

빅데이터 분석솔루션 TEXTOM 소개서

빅데이터 수집·분석·시각화를 쉽고 빠르게

TEXTOM

CONTENTS

01

TEXTOM 히스토리

02

TEXTOM 특징

03

세부기능 소개

04

TEXTOM 강점

01 TEXTOM 히스토리

웹 환경에서 데이터를 수집하고 정제하며 다양한 분석을 처리할 수 있습니다

2013.04

The SCRM 서비스 시작(텍스툼의 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

텍스툼 1.0

다국어 수집(영어, 중국어)
TF-IDF
N-gram

2015.01

텍스툼 1.5

사용자 정의 사전
전체 정제 및 타이틀 정제
가정제 데이터 편집

2015.08

텍스툼 2.0

에고네트워크 시각화 / 워드트리 시각화
서베이 / 정제모듈 MeCab-ko
개체명 인식

2018.04

텍스툼 2.5

Customized Data / Visualization_ Upload Data
News Data 채널 확장 / User Data 키워드 지정
Sentiment Analysis / 수집, 분석 처리 속도 향상

2018.12

텍스툼 3.0

UI, UX 변경 / 정제단계에서 용량 삭감
복수의 수집데이터 통합
키워드 필터링 / 분석품사 외국어, 숫자 추가
중복제거 / Window size

2019.09

텍스툼 3.5

수집데이터 관리 기능개선 / 수집량 시각화 제공
정제 단어 추천 / 옛지리스트 제공
상세품사 형태소 분석 / 형태소 분석결과 미리보기
감성분석 기본 학습데이터 제공 / 키워드 기반 감성분석 추가

2020.03

텍스툼 4.0

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

2020.09

텍스툼 4.5

키워드 미리보기 기능 제공 / TEXTOM Edu 오픈
데이터 분석 속도 향상 / 자동정제 기능 업그레이드
빅데이터 아카데미, 데이터 수집요청, 텍스툼 구축 소개 메뉴 추가
감성단어 빈도 분석 개선 / 감성단어 워드 클라우드 추가
LDA 토픽분석 정제데이터 다운로드 기능 / 분석방법 설명 버튼 추가

2021.03

텍스툼 5.0

담론분석 기능 추가 / 시계열 분석 기능 추가
토픽단어 : 클러스터 유사 단어 확인 기능 제공
감성분석 차트 고도화
Edu 사용자 편의성 향상
형태소 분석기 성능 고도화

2021.09

텍스툼 5.5

메인페이지 디자인 리뉴얼
채팅형 CS 도입
페이지 로딩 속도 개선
Edu 계정 신청 절차 간소화
Edu 실습데이터 수집 및 분석 속도 향상

2022.03

텍스툼 6.0

2-mode 바로 선택하기 기능 추가
사전기반 감성분석의 단어 포함 원문 제공
문서기반 감성분석 모델 성능 제공
중심성분석 기능 추가

02 TEXTOM의 특징

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

효율적인 데이터 저장

국문, 영문, 중문 형태소 분석기

다양한 데이터 필터링 기능



결과물 산출

토픽모델링

매트릭스

중심성분석

감성분석

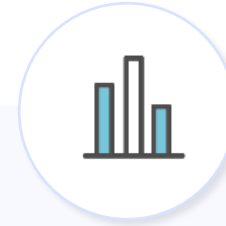
단어빈도

N-gram

TF-IDF

연결중심성

개체명인식



시각화

수집량 라인차트

워드클라우드

에고네트워크

N-gram 네트워크

토픽모델링

트리맵

매트릭스차트

감성분석

02 TEXTOM의 특징

사용이 쉽고 이용이 편리합니다.

중학생부터 기업인, 연구자까지 **사용의 폭이 넓고 다양**합니다.

사용자 환경에 최적화한 UI, UX와 상세한 매뉴얼을 통해 **누구나 쉽고 편리하게 이용**할 수 있습니다.

데이터수집

포털/SNS

뉴스

보유데이터

요청채널

데이터 전처리

수집리스트

정제 / 형태소 분석

데이터분석

텍스트 마이닝

매트릭스

담론분석

감성분석

토픽분석

시계열분석(Beta)

시각화

시각화 결과

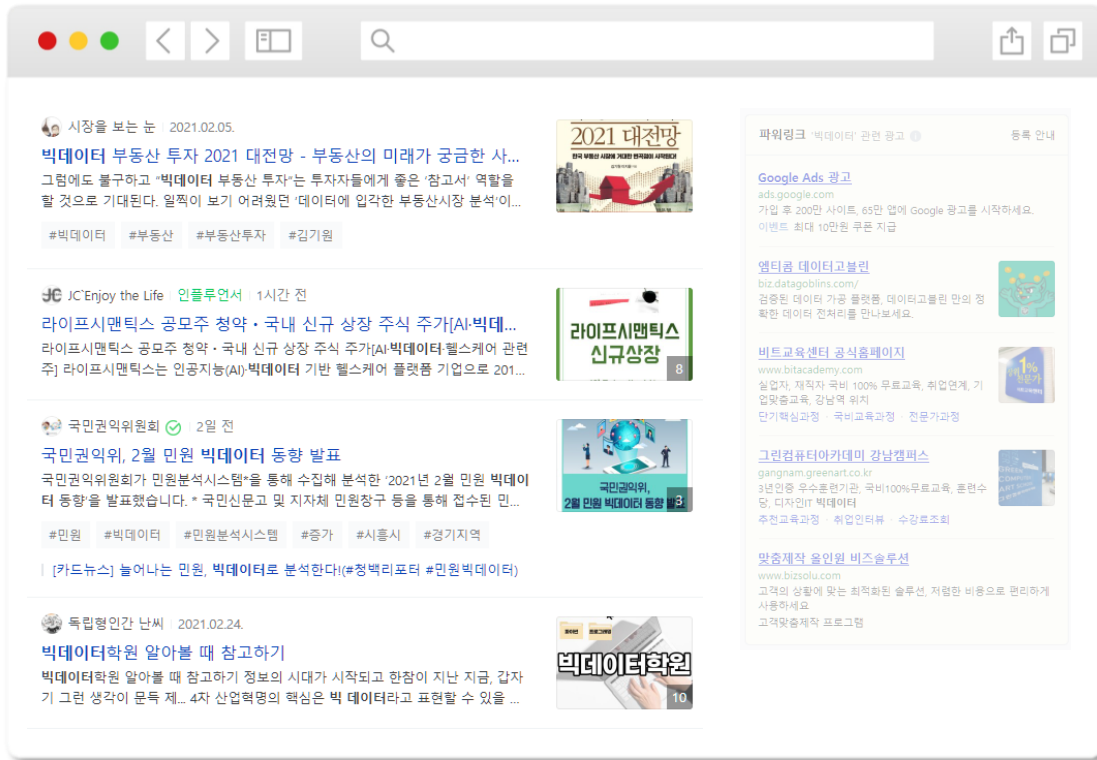
커스터마이징

03 세부기능 소개 - 1) 데이터수집

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.



원문데이터

미리보기

다운로드(Excel)

다운로드(txt)

	A	B	C	D	E
1	ai-times :: [SPSS 강좌] 조건에 맞는 레코드들	http://ai-times.tistory.c	SPSS 에서 주어	naver	web
2	개인사용자를 위한 피싱예방 가이드 - 자료실	https://www.boho.or.k	자료실 게시판	naver	web
3	'데이터품질관리와 경제, MDM' 세미나 개최	http://datastreams.co.k	- 원문보기	- naver	web
4	ai-times :: [데이터] 학생들의 체험업사에 디	http://ai-times.tistory.c	간단한 SPSS 및	naver	web
5	ai-times :: [데이터마이닝] 실험데이터정리	http://ai-times.tistory.c	데이터마이닝	naver	web
6	부동산114 창립 10주년 기념 공개사무소 회	https://www.r114.com/	종합부동산포	naver	web
7	데이터센터의 향방을 결정하는 5가지 예	http://www.itworld.co	*****@***.***	naver	web
8	iCON - 융합정보 - 연구데이터의 중요성 증	http://icon.ndsl.kr/_L_tre.aspx	관련정보	naver	web
9	데이터센터 설계 : "형태는 기능을 따른다"	http://www.itworld.co	*****@***.***	naver	web
10	[대박 갈사이벤트] 열화와 같은 성원에 감	https://www.r114.com/	종합부동산포	naver	web
11	빅뱅 (Big Bang)	https://blog.naver.com	근황 후 빅뱅	naver	web
12	[데이터마이닝기술 기법 개념]데이터마이	http://www.reportwor	데이터마이닝	naver	web
13	EMC의데이터도메인 인수와 스토리지 시	http://www.itworld.co	추천 테크라이	naver	web
14	OpenParadigm :: 빅오표기법/빅오분석법	(E http://openparadigm.ti	문다면데이터	naver	web
15	데이터센터 전력 절감, "문석 방법 톨렸다"	http://www.itworld.co	*****@***.***	naver	web
16	이상적인 홈데이터센터의 조건 - ITWorld K	http://www.itworld.co	> 뉴스 2009.	naver	web
17	데이터통합 및 거버넌스(2회) - DataStream	http://datastreams.co.k	- 원문보기-- 0	naver	web
18	오라클, 데이터센터 하드웨어의 애플을 꿈	http://www.itworld.co	*****@***.***	naver	web
19	[데이터마이닝 등장배경]데이터마이닝 기	http://www.reportwor	Big Data 특성고	naver	web
20	비활성데이터위한 클라우드 스토리지 서	http://www.itworld.co	> 뉴스 2009.	naver	web
21	[특집] 쉽고 강력한 접근 정책은 기업 DB	http://www.boannews	내부자의 데이	naver	web
22	[서울신문] [2030] 당신이 만난 최고의 리	http://www.seoul.co.kr	의료정보빅데	naver	web
23	유럽데이터센터 시장, "자체 용량은 출고	0 http://www.itworld.co	*****@***.***	naver	web



03 세부기능 소개 - 1) 데이터수집

사용자의 분석 주제에 맞는 다양한 수집채널을 선택할 수 있습니다.

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



03 세부기능 소개 - 1) 데이터수집

세부 설정기능을 통해 수집결과와 만족도를 높일 수 있습니다.

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

The screenshot shows a web interface for configuring data collection. It includes sections for: '키워드 미리보기' (Keyword Preview) with a confirmation button; '수집키워드' (Collection Keyword) with a '키워드추가' (Add Keyword) button and a dropdown for '연산자' (Operator); '기간' (Period) with date pickers and radio buttons for '1주', '3개월', and '1년'; '수집단위' (Collection Unit) with '사용' (Use) and '사용안함' (Do not use) buttons; and '채널' (Channel) with checkboxes for various sources like NAVER, Digt, Google, and YouTube. A '수집정보' (Collection Info) button is also present.

키워드 미리보기

- 수집하기 전에 수집할 키워드의 정보량, 검색추이를 확인 할 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.

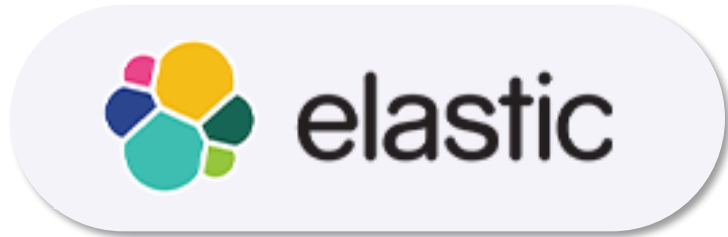
채널

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

03 세부기능 소개 - 2) 데이터 저장

분산파일 처리시스템 엘라스틱(elastic)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.



- 수집 · 정제 · 분석된 데이터의 저장과 관리를 위한 분산파일 시스템과 NoSQL 기능 구현
- 데이터의 효율적인 선택과 실시간 분석을 위한 데이터 색인 기능 제공
- 채널별 보관된 데이터량 및 수집량 확보



03 세부기능 소개 - 3) 데이터 정제

국문 뿐만 아니라 영문, 중문 데이터도 분석이 가능합니다.

국문, 영문, 중문 형태소분석기를 적용하여 **다국어 분석**이 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](#)

03 세부기능 소개 - 3) 데이터 정제

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

- 텍스트가 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.

키워드 필터링

- 특정 키워드가 포함된 문서 자체를 정제데이터에서 제외하거나 특정 키워드가 포함된 문서만 추출하여 정제 결과에 반영할 수 있습니다.

중복제거

- 수집된 데이터에서 중복되는 URL이나 내용을 제외합니다.

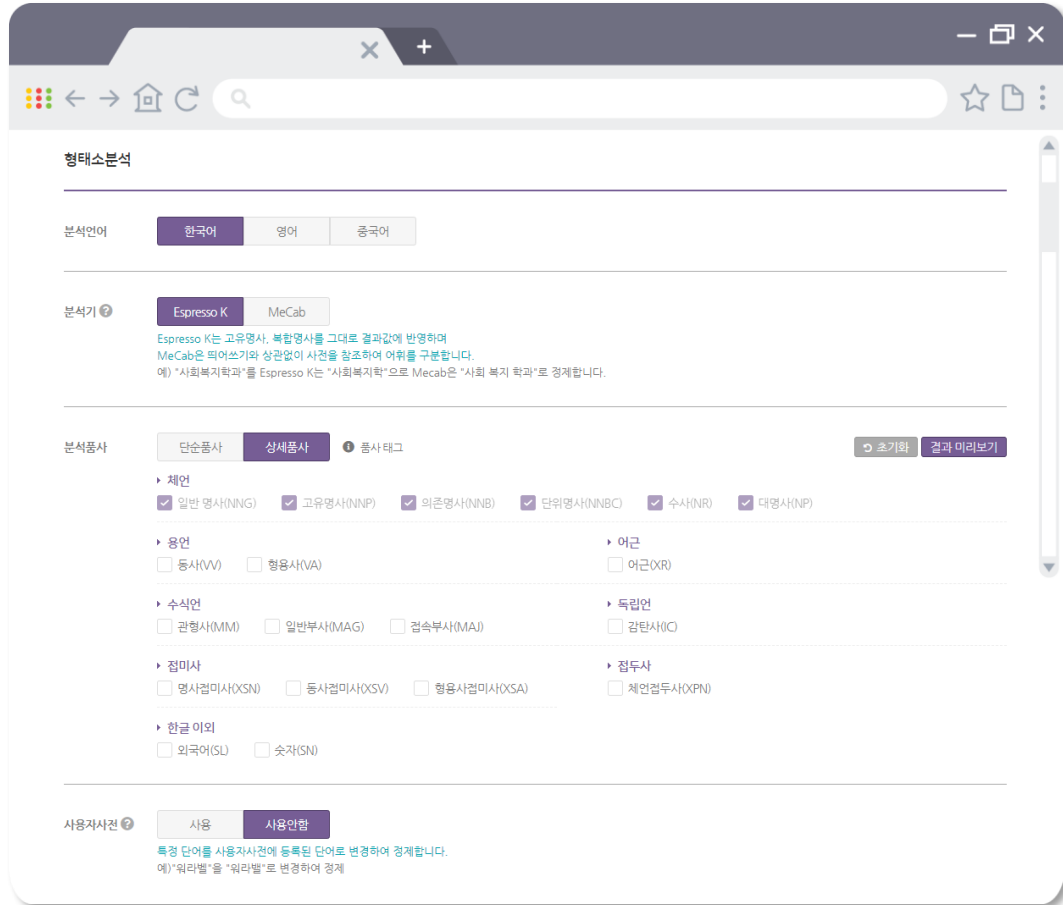
선택안함

- 이미 정제한 데이터 즉, 전처리가 필요하지 않은 데이터를 데이터분석 단계로 넘길 때 이용할 수 있습니다.

03 세부기능 소개 - 3) 데이터 정제

한국어 자연어 처리에 뛰어난 형태소분석 기능을 제공합니다.

세밀한 분석품사 설정을 통해 정제 데이터의 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

단어품사 : 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 등록된 특정 단어들을 정제데이터에 미리 반영할 수 있습니다.

- 동일 · 유사한 데이터를 반복 정제할 때 유용하게 사용할 수 있습니다.

03 세부기능 소개 - 3) 데이터 정제

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

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

정제데이터 (텍스트마이닝)

데이터명	생성날짜	용량
4차 산업혁명 +일자리	2019-09-16	2.26 MB

변경한 내용 적용

미리보기 바로편집하기

● 정확한일치 ① 부분일치 ⓘ 사용자사전 미지정 ▼ 적용

변경할 단어 + → 수정단어 변경

사용자사전 수정은 사용자사전 메뉴에서 가능합니다. 사용자사전 바로가기 >

추천단어 ⓘ

- 4차 산업혁명 1171
- 4차산업혁명
- 일자리 창출 197
- 일자리창출
- 산업혁명시대 95
- 산업혁명시대

수정내역

- 빅데이터
- 빅데이터
- 텍스트
- TEXTOM

4차 산업혁명 미래 관광일자리창출 방안 연구 표제지목록연구개 4제1장 서론 연구 배경 목적 연구 배경 연구 목적 연구 범위 방법 연구 범위 연구 방법 29제2장 4차 산업혁명 관광일자리 동향 4차 산업혁명 일자리 4차 산업혁명 개요 4차 산업혁명 일자리 4차 산업혁명 관광일자리 4차 산업혁명시 대 관광산업 4차 산업혁명 관련 관광정책 4차 산업혁명 관광일자리 동향 47제3장 전문가 의견 조사 주요 사례 전문가 의견 조사 조사 개요 조사 결과 사례 분석 국내 관광기업 사례 해외 관광기업 사례 해외 일자리 정책 73제4장 기본 방향 추진 과제 종합 분석 4차 산업혁명 관광산업 변화 4차 산업혁명 관광산업 비즈니스 프로세스 변화 4차 산업혁명 일자리 직무 변화 종합 분석 기본 방향 4차 산업혁명 선도 관광산업 혁신 생태계 구축 관광혁신 일자리 추진 전략 주요 추진 과제 관광산업 혁신 지원 구조 재정 일자리 창출 다각화 지원형태 고 지원 지역관광 안전망 연계 100제5장 결론 정책 제언 103참고문헌 부록 전문가 의견 조사 설문지 112판권기 산업혁명 주요 특징 기술변화 일자리 영향 산업별 취업 전망 4차 산업혁명 시대 신직업 4차 산업혁명

· 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.

03 세부기능 소개 - 3) 데이터 정제

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

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



분석결과

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.

개체명인식

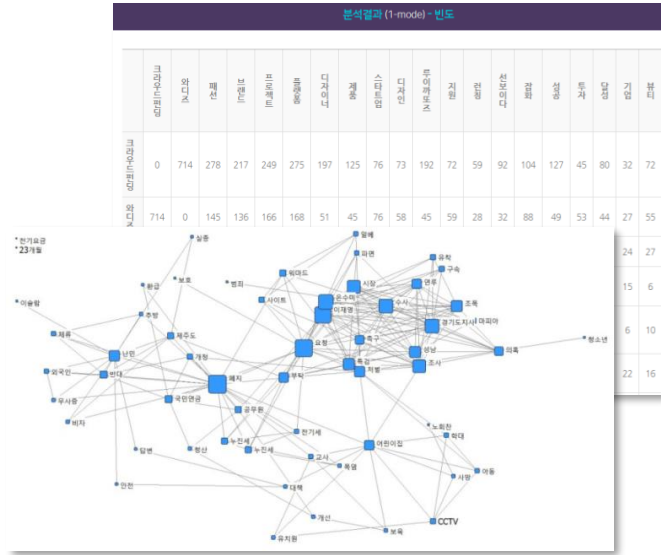
- 자동 정제 결과에 따라 분리된 단어를 14개의 범주에 따라 분류합니다.
(사람, 학문, 대상물, 기관, 지역, 문명, 날짜, 시간, 숫자, 사건/사고, 식물, 금속, 용어)

03 세부기능 소개 - 3) 데이터 정제

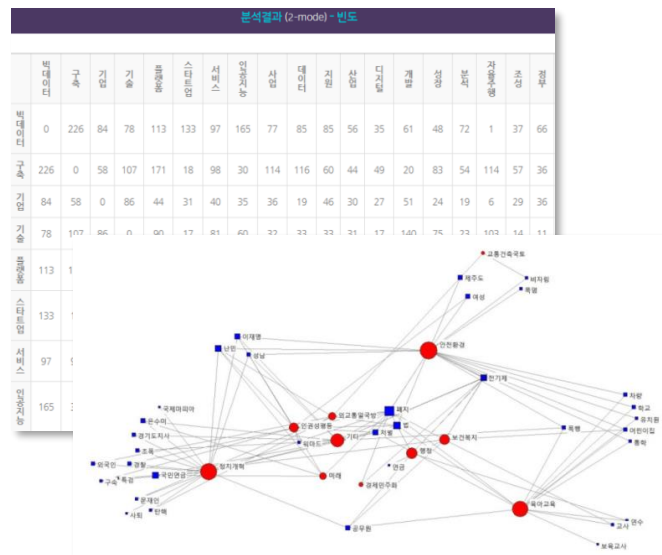
데이터 편집 기능으로 2차 정제가 가능하여 결과물의 정확도를 높일 수 있습니다.

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

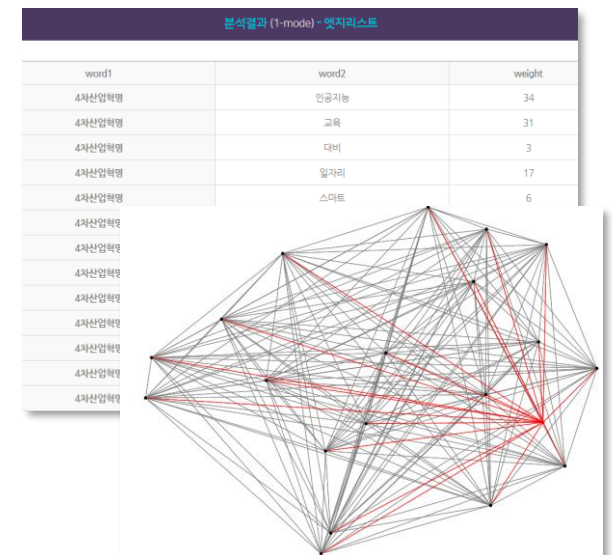
1-mode 분석



2-mode 분석



엣지리스트



◆ 호환 가능한 프로그램

UCINET Software

NODEXL

PAJEK

Gephi

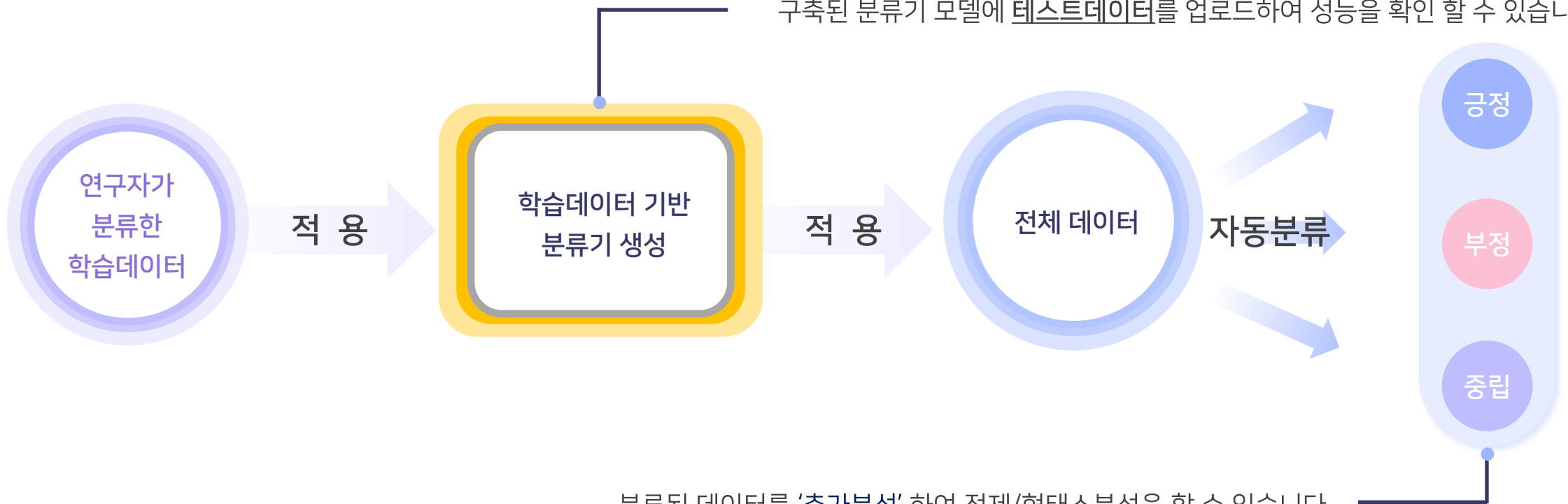
03 세부기능 소개 - 4) 결과물 산출

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.

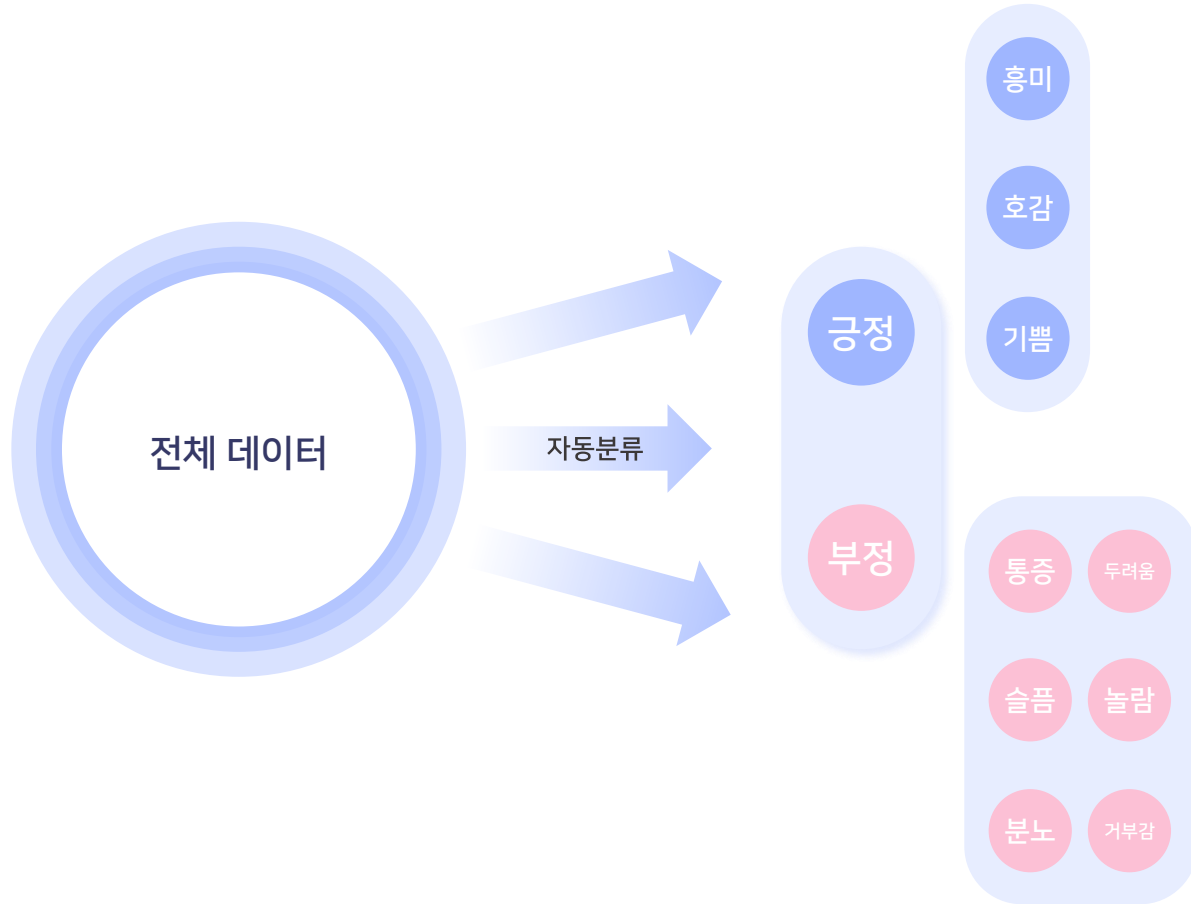
구축된 분류기 모델에 테스트데이터를 업로드하여 성능을 확인 할 수 있습니다.



03 세부기능 소개 - 4) 결과물 산출

키워드 기반 감성분석이 가능합니다.

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



The screenshot shows a web interface for sentiment analysis results. The main table displays overall statistics:

구분	빈도(건)	감성강도비율(%)	빈도비율(%)
긍정	7470 / 8549	88.37 / 100.0	87.38 / 100.0
부정	1079 / 8549	11.63 / 100.0	12.62 / 100.0

Below this, there are tabs for '긍정 키워드', '부정 키워드', and '세부감성'. The '세부감성' tab is active, showing a table for the keyword '흥미' (Interest):

감성분류	빈도(건)	감정강도	빈도 + 감정강도	빈도비율(%)
희신적	374	3.88889	1454.44486	4.37
원하다	172	5.0	860	2.01
새롭다	94	2.7778	261.1132	1.1
특별하다	84	3.7778	317.33352	0.98
기대하다	49	4.6667	228.66683	0.57
재미있다	27	2.6667	72.0009	0.32
특이하다	26	4.0	104	0.3
환상적이다	19	5.7778	109.77782	0.22
이상적이다	19	3.4444	65.4436	0.22
사교적이다	12	3.6667	44.0004	0.14
신기하다	12	5.3333	63.9996	0.14
파격적	11	4.7778	52.5558	0.13

감성키워드를 클릭하면 해당 키워드가 언급된 원문을 확인할 수 있습니다.

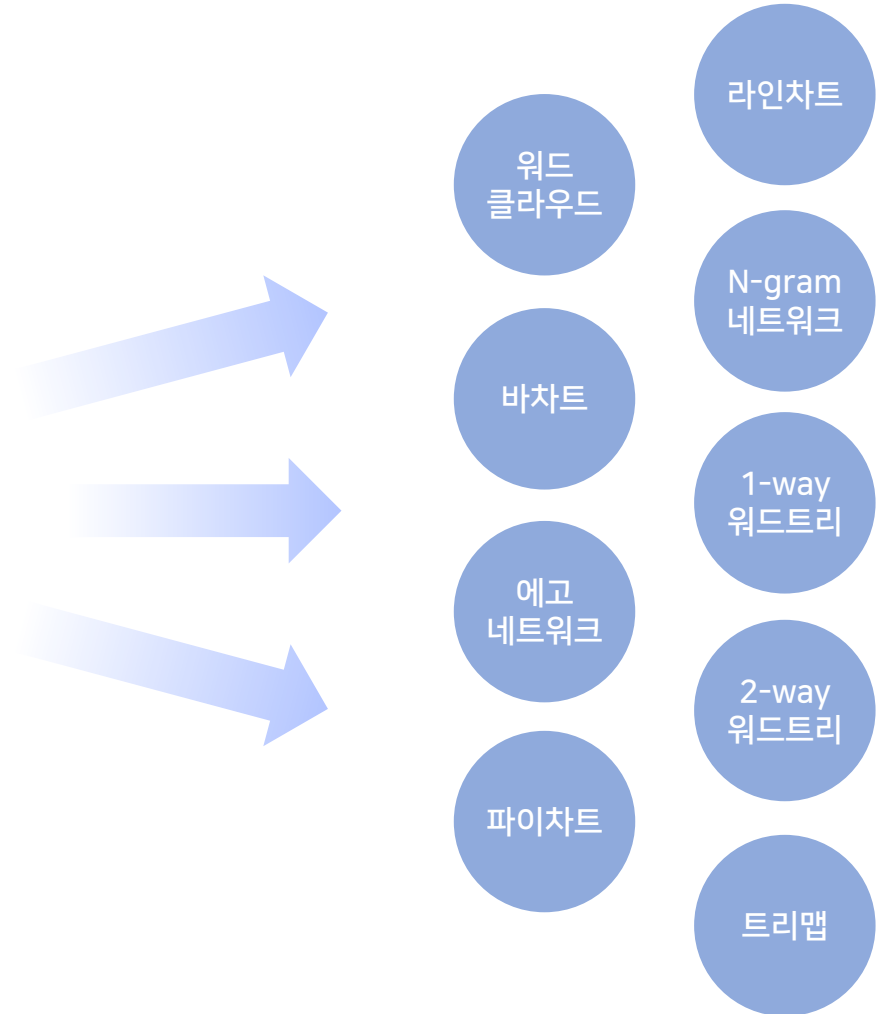
03 세부기능 소개 - 4) 결과물 산출

형식에 맞는 엑셀 파일을 업로드하면 다양한 유형의 시각화 결과물을 바로 산출 할 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



04 TEXTOM의 강점

텍스툼의 강점은 크게 4가지로 요약할 수 있습니다.



데이터 수집·분석 작업에 소모되는 시간 단축

데이터 수집부터 정제, 분석, 시각화 작업까지
한 곳에서 처리 가능



다국어 텍스트 분석

한국어, 영어, 중국어를 지원하며 텍스툼 차이나를
통해 다양한 중국채널 데이터 수집가능



다양한 데이터 분석 가능

단어 간 관계를 보는 매트릭스,
텍스트에서 감성을 발굴해주는 감성분석 등



다양한 통계분석 프로그램에 적용 가능

호환성이 높은 결과값을 제공

04 TEXTOM의 강점

텍스톰의 수 많은 연구논문과 활용사례를 통해 검증되고 있습니다.

A screenshot of the TEXTOM search results page. It shows a list of research papers with filters for years (2021, 2020, 2019, 2018, 2017, 2016-2013). The list includes titles, authors, and publication details.

Year	Title	Author	Journal
2021	소셜네트워킹을 기반으로 한 도시 문화 관광기생종의 키워드 분석, 초항미디어학, 1229-5560	정수근, 이상영	초항미디어학
2021	언어 네트워크 분석을 활용한 대학 실시간 화상강의와 녹화강의에 대한 학습자 인식 분석, 교육정보미디어연구, 1229-7291	김영지, 김영성	교육정보미디어연구
2021	코로나19 시대 빅데이터에 나타난 유아안전 의미연결망 분석, Culinary Science & Hospitality Research, 1226-2722	이경선	Culinary Science & Hospitality Research
2021	한국미용모에 대한 관심 변화와 정부정책의 방향: 1995년-2020년 소셜미디어 빅데이터 분석, 한국융합학회논문지, 2233-4890	서동희, 진복선	한국융합학회논문지
2021	빅데이터 분석을 이용한 로스트 COVID 시대의 여행 스트레스 연관 키워드 분석에 관한 연구, 융복합지식학회논문지, 2287-6920	이정섭, 윤석재	융복합지식학회논문지
2021	이동확대 온톨로지 개발, 한국이동복지학, 1226-2609	최익훈, 이세원, 오세현, 김지선, 현종섭, 김현아, 노종래	한국이동복지학



A grid of logos for various educational institutions and research centers that utilize TEXTOM. The logos are arranged in two rows under the heading '교육기관' (Educational Institutions).

Row 1	Row 2
XNU	5차공사
KU	KRIS
경성대학교	ETRI
계명대학교	한국문화연구원
고려대학교	KRIF
강원대학교	한국과학기술연구원
연세대학교	한국과학기술연구원
UNIST	한국과학기술연구원
한국과학기술연구원	농수산식품연구원
한국과학기술연구원	국립기술물질원
한국과학기술연구원	농수산식품교육문화정보원
KOFRUM	한국과학기술연구원
한국과학기술연구원	한국과학기술연구원
한국과학기술연구원	한국과학기술연구원

2022년 3월 기준

TEXTOM을 활용해 작성된 논문 300여 편

약 100여개의 기관에서 TEXTOM을 활용해 연구/교육을 진행하였습니다.

감사합니다.

문의 : textom@theimc.co.kr