



내 분야의 텍스트로 논문쓰기

# 텍스트마이닝 학습하기

강의와 교재를 평생소장하고 공부하세요

## 코딩 없이, 클릭만으로 할 수 있는 머신러닝과 딥러닝



### 데이터분석 분야 최고 실력자의 직강

23년간의 1,000건 이상의 데이터 분석 프로젝트 경험을 토대로 만든 고품질의 강의 제공

### 평생 소장 가능한 강의

평생 옆에 두고 전문가의 고품질 강의를 소장할 수 있는 기회!  
(전과목 USB+교재 제공)

### 데이터를 활용한 실습

데이터를 활용한 실습을 통해 AI분석의 이해도 향상!

# 클릭으로 완성하는 머신러닝과 딥러닝 올인원 패키지

USB(강의영상·실습파일) + 교재 제공!

## 1권. ORANGE와 핵심 마이닝

총 20강

## 2권. ORANGE 지도학습 마스터

총 35강

## 3권. ORANGE 비지도학습 마스터

총 19강

## 4권. ORANGE 텍스트와 이미지분석 마스터

총 30강

평생소장 가능 머신러닝 툴 ORANGE 강의와 함께라면  
머신러닝과 딥러닝 마스터 가능합니다😊



## ☰ 구매 가격

	정가	할인율	판매가
1과목 구매시	300,000	0%	300,000
2과목 구매시	600,000	10%	540,000
3과목 구매시	900,000	15%	765,000
4과목 구매시	1,200,000	20%	960,000

## ☰ 입금 및 문의 안내

### 상품구성

- 본 상품은 USB+교재로 구성되어 있습니다
- 상품은 입금 확인 후 택배로 배송됩니다

### 계좌이체 안내

- 우리은행, 1005-402-421172, (주)와이즈인컴퍼니

### 문의사항 및 계산서/견적서 요청

- 연락처: 070-8676-1312
- 이메일: hs9177@wiseinc.co.kr

대학의 경우 각 대학 도서관에 구매요청을 하실 수 있습니다



# 김 원 표

現 (주)와이즈인컴퍼니 대표  
한양대 겸임교수

### 【 주요 경력 】

- 20년간 [2,000건](#) 이상의 통계분석/ 빅데이터 프로젝트 수행
- [20권](#) 이상 통계분석/ 빅데이터 관련 서적 출간
- 연간 [2만명](#) 이상의 수강생이 검증한 전문 강사
- AI 통계, 데이터분석 솔루션 " [데이터인\(DataIN\)](#) " 개발 기획 총괄
- 국책연구소 인공지능 (딥러닝) 프로젝트 다수 수행

# CONTENTS

---



## 텍스트마이닝 학습하기

1. 텍스트마이닝의 핵심지식 담기
2. 데이터 불러오기와 워드 클라우드
3. 감성분석
4. 토픽모델링과 LDA시각화



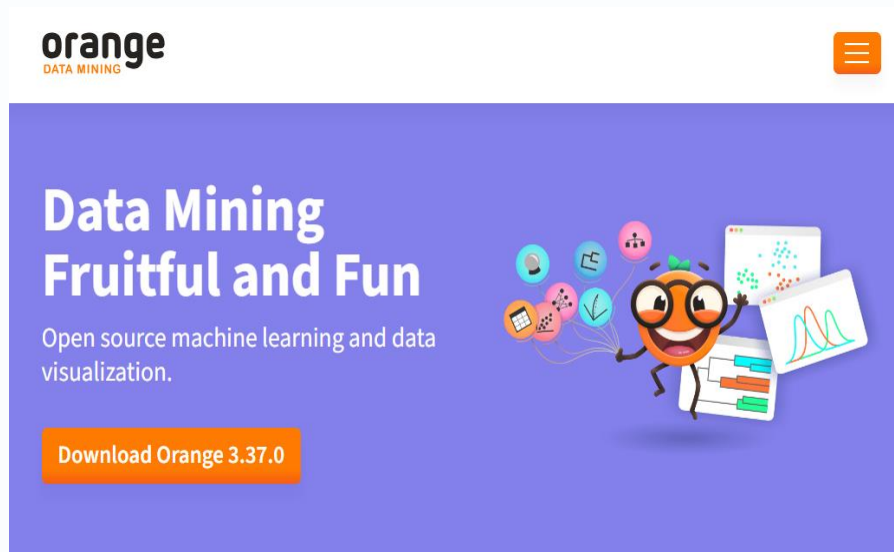
# Part 01

## 텍스트마이닝의 핵심지식 담기



## 1. 무료 머신러닝 솔루션 ORANGE3 활용

### ORANGE3

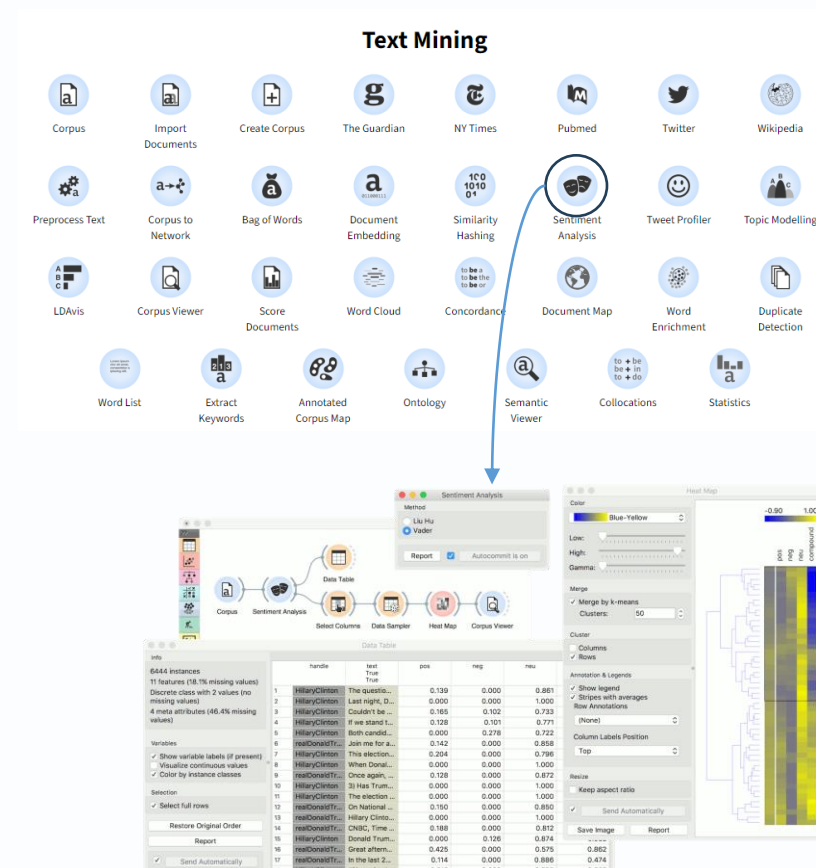


다운로드

<https://orangedatamining.com/download/>

텍스트마이닝 상세 설명법

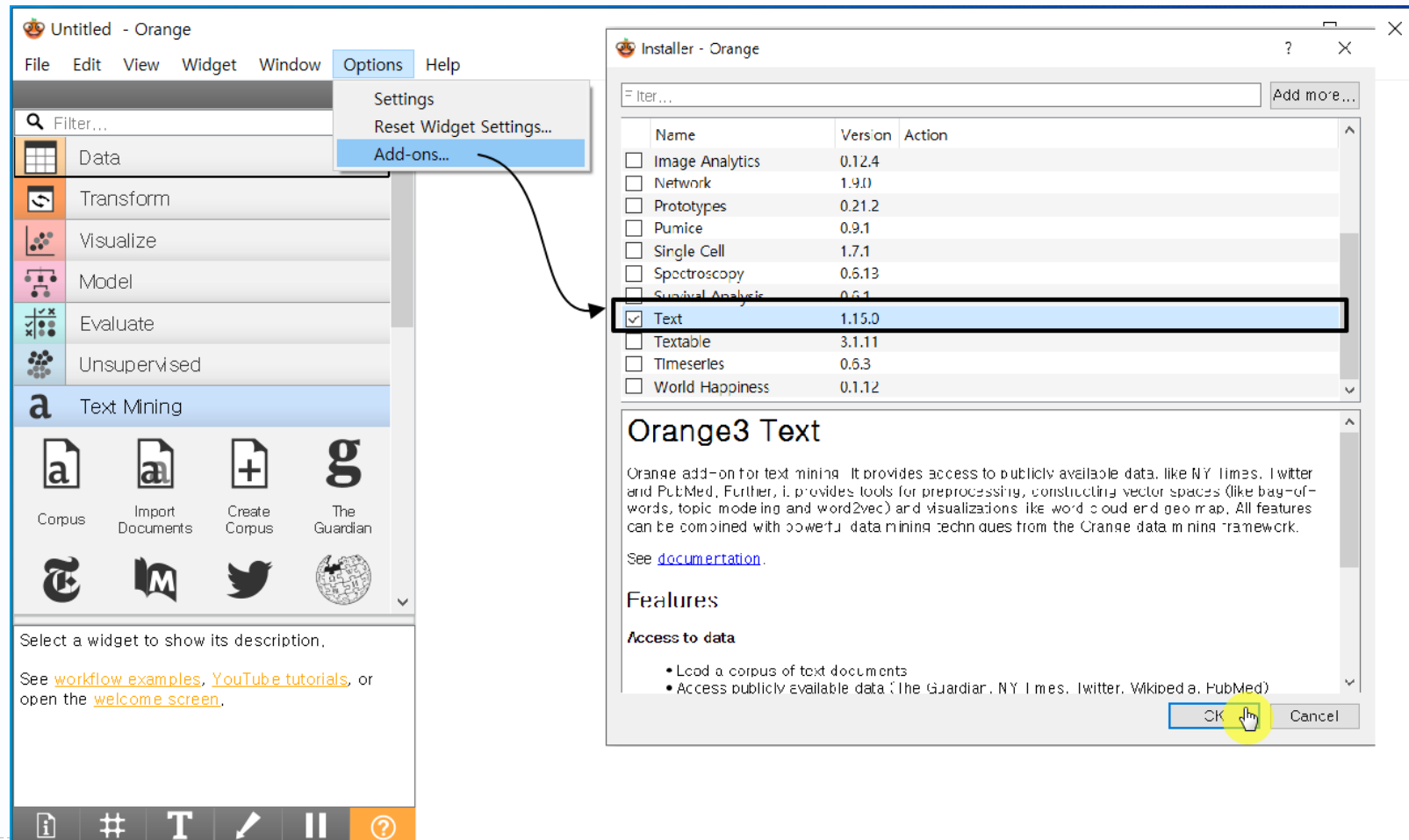
<https://orangedatamining.com/widget-catalog/>





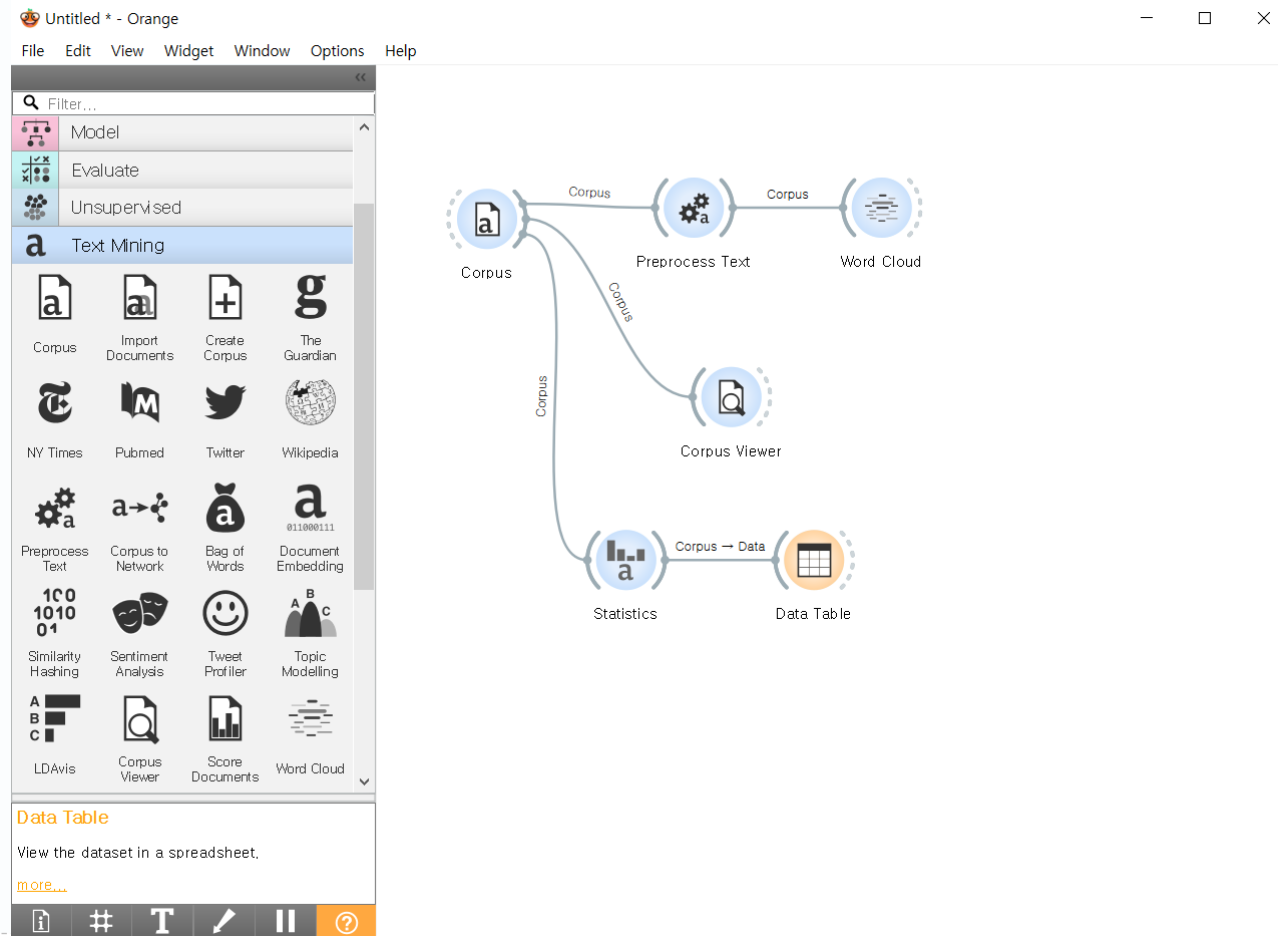
## 1. 무료 머신러닝 솔루션 ORANGE3 활용

- 텍스트마이닝 위젯 추가방법: Options > Add-ons 클릭 후, Text 선택하여 'OK'로 추가 설치 (꽤 많은 시간이 소요됨)



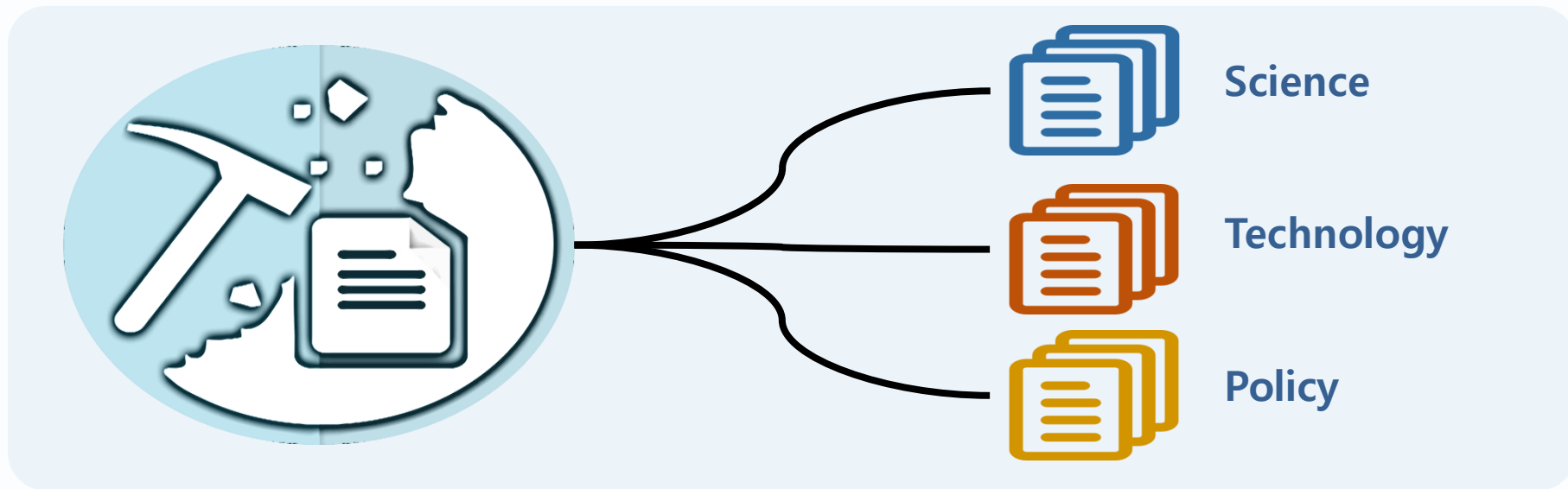
## 1. 무료 머신러닝 솔루션 ORANGE3 활용

- Text Mining 기능이 추가 설치됨. 왼쪽의 위젯을 오른쪽 그림으로 프로세스에 따라 설정하여 분석 수행



### 2. 텍스트마이닝 기본 개념

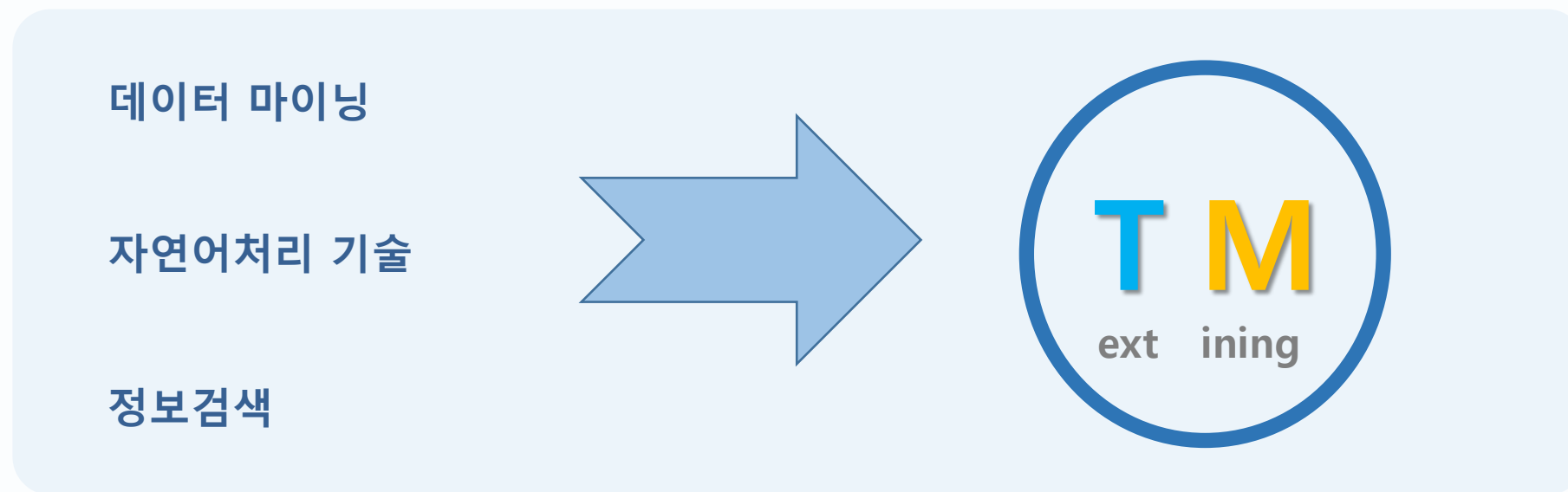
- 텍스트 형태로 이루어진 정형/비정형 데이터들을 자연어 처리 방식(Natural Language Processing)과 문서처리 방법을 적용하여 유용한 정보를 추출하고 가공하는 기술



- 텍스트란 일반 문서나 도서뿐만 아니라 웹페이지, 블로그, 전자저널, 이메일 등 전자문서도 포괄한 자료원천으로 우리 일상에서 가장 사람과 가까운 형태의 정보

### 2. 텍스트마이닝 기본 개념

- 텍스트 마이닝은 데이터로부터 유용한 인사이트를 발굴하는 **데이터마이닝(Data Mining)**, 언어를 정보로 변환하기 위한 **자연어처리(Natural Language Process)**, **정보검색** 등 다양한 분야가 접목되어 발전한 학문, 기술

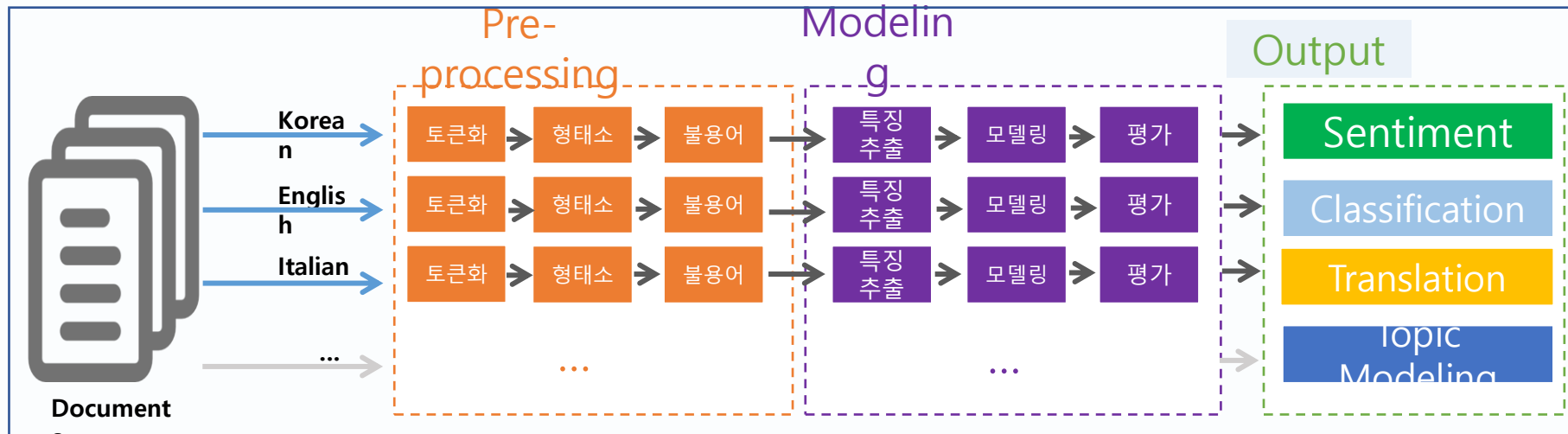
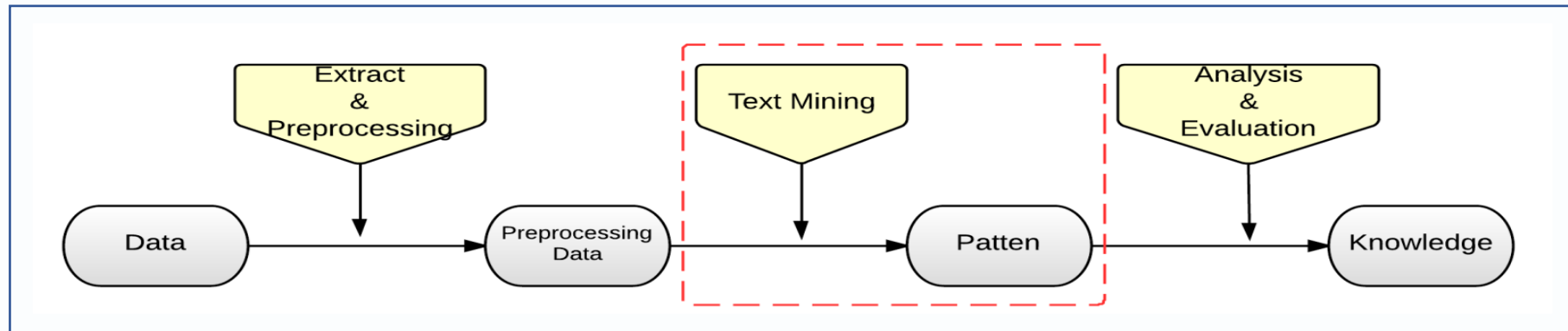


- 즉, 텍스트마이닝이란 텍스트 데이터로부터 새로운 고급 정보를 이끌어 내는 과정이라 할 수 있음

### 2. 텍스트마이닝 기본 개념

- **문서 요약** (Summarization) : 논문/신문/보고서 요약
- **문서 분류** (Classification) : 자동 범주화 (예: 뉴스 기사 분석 -> 사회/경제/금융 등으로 분류)
- **문서 군집** (Clustering) : 유사 단어 또는 유사 문서 간의 군집 분석
- **특징 추출** (Feature Extraction) : 주요 키워드 추출

## 2. 텍스트마이닝 기본 개념





## 2. 텍스트마이닝 기본 개념

### 가. Sentiment Analysis

- 텍스트 마이닝의 한 분야로 **사용자의 의견, 감정, 태도** 등을 텍스트로부터 분석하는 방법
- 텍스트 내 키워드와 관련된 감성 어휘의 빈도수를 분석하여 **긍정/부정/중립**으로 분류하고 사용자의 의견을 분석함  
(예 :상품평, 후기, 선호도 등)

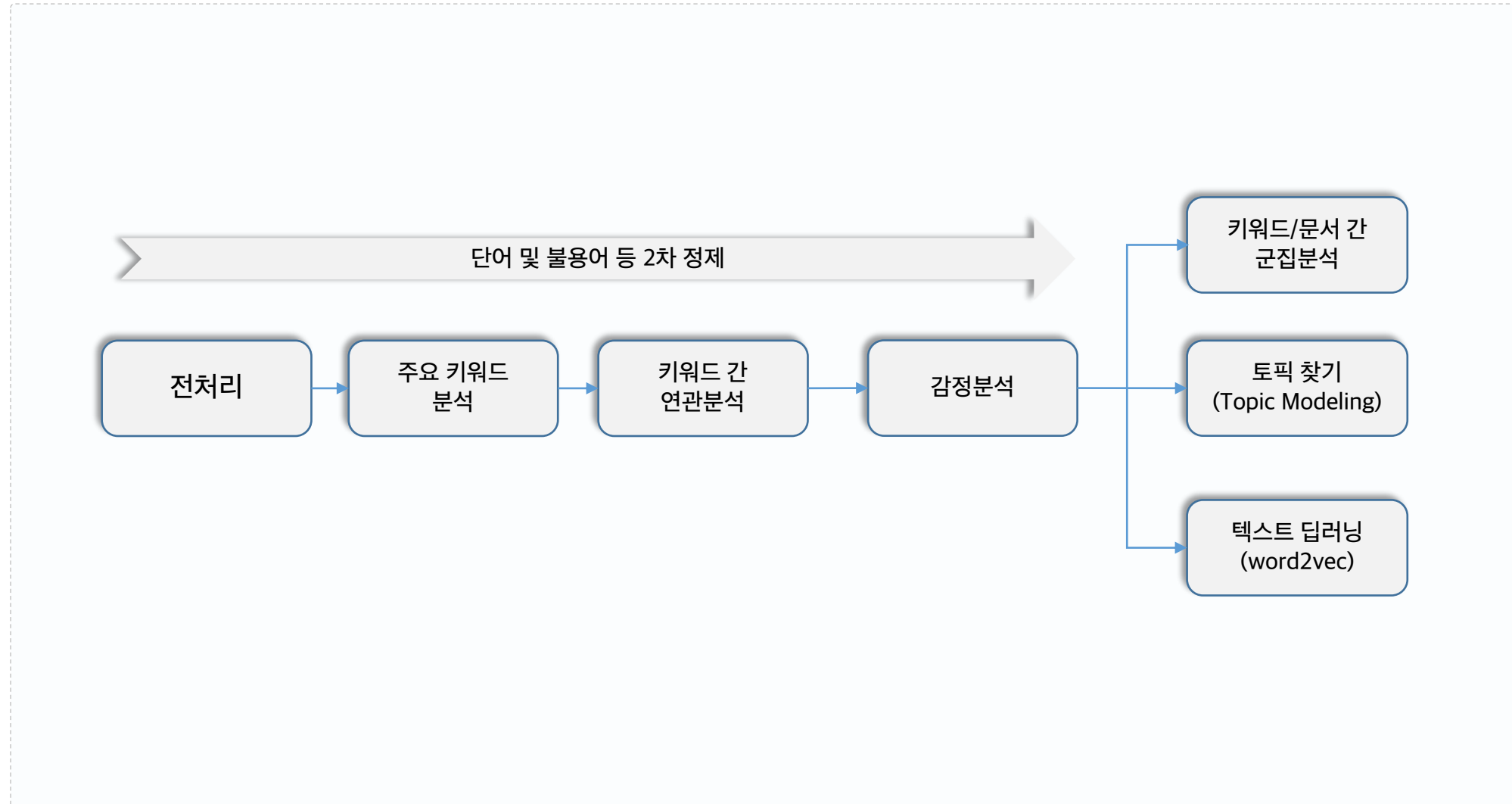
### 나. Topic Modeling

- 문서를 이루고 있는 키워드들을 바탕으로 문서에서 **주제(Topic)**를 도출하기 위해 사용되는 **통계적 분석방법**
- **비정형 텍스트 분석**에 많이 사용되고 있으며 다양한 종류의 데이터에도 적용이 가능

### 다. Word2Vec/Doc2Vec

- 단어를 벡터 공간으로 표현하는 방법으로 인공지능 기반의 임베딩 기법
- 단어 사이의 **분포 관계**를 바탕으로 단어를 일정한 의미를 갖는 **벡터로 변환**하는 기술
- Word2Vec 알고리즘의 확장으로 **문장, 단락 또는 전체 문서**와 같이 더 큰 텍스트를 표현

### 3. 텍스트마이닝 프로세스



# Part 02

## 데이터 불러오기와 워드 클라우딩

## 1. ORANGE 구성

The screenshot displays the Orange3 data mining software interface, illustrating the workflow for generating a word cloud from a corpus.

**Workflow Overview:**

- Corpus:** The initial data source, represented by a document icon.
- Preprocess Text - Orange:** A widget used for text preprocessing, indicated by a blue box and a red circle labeled '1'.
- Preprocessors:** A list of available preprocessing widgets, including 'Preprocess Text', 'Corpus to Network', 'Bag of Words', and 'Document Embedding'.
- Word Cloud - Orange:** A widget for generating word clouds, indicated by a red circle labeled '2'.
- Transformation:** A widget for applying transformations to the data, indicated by a red circle labeled '3'.
- Tokenization:** A widget for tokenizing the text, indicated by a red circle labeled '3'.
- Filtering:** A widget for filtering the data, indicated by a red circle labeled '3'.

**Widget Settings:**

- Corpus - Orange:**
  - Corpus file: `ndsl_korean_t20.csv` (indicated by a red circle labeled '1')
  - Language: `Korean`
  - Used text features: `제목` (indicated by a red circle labeled '2')
  - Ignored text features: `논문명`, `저자`
- Word Cloud - Orange:**
  - Cloud preferences: ☒ Color words
  - Words tilt: `30°`
  - Words & weights:
 

Weight	Word
138	인공지능
53	데이터
- Transformation:**
  - ☒ Lowercase
  - ☐ Remove accents
  - ☐ Parse html
  - ☐ Remove urls
- Tokenization:**
  - ☐ Word Punctuation
  - ☐ Whitespace
  - ☐ Sentence
  - ☒ Regexp
  - Pattern: `ww+`
  - ☐ Tweet
- Filtering:**
  - ☒ Stopwords: `English`, `쓸모없는단어.txt` (indicated by a red circle labeled '3')
  - ☐ Lexicon: `(none)`
  - ☒ Numbers: ☐ Includes Numbers
  - ☒ Regexp: `[W|W'|'|'|'W|W'|'W-|-|W$|&|W+|>|<|W|W|W]`
  - ☐ Document frequency: ☒ Relative: `0.10`, `0.90`
  - ☐ Absolute: `1`, `10`
  - ☐ Most frequent tokens: `100`
  - ☐ POS tags: `NOUN, VERB`

**Output:**

- Tokens: 21931
- Types: 10277
- Apply Automatically: ☒

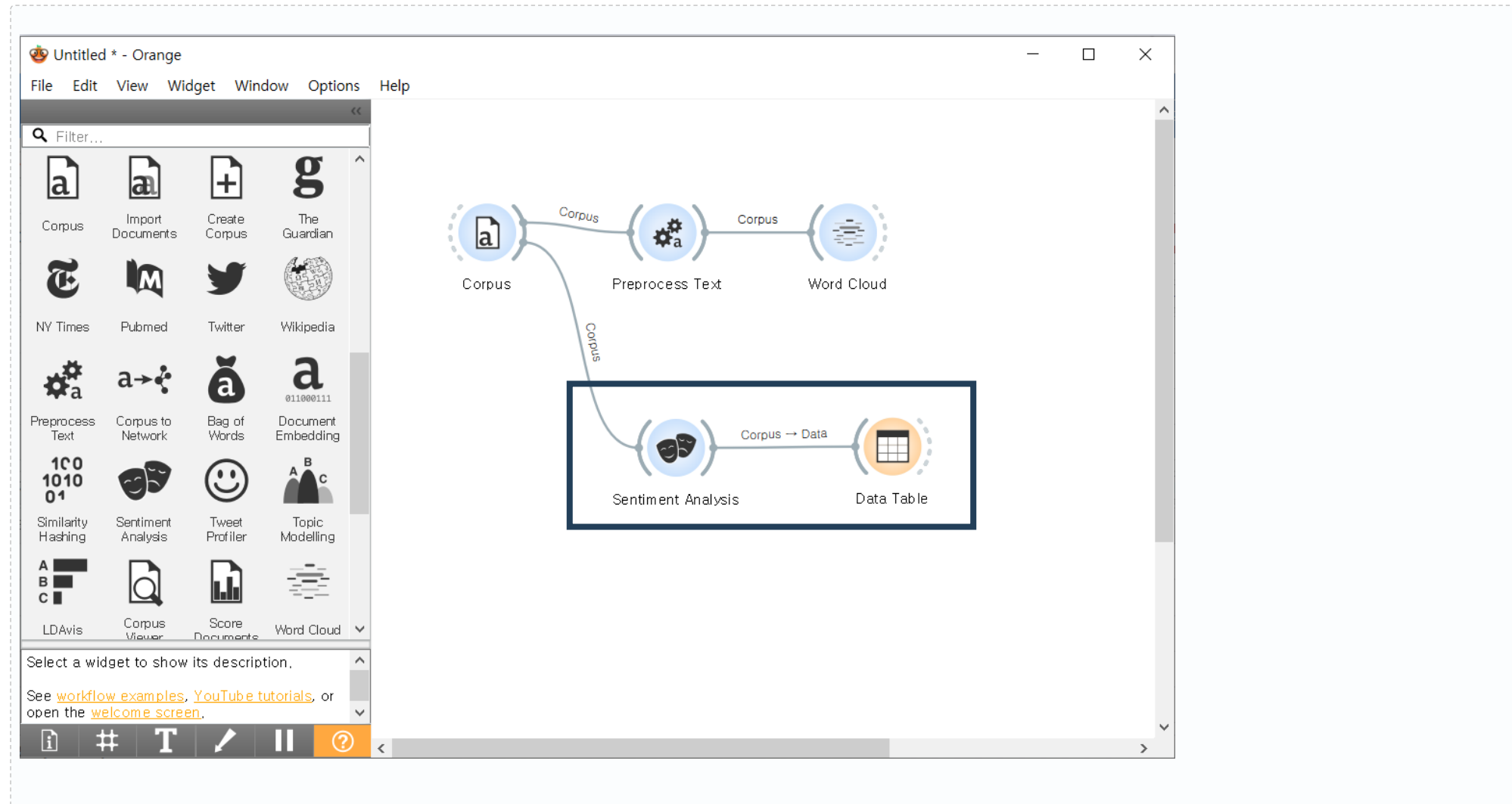
**Word Cloud:**

The word cloud displays the results of the preprocessing and filtering steps, showing the frequency of words in the corpus. The words are colored and sized according to their frequency, with '인공지능' (Artificial Intelligence) and '데이터' (Data) being the most prominent.

# Part 03

## 감성분석

## 1. ORANGE 구성

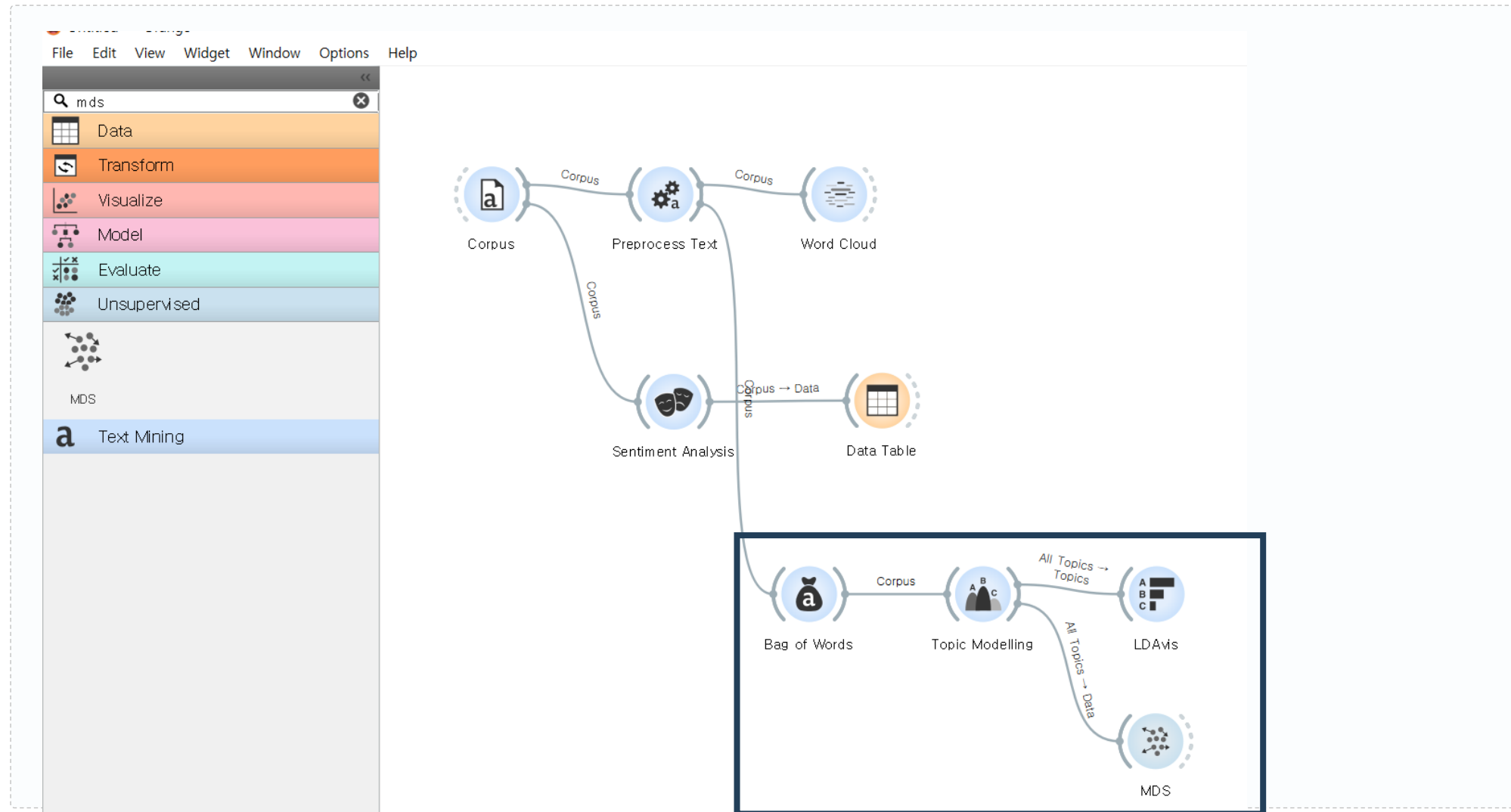




# Part 04

## 토픽모델링과 LDA시각화

## 1. ORANGE 구성





# THANK YOU

(주)와이즈인컴퍼니 / 서울시 강남구 언주로 309, 기성빌딩 3층 / T 02.558.5144 / F 02.558.5146