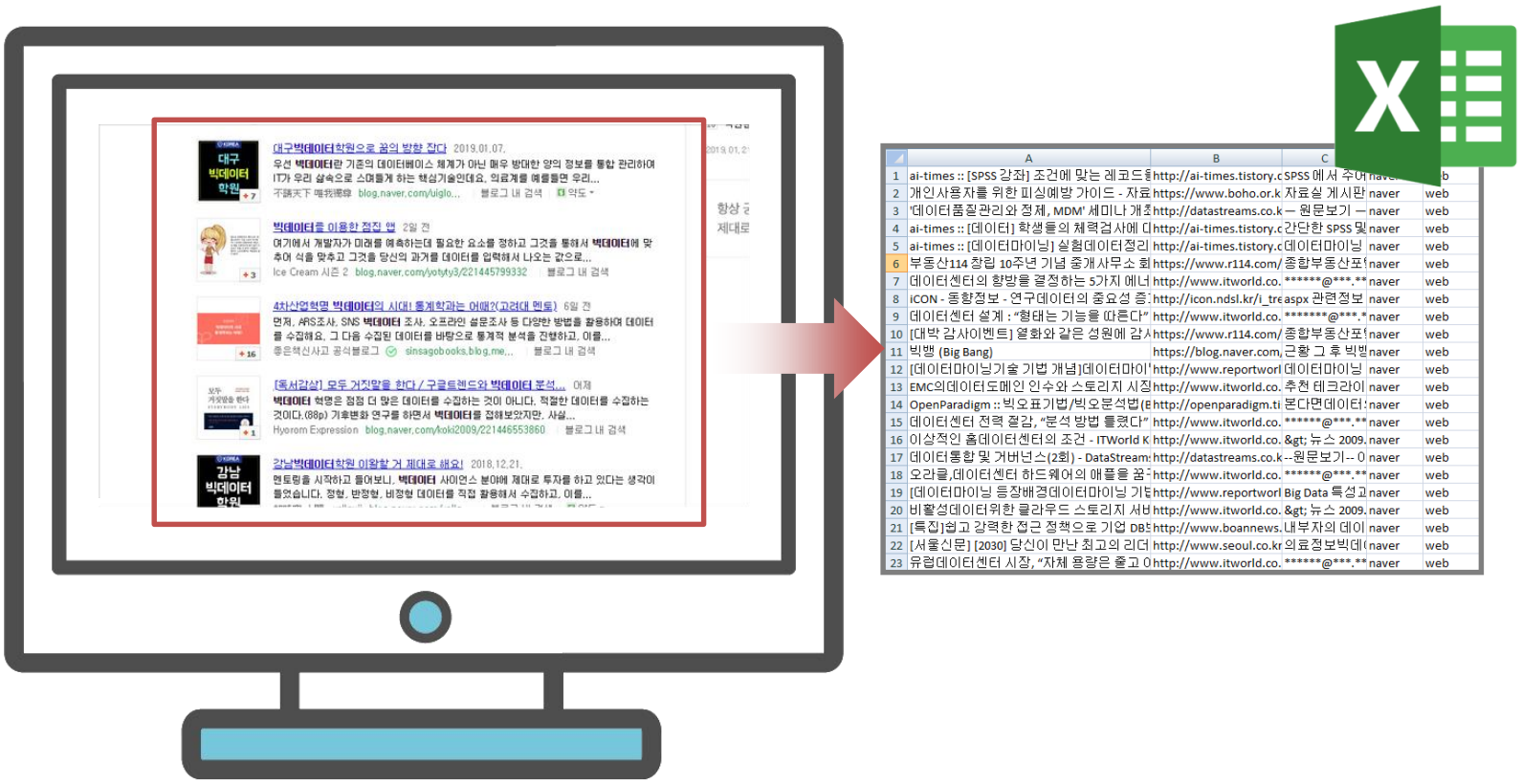
중학생부터 기업인, 연구자까지 사용의 폭이 넓고 다양합니다.
사용자 환경에 최적화한 UI, UX와 상세한 매뉴얼을 통해 누구나 쉽고 편리하게 이용할 수 있습니다.



웹 상의 데이터를 빠른 속도로 수집하여 데이터 셋(Data Set)을 제공합니다

요약문서 1건 당 0.5초의 수집속도를 자랑합니다.

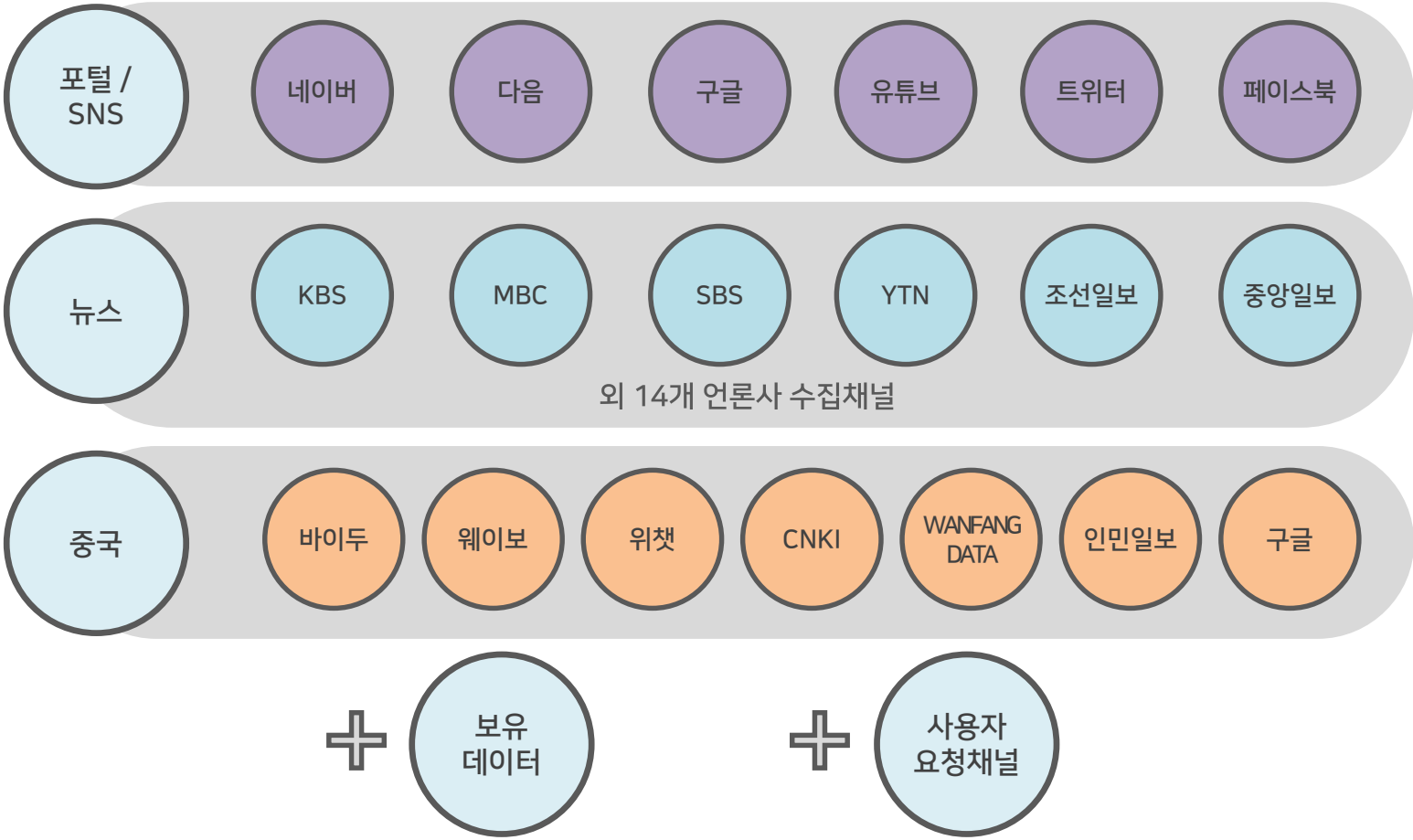
수집은 무료로 이용 가능하며, 수집된 데이터로 분석을 진행할 경우 원문데이터를 엑셀파일 또는 텍스트파일로 다운로드 받을 수 있습니다.



TEXTOM

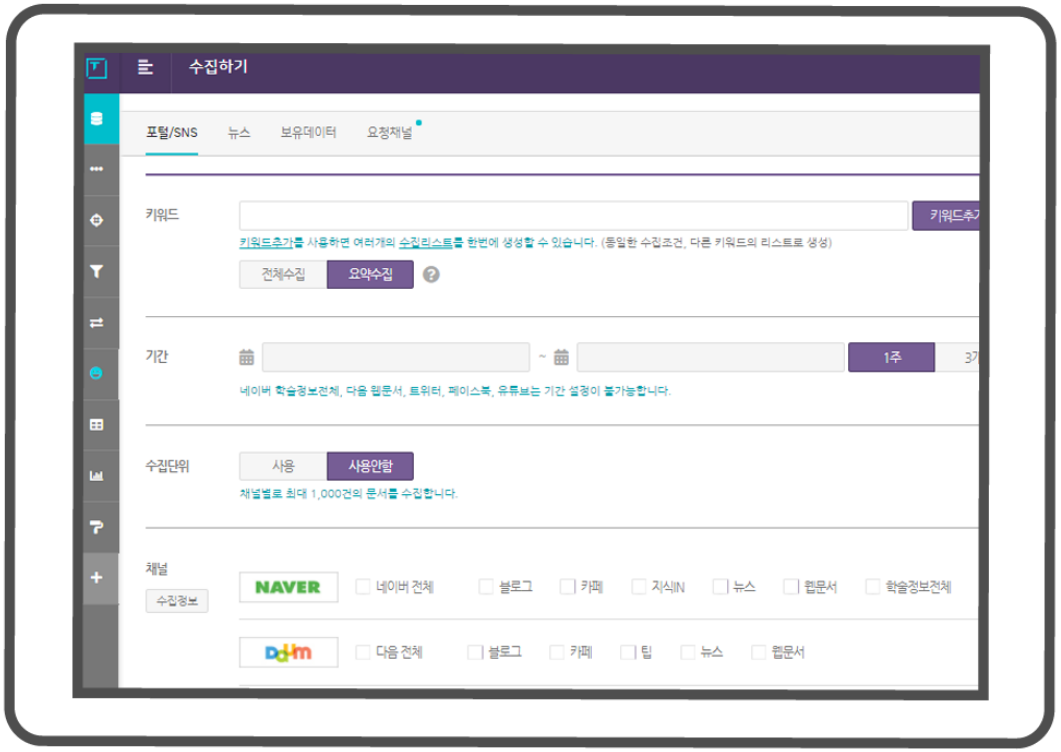
사용자의 입맛에 맞는 다양한 수집채널을 선택할 수 있습니다

포털사이트, 소셜미디어, 다양한 언론사의 데이터를 수집합니다.
보유데이터 업로드를 통한 분석 또한 가능하며, 사용자가 원하는 수집채널을 의뢰할 수도 있습니다.
중국 버전의 텍스톰을 통해 다양한 중문 채널의 데이터를 수집할 수 있습니다.



세부 설정기능을 통해 수집결과의 만족도를 높일 수 있습니다

제목, 날짜, 본문, URL 수집이 가능하며, 검색 연산자가 수집키워드에 반영되어 수집결과 데이터의 정확도를 높일 수 있습니다.
수집하고자 하는 데이터가 생성된 기간을 설정할 수 있으며, '수집단위' 기능을 통해 데이터 수집량을 조절할 수 있습니다.



◆ 키워드

입력한 키워드로 실제 각 채널에 나타나는 검색 결과가 수집됩니다.

◆ 기간

포털사이트의 경우 1991년부터 현재까지의 데이터를 수집할 수 있습니다.

◆ 수집단위

일, 주, 월, 년 단위 중 선택한 단위 당 최대 1,000건의 문서를 수집합니다.

◆ 채널

채널 하위에 있는 섹션(블로그, 카페, 뉴스 등)의 데이터를 선택할 수 있습니다.

분산파일 처리시스템 하둡(Hadoop)을 기반으로 대용량 파일 보관에 뛰어납니다

수집 · 정제 · 분석된 데이터의 저장과 관리를 위한 분산파일 시스템과 NoSQL 기능을 구현하여 대규모의 데이터를 유연하게 처리합니다.
데이터의 효율적인 선택과 실시간 분석을 위한 데이터 색인 기능을 제공합니다.

- Hadoop Distributed File System
  대용량 파일을 안전하게 저장하고, 처리하기 위한
  하둡 분산 파일 시스템

▶ 수집된 모든 데이터 1개월 보관

▶ 영구보존

국문 뿐만 아니라 영문, 중문 데이터 분석도 가능합니다

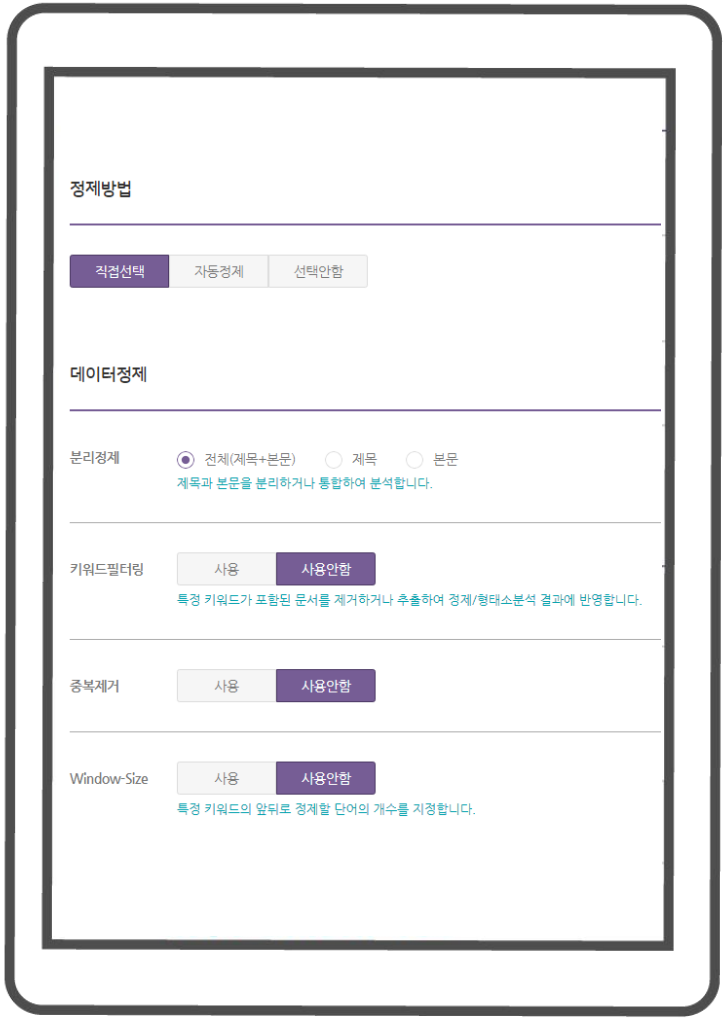
국문, 영문, 중문 형태소분석기를 적용하여 다국어 분석이 가능합니다.
국문의 경우 두 가지 형태소분석기 중에서 원하는 결과에 적합한 형태소 분석기를 선택할 수 있습니다.

- **Espresso K** 고유명사, 복합명사를 그대로 결과값에 반영합니다.
- **MeCab** 띄어쓰기와 상관없이 사전을 참조하여 어휘를 구분합니다.

[두 형태소분석기의 차이 살펴보기]

정교한 정제와 자동 정제 모두 가능합니다

사용자 분석 목적에 맞게 정제데이터를 생성할 수 있도록 다양한 전처리 설정 기능을 제공합니다.



정제방법

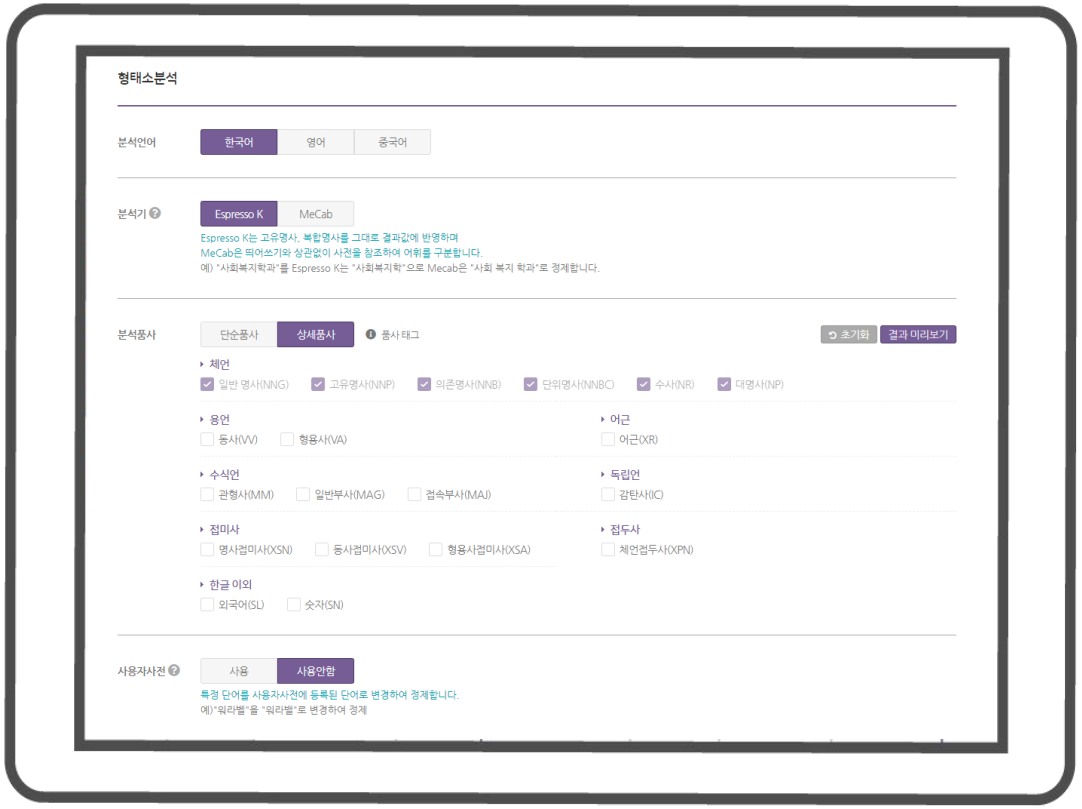
직접선택	이용자가 원하는 분석에 맞게 세부설정을 직접 고를 수 있습니다.
자동정제	텍스톰이 지정한 기본설정(제목+본문/MeCab/명사)으로 정제가 됩니다.
선택안함	이미 정제한 데이터 즉, 전처리가 필요하지 않은 데이터를 데이터분석 단계로 넘길 때 이용할 수 있습니다.

데이터정제

분리정제	수집된 데이터의 제목과 본문을 분리하거나 통합하여 정제할 수 있습니다.
키워드필터링	특정 키워드가 포함된 문서 자체를 정제데이터에서 제외하거나 특정 키워드가 포함된 문서만 추출하여 정제 결과에 반영할 수 있습니다.
중복제거	수집된 데이터에서 중복되는 URL이나 내용을 제외합니다.
Window-Size	특정 키워드의 앞뒤에 위치한 단어의 개수를 지정해 해당 개수까지의 단어만 정제 결과에 반영할 수 있습니다.

한국어 자연어 처리에 뛰어난 형태소분석 기능을 제공합니다

세밀한 분석품사 설정을 통해 정제데이터의 결과물이 뛰어납니다.

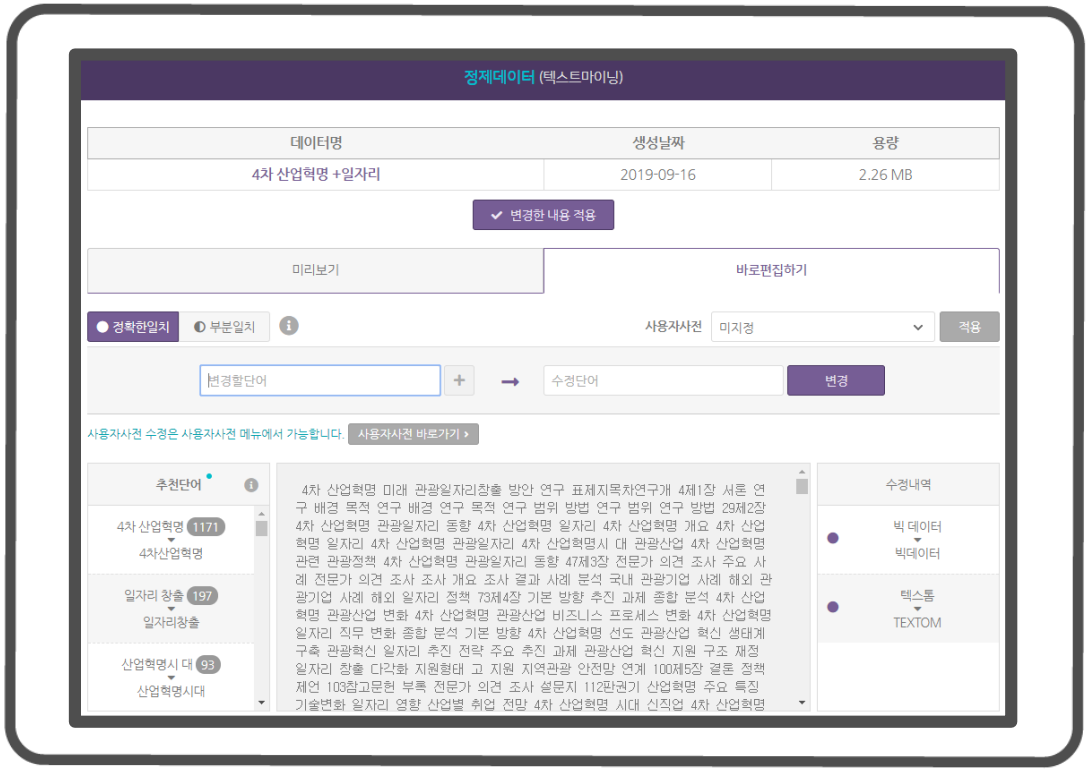


형태소분석

분석언어		한국어, 영어, 중국어 중에서 분석에 사용할 언어를 선택할 수 있습니다.
분석기		Espresso K, MeCab 중에서 원하는 결과값에 맞는 형태소분석기를 선택할 수 있습니다.
분석품사	단순품사	명사, 형용사, 동사, 외국어, 숫자 중에서 정제 결과에 반영할 품사를 선택할 수 있습니다.
	상세품사	체언, 용언, 어근, 수식언, 독립언 등 다양한 품사를 정교하게 선택하여 반영할 수 있습니다.
사용자사전		사용자가 사용자사전에 등록한 특정 단어들을 정제 데이터에 미리 반영할 수 있습니다. 동일 · 유사한 데이터를 반복 정제할 때 유용하게 사용할 수 있습니다.

데이터 편집 기능으로 2차 정제가 가능하여 결과물의 정확도를 높일 수 있습니다

사용자가 정제된 데이터를 더 정교하게 편집할 수 있어 결과물의 정확도를 높입니다.



- 자동 정제된 데이터의 결과값을 보고 여전히 남아 있는 불용어나 복합명사, 완전하지 않은 단어를 정제하는 작업을 진행할 수 있습니다.
- 웹 상에서 바로 편집하거나 정제 데이터를 다운받아 오프라인 환경에서도 편집한 후 업로드 할 수 있습니다.
- N-gram 기반의 확률 모델을 통해 단어쌍을 선별하여 편집할 단어를 추천합니다.
- 수정내역의 되돌리기, 정확한일치, 부분일치 등 이용자의 편의를 고려한 다양한 편집 기능을 지원합니다.
- 등록된 사용자사전 그룹을 적용하여 편집에 소모되는 시간을 최소화 할 수 있습니다.

텍스트마이닝의 다양한 결과값을 제공합니다

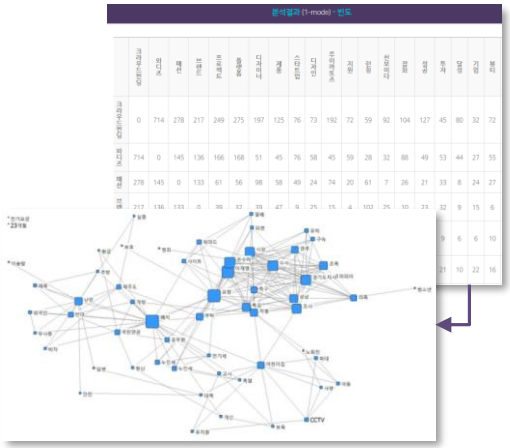
단어빈도, N-gram, TF-IDF, 연결중심성, 개체명인식 결과를 자동으로 산출합니다.

단어빈도	전체 문서 내에서 단어의 출현빈도가 높은 순서대로 단어와 빈도수를 표시합니다.
N-gram	두 개의 단어가 연쇄적으로 출현한 빈도수를 나타냅니다.
TF-IDF	문서 내에서 단어의 출현빈도를 나타냅니다. 값이 클수록 해당 문서 내에서 중요한 단어라고 해석할 수 있습니다.
연결중심성	다른 단어와 직접 연결된 정도를 나타냅니다.
개체명인식	자동 정제 결과에 따라 분리된 단어를 14개의 범주에 따라 분류합니다. (사람, 학문, 대상물, 기관, 지역, 문명, 날짜, 시간, 숫자, 사건/사고, 식물, 금속, 용어)

다양한 통계분석 프로그램과 연계하여 심층분석이 가능하도록 호환성 높은 데이터를 제공합니다

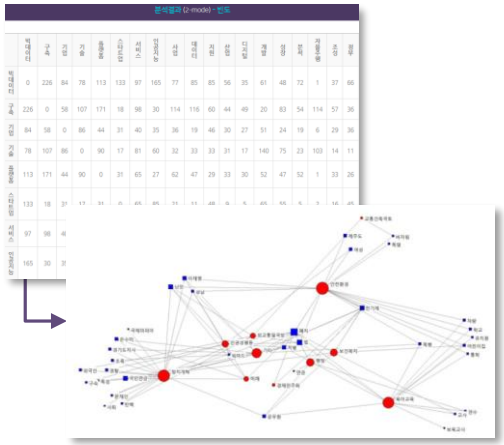
분석하고 싶은 단어를 선택하거나 단어 파일을 업로드 하면 결과값을 자동으로 산출합니다.

1-mode 분석



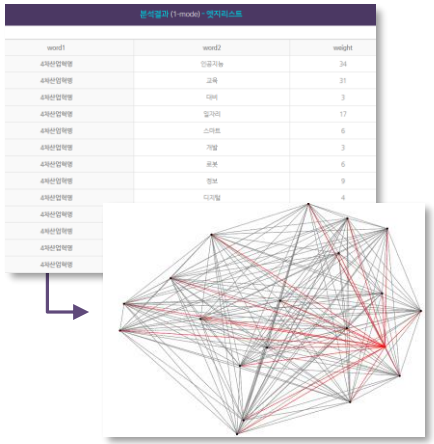
UCINET 활용

2-mode 분석



UCINET 활용

엣지리스트

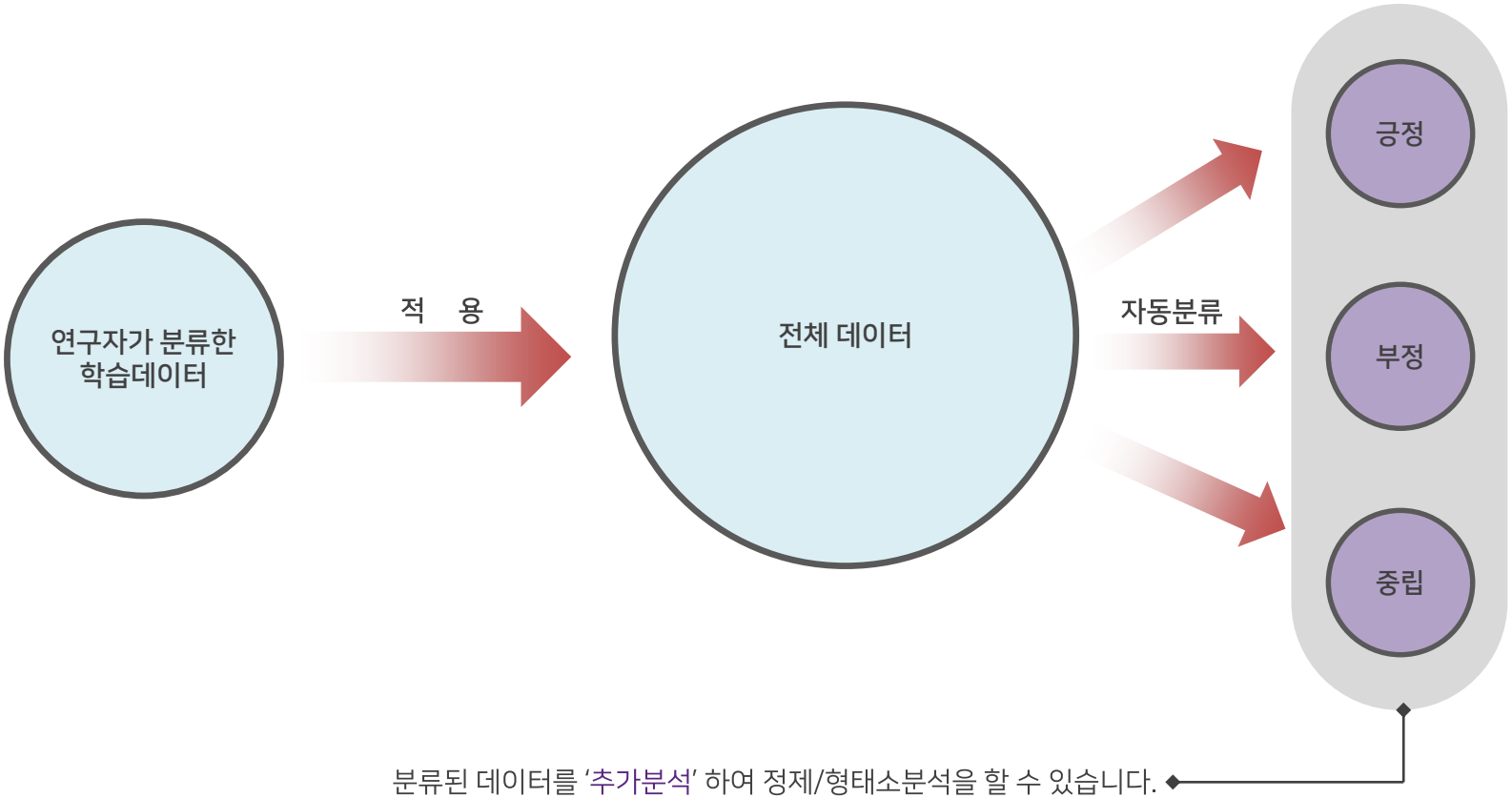


NODEXL 활용

◆ 호환 가능한 프로그램 UCINET Software NODEXL PAJEK Gephi

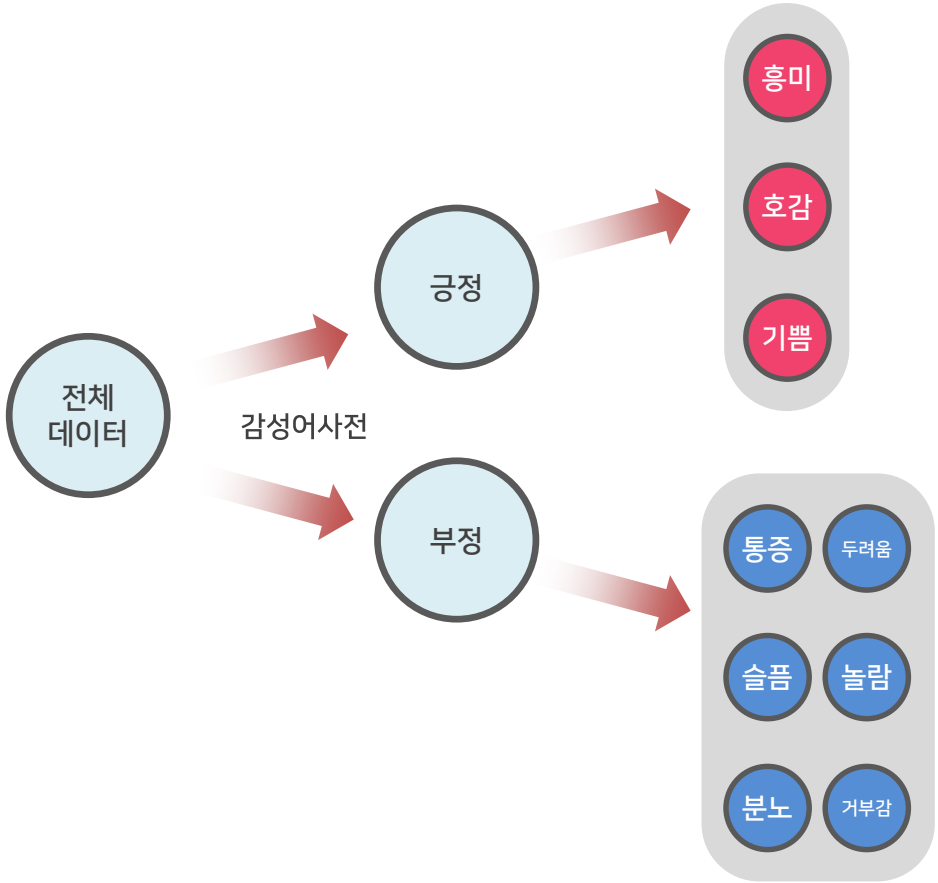
기계학습 기법의 감성분석이 가능합니다

베이지안 분류기(Bayes Classifier)를 통해 기계학습 기법의 감성분석 기능을 제공합니다.
연구자가 직접 학습데이터를 구성하여 적용함으로써 분석 주제의 제한 없이 모든 분야의 데이터에서 감성분석이 가능합니다.



키워드 기반 감성분석이 가능합니다

텍스톰이 보유하고 있는 감성어 사전을 기반으로 하여 전체 데이터의 긍정/부정 키워드에 대한 빈도분석이 가능합니다.



형태소 분석 결과 미리보기 (분석품사)

구분	빈도(건)	비율(%)
긍정	380 / 482	78.84 / 100.0
부정	102 / 482	21.16 / 100.0

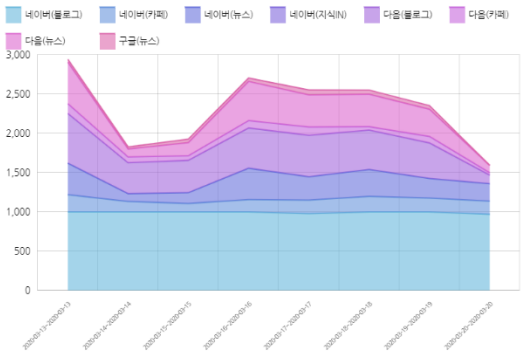
긍정 키워드 | 부정 키워드

흥미 | 호감 | 기쁨

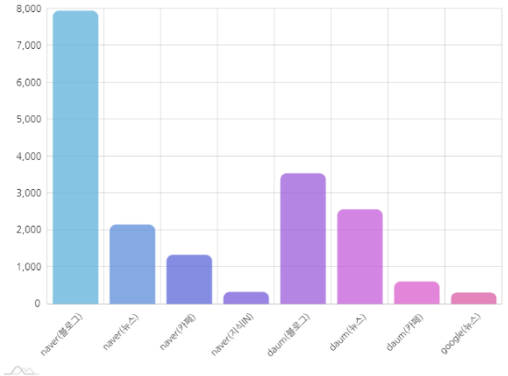
감정분류	빈도(건)	비율(%)
기대하다	21	4.36
새롭다	15	3.11
특별하다	11	2.28
파격적	8	1.66
인상적이다	6	1.24
이국적	2	0.41
혁신적	2	0.41
짜릿하다	2	0.41
독특하다	1	0.21

분석결과를 다양한 시각화 결과물로 나타낼 수 있습니다

수집량, 단어빈도, N-gram, Topic Modeling, 개체명인식 결과값을 다양한 종류의 차트로 표현할 수 있습니다.

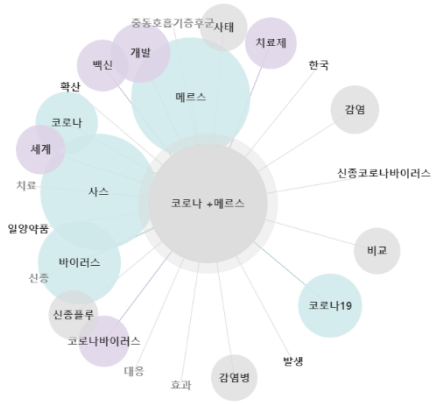


기간별 수집량그래프

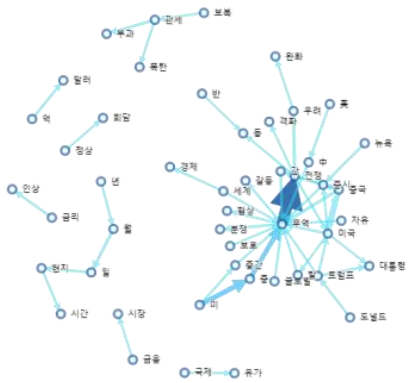


채널별 수집량그래프

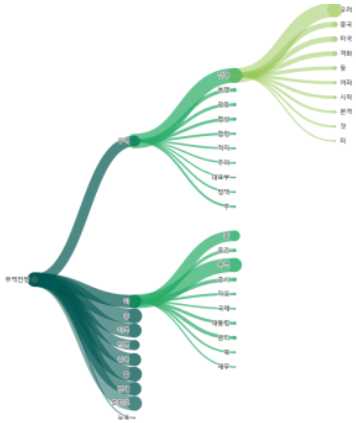
워드클라우드



에고네트워크



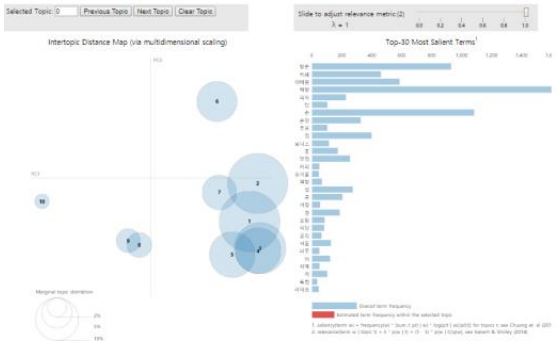
N-gram 네트워크



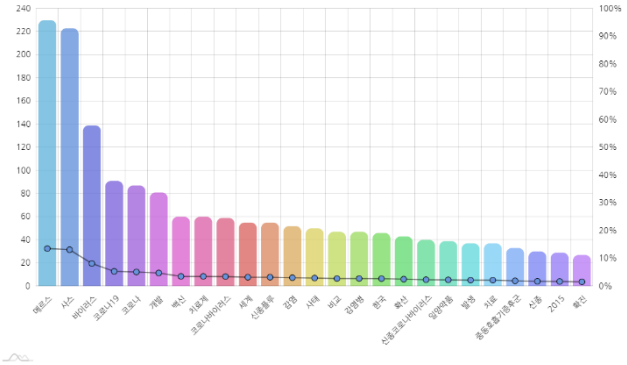
1-way 워드트리

분석결과를 다양한 시각화 결과물로 나타낼 수 있습니다

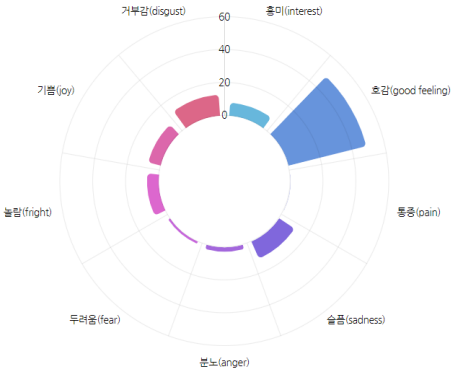
수집량, 단어빈도, N-gram, Topic Modeling, 개체명인식 결과값을 다양한 종류의 차트로 표현할 수 있습니다.



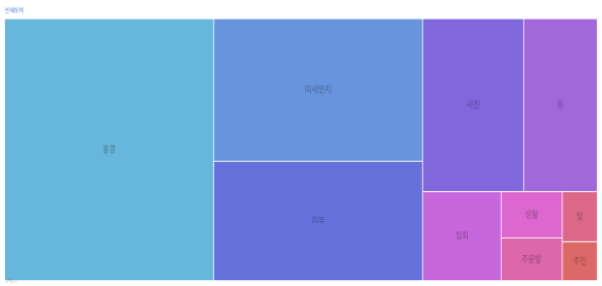
토픽모델링



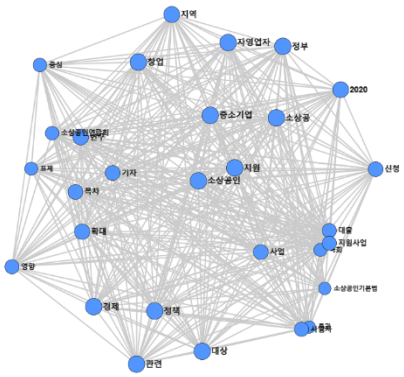
바차트



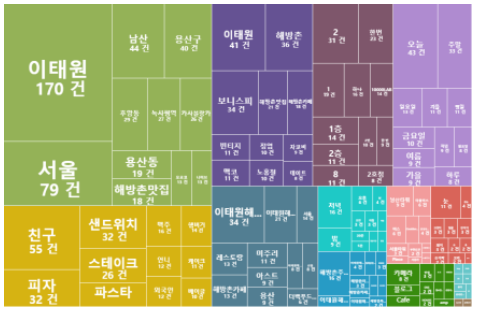
어휘 감성분석



클러스터링



매트릭스 차트



개체명인식

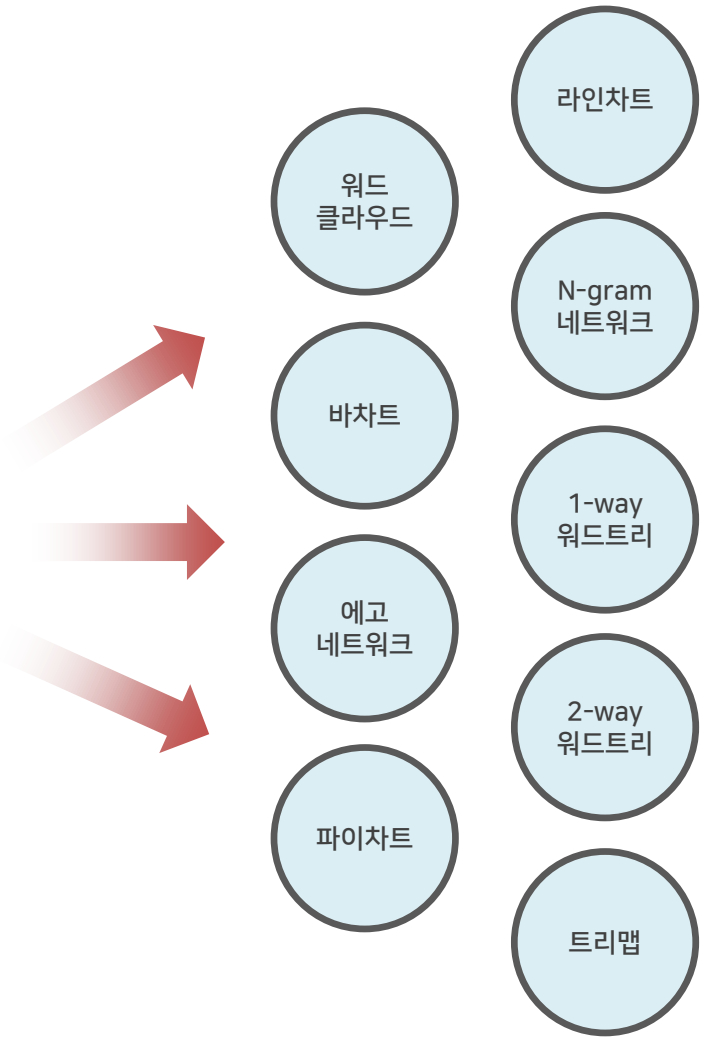
형식에 맞는 엑셀 파일을 업로드하면 다양한 유형의 시각화 결과물을 바로 산출할 수 있습니다

'시각화 커스터마이징'을 통해 원하는 시각화 결과물을 얻을 수 있습니다.

Sample

	A	B	C	D	E	F
1	빅데이터	29871				
2	분석	6118				
3	활용	3456				
4	정보	2647				
5	전문가	2408				
6	기술	2086				
7	산업	1911				
8	기반	1576				
9	연구	1544				
10	교육	1431				

해당 시각화 유형에 맞는 엑셀파일을 업로드



텍스톰의 강점은 크게 네 가지로 요약할 수 있습니다

1

데이터 수집 · 분석 작업에 소모되는 시간 단축

데이터 수집부터 정제, 분석, 시각화 작업까지 한 곳에서 처리 가능

2

다양한 데이터 분석 가능

웹사이트의 데이터 뿐만 아니라 사용자 보유데이터와 원하는 채널 수집까지 지원

3

다국어 텍스트 분석

한국어, 영어, 중국어를 지원하며 텍스톰 차이나를 통해 다양한 중국 채널에서 데이터 수집 가능

4

다양한 통계분석 프로그램에 적용 가능

호환성 높은 결과값을 제공

텍스톰의 강점은 수 많은 연구논문과 활용사례를 통해 검증됩니다

2020 | 2019 | 2018 | 2017 | 2016~2013

2020	Social media, media and urban transformation in the context of overtourism, International Journal of Tourism Cities, 2056-5607 장호찬, 박민경
2020	키워드 네트워크 분석을 활용한 영유아 놀이 관련 연구동향 분석, 학습지중심교과교육학회 최지은
2020	2015 개정 고등학교 가정 교과서의 '생활문화' 핵심개념 단원 분석, 학습지중심교과교육학회 김샛별, 채정현
2020	유아교육재정에 대한 키워드 네트워크 분석: 빅데이터를 중심으로, 학습지중심교과교육학회 서원석, 이강훈, 김석우
2020	빅데이터를 통해 바라본 유아 창의성과 놀이에 대한 사회적 인식 네트워크 분석 연구, 학습지중심교과교육학회 유효인, 문가영

텍스톰을 활용해 작성된 논문 210여 편
(온라인에서 공개된 논문 수)
약 60여 기관에서 텍스톰을 활용해 연구/교육을 진행

THANK YOU