

# 아이 발용 데이터 분석 및 머신 러닝



**차시** 텍스트분석1



## 텍스트 데이터 다루기

- 텍스트 데이터 분석을 위한 전처리과정
- 워드 클라우드를 활용한 텍스트 데이터 시각화

#### 뉴스기반 통계 검색 서비스





#### <u>뉴스기반 통계 검색 서비스</u>



금리 - 관련 뉴스 2,545 건 2022/10/16~2022/10/22			+
제목		출처	날짜
합작품? 엔 달러 환율, 두시간 만에 7엔 '뚝' 무슨 일?		머니투데이	2022/10/22
팬데믹 2년간 한국 집값 상승률 '15.8%' 미국 호주의 절반 …	ď	머니S	2022/10/22
파운드화 추락에 웃는 스카치 위스키 업계 수출 10%		연합뉴스	2022/10/22
테슬라 3.45%-루시드 5.63%, 전기차 일제 급등(상보)	Z.	뉴스1	2022/10/22
코스피 2200 줄다리기 개인 vs 기관 엇갈린 베팅		매경이코노미	2022/10/22
추락과 반등의 갈림길에 선 카카오 왕국	ď	시사저널	2022/10/22
채권시장 패닉 A등급 회사채 58% 안팔렸다	ď	조선일보	2022/10/22
집값 하락 지금부터 강남권도 6억씩 뚝뚝' 얼어붙는 부동산…		뉴스1	2022/10/22

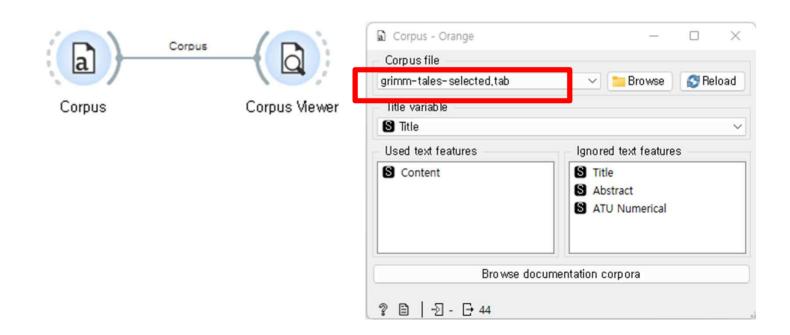


○ 단어 토큰화 : 어간분석을 통해 주요 단어 추출, 구두점과 특수문자 및 의미 없는 단어 (불용어-Stopword) 의 제거, 형태소 분석 등의 전처리 과정이 필요.

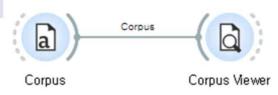
# 텍스트 가져오기

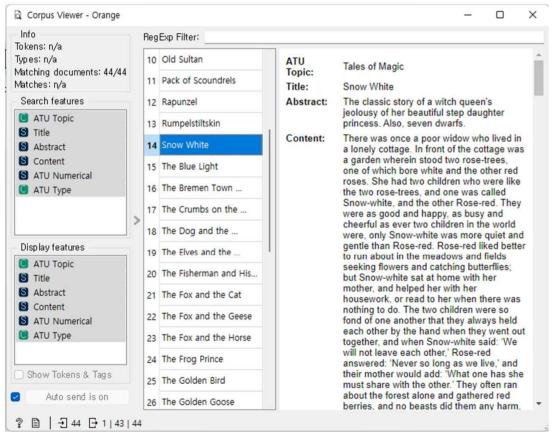


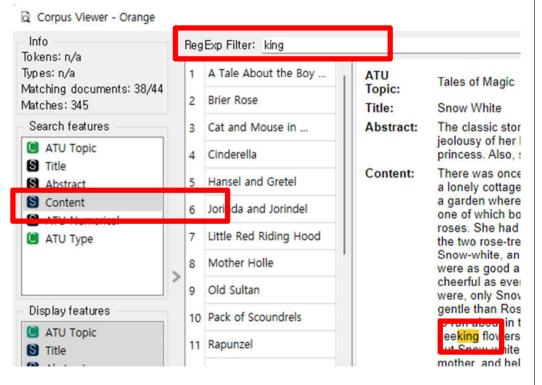
◦ Corpus (말뭉치) : 언어연구를 위해 컴퓨터가 텍스트를 처리하거나 분석할 수 있도록 한 자료모음



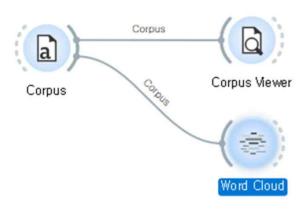
# 코퍼스뷰어 활용

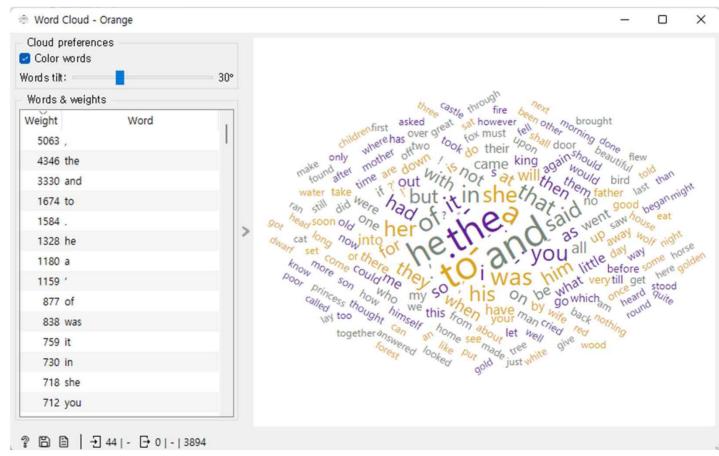


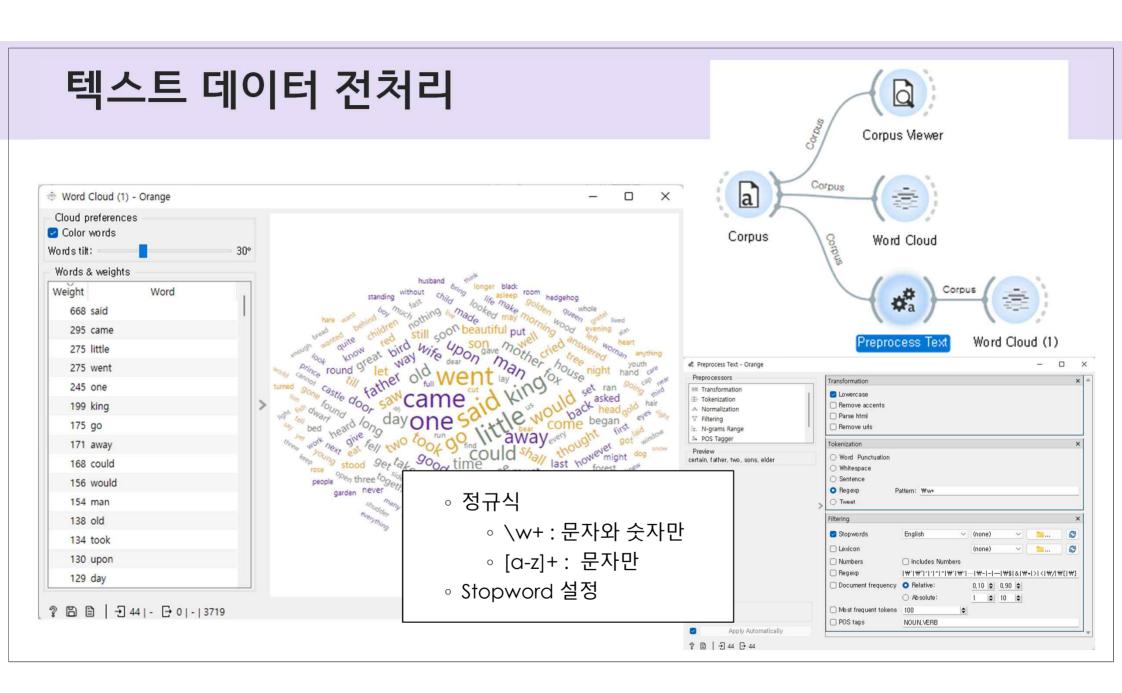




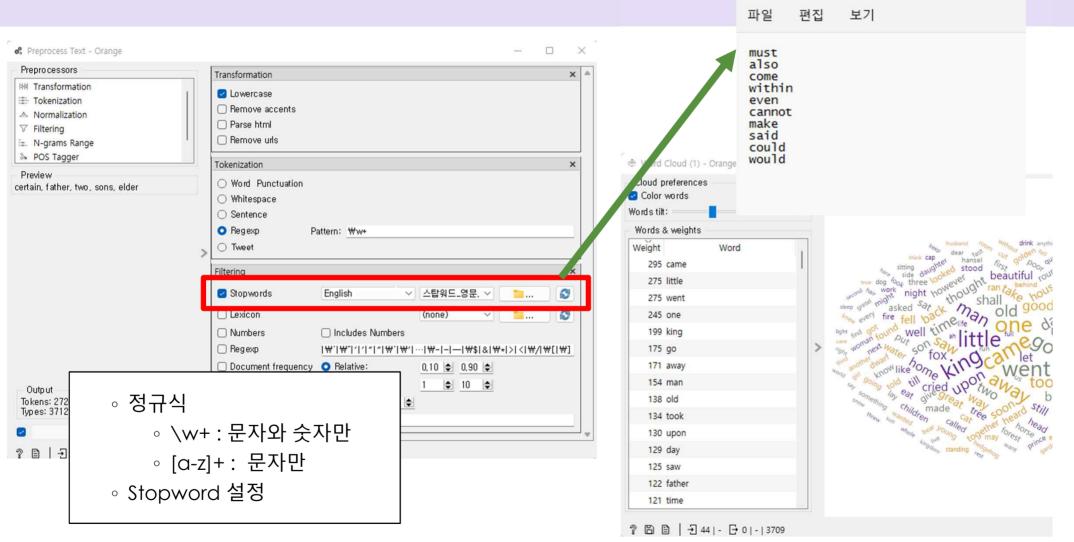
# 워드 클라우드로 문서 시각화







# 불용어 처리 - Stopword



스탑워드\_영문.txt - 메모장

# Bag of Words

- Bag of Words
  - Bag of Words란 단어들의 순서는 전혀 고려하지 않고, 단어들의 출현 빈도 (frequency)에만 집중하는 텍스트 데이터의 수치화 표현 방법



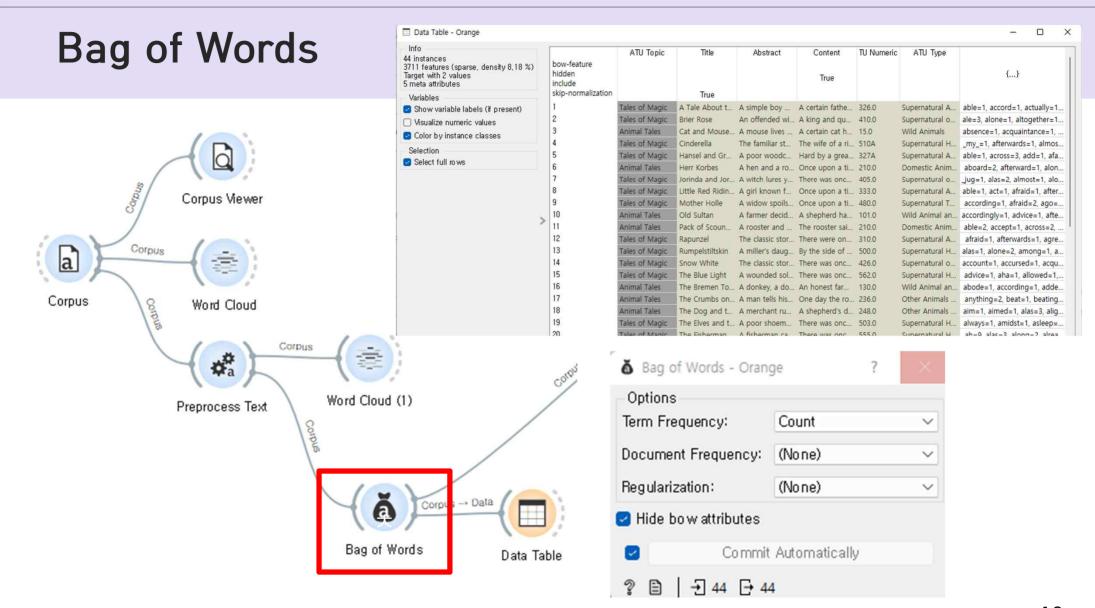
# Bag of words 의 단어당 빈도수를 활용한 유사도 검증

### **Bag of Words**

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

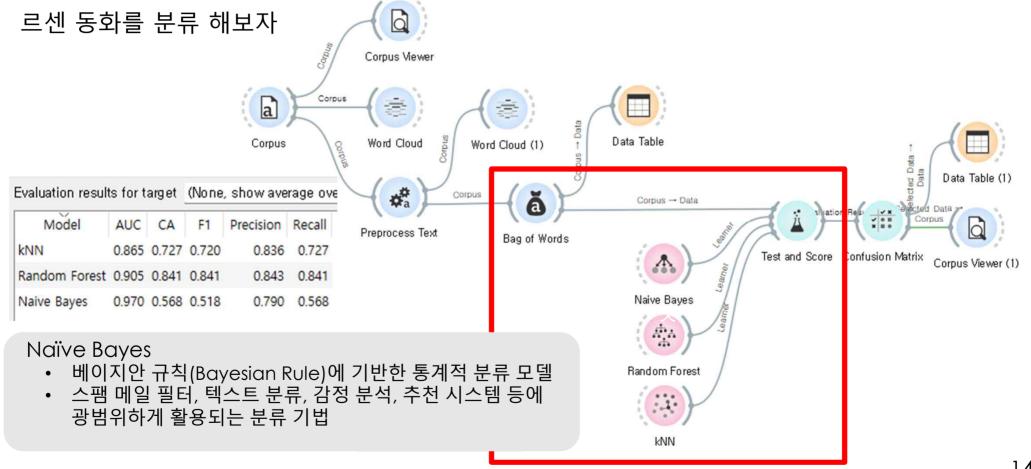
Stopword List: a, the, is, to, in, with, ...

Document	cat	sat	hat
the cat sat	1	1	0
the cat sat in the hat	1	1	1
the cat with the hat	1	0	1

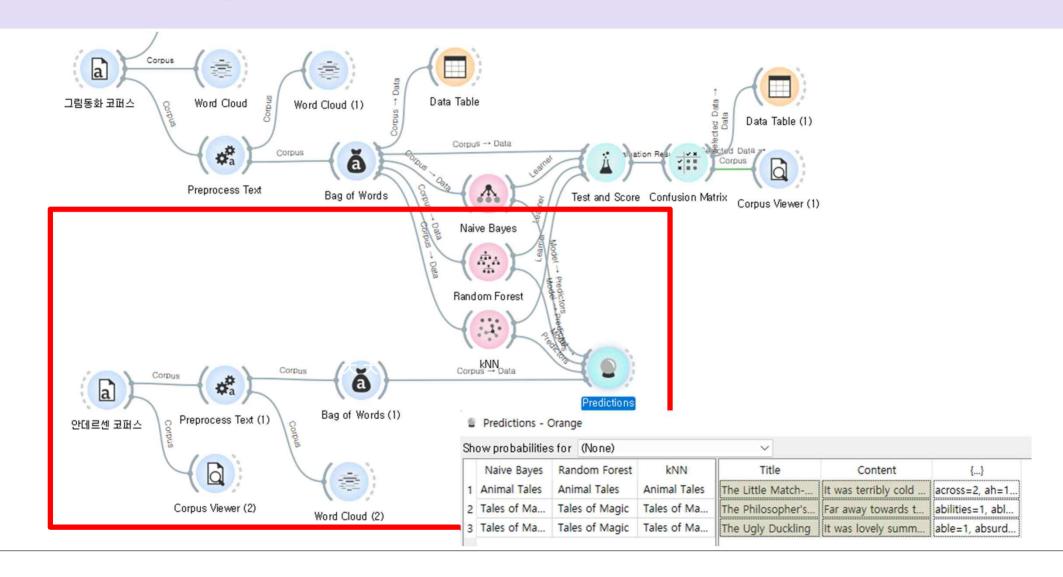


# 그림동화 분류

◦ grimm-tales-selected.tab Corpus 에서 44개 동화를 ATU Topic 으로 레이블로 학습 시킨 후 안데



# 안데르센 동화 분류



# orange

다음 시간에는 텍스트데이터를 군집 화 해보고 SNS를 통한 선호도 평가 방법을 다루어 보도록 하겠습니다.



# 아이 발용 데이터 분석 및 머신 러닝



**11차시** 텍스트분석2



# 텍스트 데이터 다루기2

- 텍스트 문서의 유사도 검증
- SNS를 활용한 텍스트 분석
- 감정분석 (Sentiment Analysis)

# 역대대통령의취임사를 통해 연설의 특징 을 탐색하고 군집화해 봅시다.

- 대통령취임사1998\_국문.txt
- 대통령취임사2003\_국문.txt
- 대통령취임사2008\_국문.txt
- 대통령취임사2013\_국문.txt
- 대통령취임사2018\_국문.txt
- 대통령취임사2022\_국문.txt

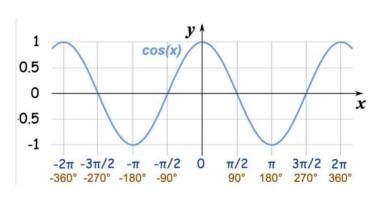
■ 대통령취임사\_영문.txt

한글과 영문 문서에서 분석의 차이 ? 형태소 분석은 안됨.

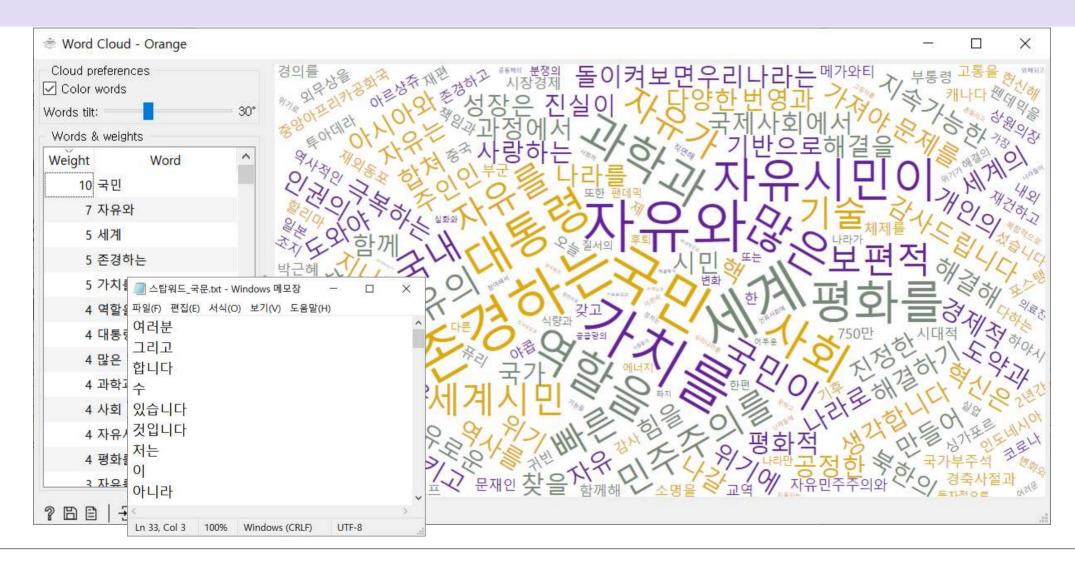
# 문서 간의 유사도 검증 방법

- 단어의 의미적인 관계를 고려해야 하기 때문에 단어를 벡터화 하여 벡터들 간의 거리를 측정한다.
- 유클리디안 거리는 단어 빈도수 측면에서 한계가 존재
- 코사인 유사도 (Cosine Similarity): 벡터간의 사잇값(각도)을 계산하여 유사도를 측정 -1~1 사이의 값으로 계산. 이 각도에 따라 두 벡터의 관계를 확인할 수 있음. (각도가 0 이면 cos(0)=1, 각도가 180 이면 cos(180)=-1 각도가 90이면 cos(90)=0 따라서 1에 가까울 수록 유사하다. 단 문서를 featur화시킬 때 음수값이 없으므로 각도는 0~90 따라서 코사인 유사도는 0~1 값을 가진다.





# 대통령 취임사 (2022.05.) - 불용어(stopwords)



# 자연어처리 단계

단계	설명	보기: 나는 그 과자를 먹었다.		
형태소 분석	문장을 형태소 열로 분리하고 품사를 부착하는 단계	나(대명사)+는(조사) 그(대명사) 과자(명사)+를(조사) 먹-(동사)+-었-(선어말 어미)+-다(어말 어미)+.(문장 부호)		
구문 분석	문장의 문법적 적합성과 어절 의 구문적 역할(주어, 목적어 등)을 찾는 단계	ISURT: 나는   MOD: 그   ORT: 과자들		
의미 분석	문장을 구성하는 술어와 인자 들 사이의 의미적 적합성을 분석하는 단계	PREDICATE: 먹다 AGENT: 나/ANIMATE OBJECT: 그 과자/EATABLE		
담화 분석	대화 문맥을 파악하여 상호 참조를 해결하고 의도를 파악 하는 단계	SPEECH ACT: STATEMENT PREDICATE: 먹다 AGENT: 홍길동/ANIMATE OBJECT: 과자/EATABLE		

https://www.korean.go.kr/nkview/nklife/2017\_4/27\_0404.pdf

# 컴퓨터는 텍스트를 어떻게 처리할까?

토큰 텍스트 데이터에서 의미를 가지는 최소한의 덩어리

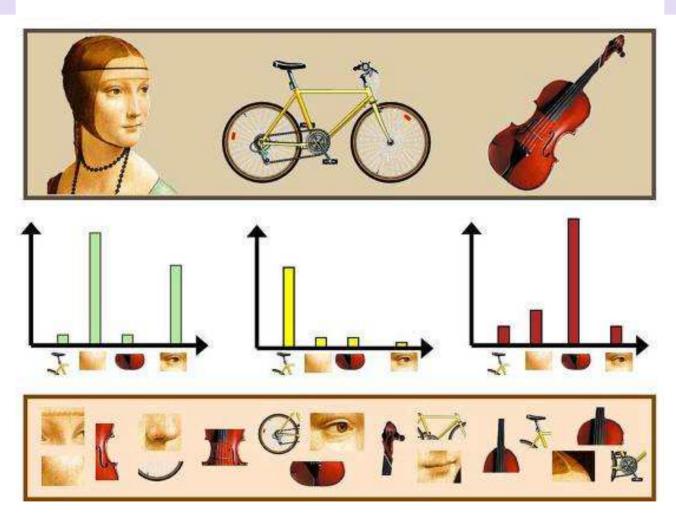
토큰화 텍스트 데이터를 토큰 단위로 나누는 과정

나는 너를 좋아한다



나/는/너/를/좋아한다

# Bag of Words를 이미지 분류에 사용한 예



# Bag of Words

나, 는, 너, 를, 좋아한다 너, 는, 나, 를, 좋아한다

나	느	너 를		좋아한다
2	2	2	2	2

약,이,쓰다 글,을,쓰다

약	0	쓰다	글	이미
1	1	2	1	1

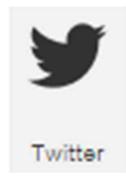
# SNS = PST EL

**g**The
Guardian





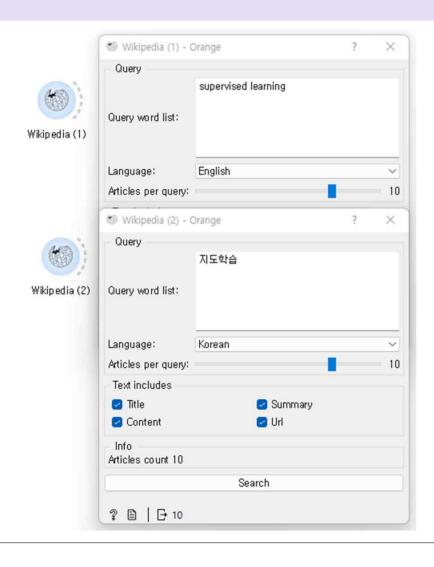


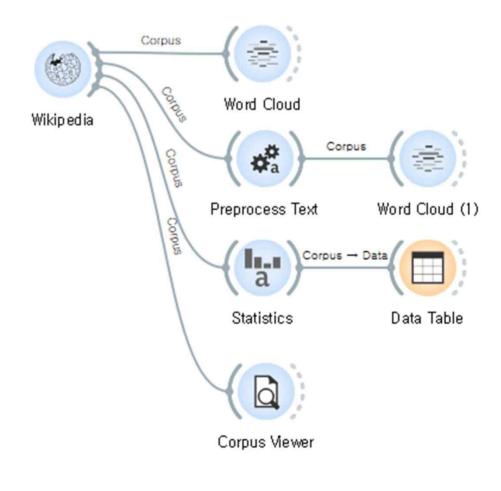




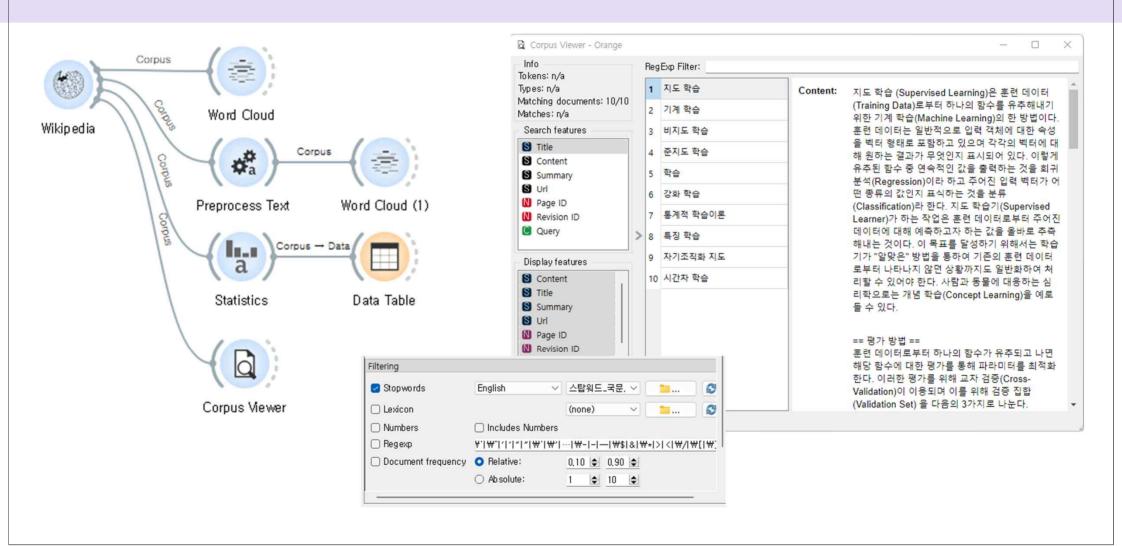


# 위키피디아에서 데이터 가져오기

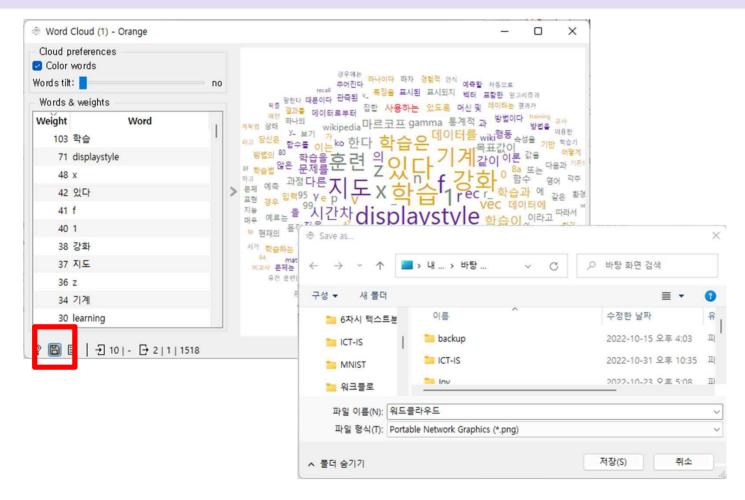




# 결과보기 및 전처리



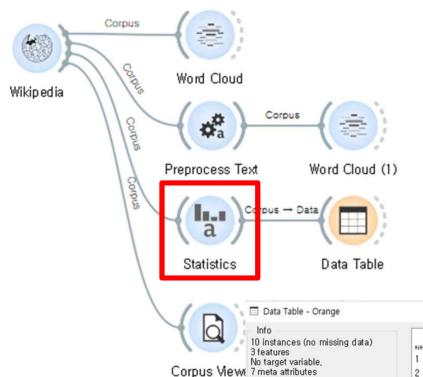
# 워드 클라우드 저장하기



경우에는 하나이다 하자 경험적 연석 역측할 자동으로 무어진다 "Fecal Profile Prof

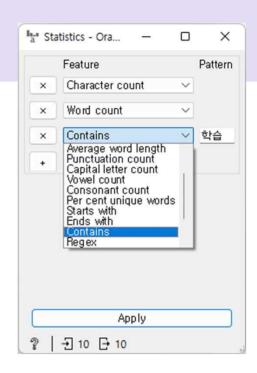
워드클라우드.png

# Statistics 위젯



✓ Show variable labels (if present)
 ✓ Msualize numeric values
 ✓ Color by instance classes

Selection
Select full rows



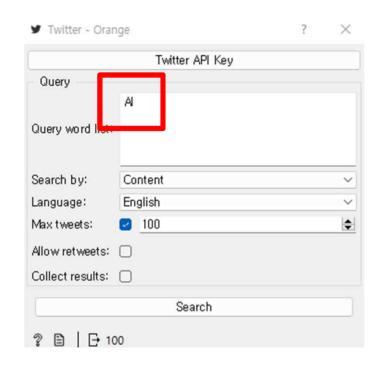
titla	Title True	Content	Summary	Url	Page ID	Revision ID	Query	Character count	Word count	Contains 학습
1	지도 학습	지도 학습 (Sup	지도 학습 (Sup	https://ko.wikip	383455	31553055	지도학습	1351	356	16
2	기계 학습	기계 학습(機械	기계 학습(機械	https://ko.wikip	128889	33055586	지도학습	1591	523	41
3	비지도 학습	비 지도 학습(U	비 지도 학습(U	https://ko.wikip	383468	31553000	지도학습	662	160	18
4	준지도 학습	준 지도 학습(영	준 지도 학습(영	https://ko.wikip	383481	31937219	지도학습	1159	369	27
5	학습	학습(學習) 또는	학습(學習) 또는	https://ko.wikip	686	33244805	지도학습	954	310	38
6	강화 학습	강화 학습(Reinf	강화 학습(Reinf	https://ko.wikip	49021	32844089	지도학습	2319	719	41
7	통계적 학습이론	통계적 학습이	통계적 학습이	https://ko.wikip	2494690	31603763	지도학습	2996	1022	31
8	특징 학습	특징 학습(featu	특징 학습(featu	https://ko.wikip	3088270	32170748	지도학습	225	45	11
9	자기조직화 지도	자기조직화지도	자기조직화지도	https://ko.wikip	111586	33508887	지도학습	892	213	4
10	시간차 학습	시간차 학습(Te	시간차 학습(Te	https://ko.wikip	1580458	31697575	지도학습	2293	773	35

# 감성 분석 (Sentiment Analysis)

# 트위터

#### API키 받기 설명서: https://blog.naver.com/aimento/222804924834



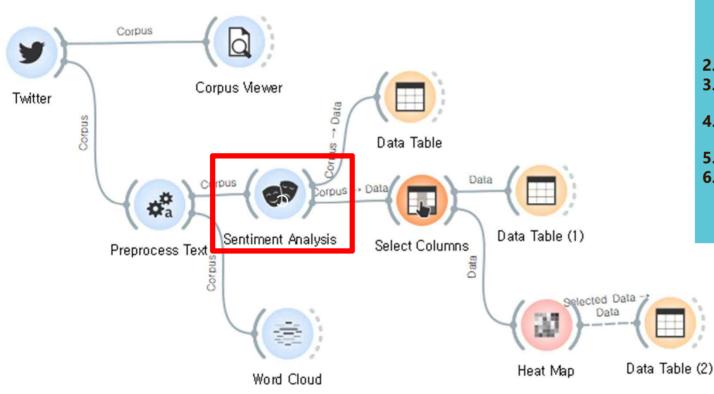




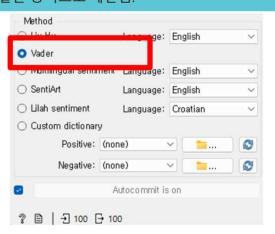


API 키 받는 방법 https://developer.nytimes.com/get-started

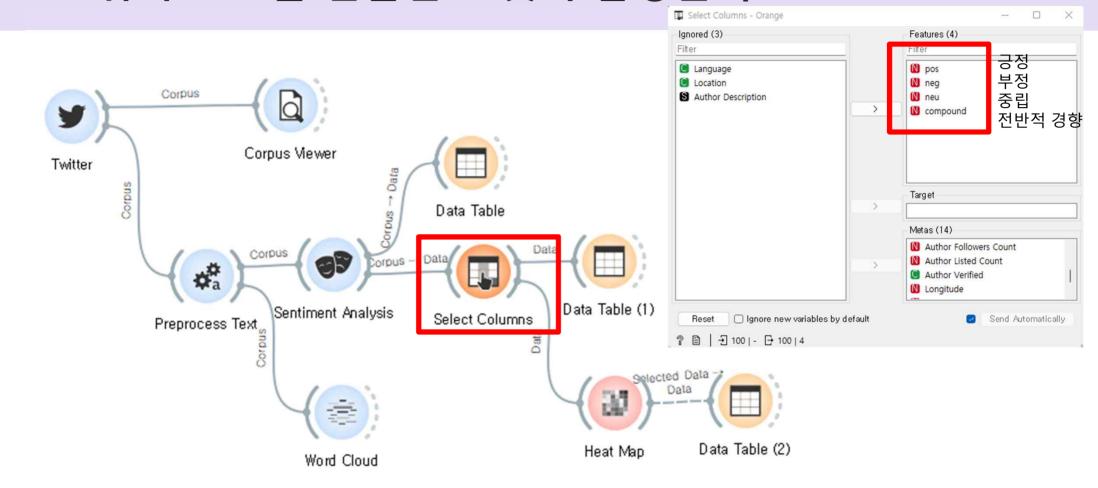
# 감성분석 Sentiment Analysis catalog/text-mining/sentimentanalysis/



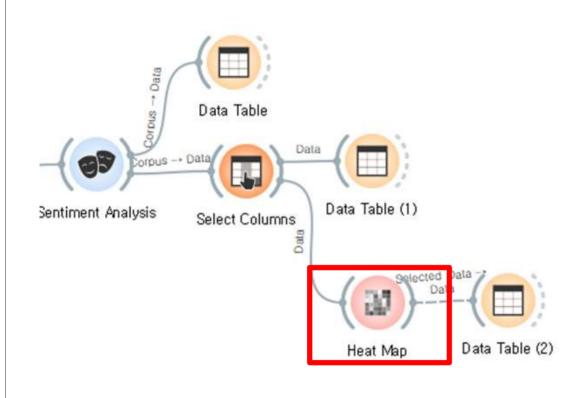
- 1. Liu Hu: 사전 기반 감정 분석(영어 및 슬로베니아어 지원). 최종 점수는 문서 길이로 정규화되고 100을 곱한 긍정적인 단어의 합계와 부정적인 단어의 합계 간의 차이로 나타냄 . 최종 점수는 문서의 감정 차이의 백분율을 반영.
- 2. Vader: 사전 및 규칙 기반 감정 분석
- **3. Multilingual sentiment** : 다국어 어휘집 기반 감성분석
- 4. SentiArt : 텍스트 원자가를 반환하는 벡터 공간 모델 기반 감정 분석
- 5. Lilah sentiment : NRC 감정 사전의 수동 번역
- 6. Custom dictionary: 자신의 긍정적이고 부정적인 감정 사전을 추가합니다. 허용되는 소스 유형은 한 줄에 각 단어가 포함된 .txt 파일입니다. 최종 점수는 Liu Hu와 같은 방식으로 계산됨.

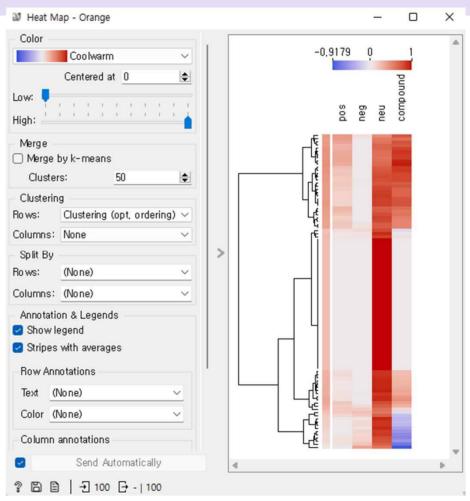


# 트위터 - AI 을 언급한 트윗의 감성분석

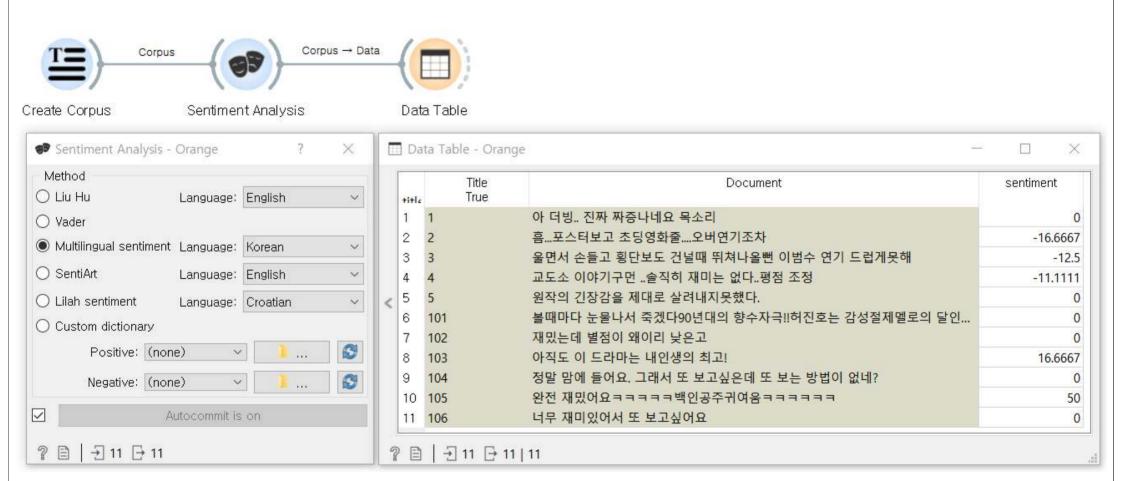


# Heatmap으로 나타내기



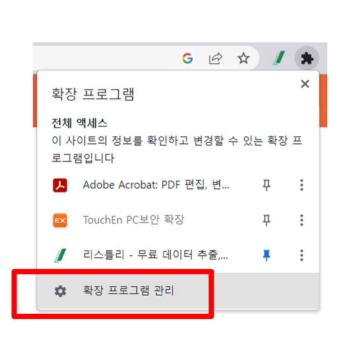


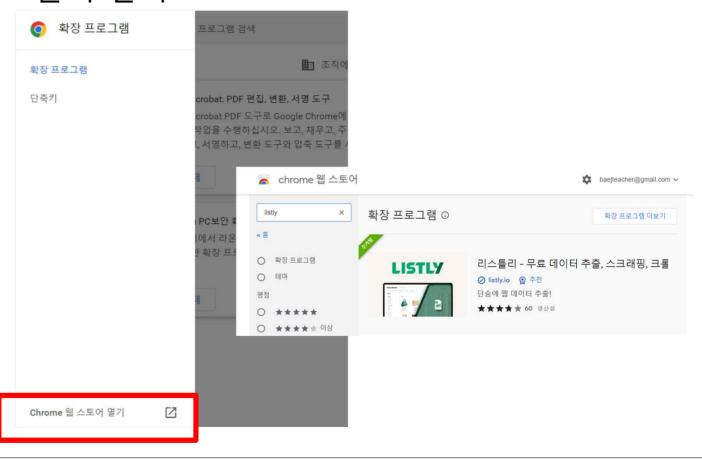
# 감성분석 - 한국어



# 댓글 분석을 위한 간단한 웹크롤링 도구

### ○구글의 확장프로그램 리스틀리 설치





# orange

다음 시간에는 연관분석을 활용한 추천시스템을 다루어 보도록 하겠 습니다.