

텍스트 마이닝

홍익 대학교
Hyun-Sun Ryu

텍스트 마이닝(Text mining)의 실습

텍스트 마이닝 프로세스



텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

John F. Kennedy 연설문

< John F. Kennedy 연설문을 바탕으로 연설문 분석하고
분류하기>

텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

- https://www.jfklibrary.org/archives/other-resources/john-f-kennedy-speeches



Home > Archives > Other Resources

Research Guides by Subject

John F. Kennedy: Speeches

John F. Kennedy: Press
Conferences

Historypin

Goodreads

NARA Catalog

National Security Action
Memoranda (NSAMs)

John F. Kennedy: Executive
Orders

All

Remarks of John F. Kennedy at an Induction Ceremony for Navy Recruits, Charleston, South
Carolina, July 4, 1942

July 4, 1942

[LEARN MORE](#)

Remarks of John F. Kennedy, United War Fund Appeal, Boston, Massachusetts, October 8, 1945

October 8, 1945

[LEARN MORE](#)

텍스트 마이닝 프로세스

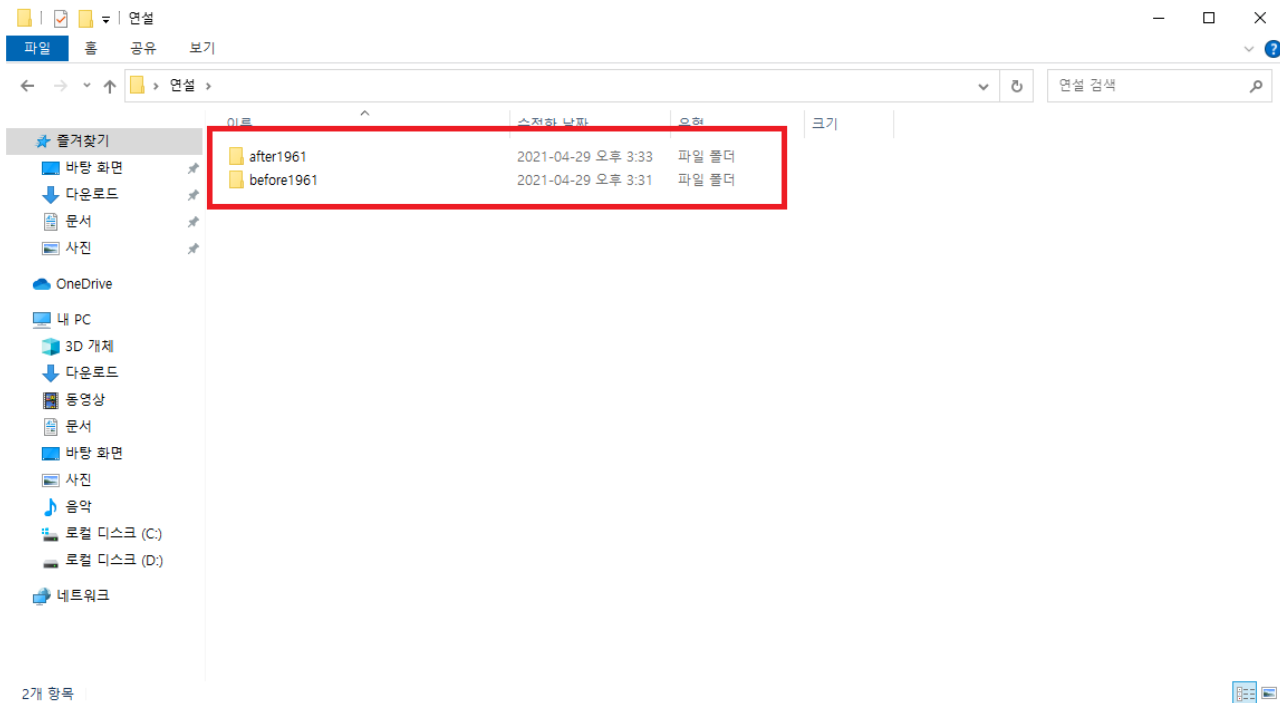
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스


데이터 수집

전처리

모델생성

학습/분류

평가

위젯	설명	입력	출력
 Import Documents	폴더에서 텍스트 문서를 가져온다.	x	Corpus, Skipped documents

- Import Documents 위젯을 사용하면 텍스트 파일을 검색하고 말뭉치를 작성
- 위젯에서는 .txt, .docx, .odt, .pdf 및 .xml 파일을 읽음.
- 폴더에 하위 폴더가 있으면 클래스 레이블로 사용

텍스트 마이닝 프로세스

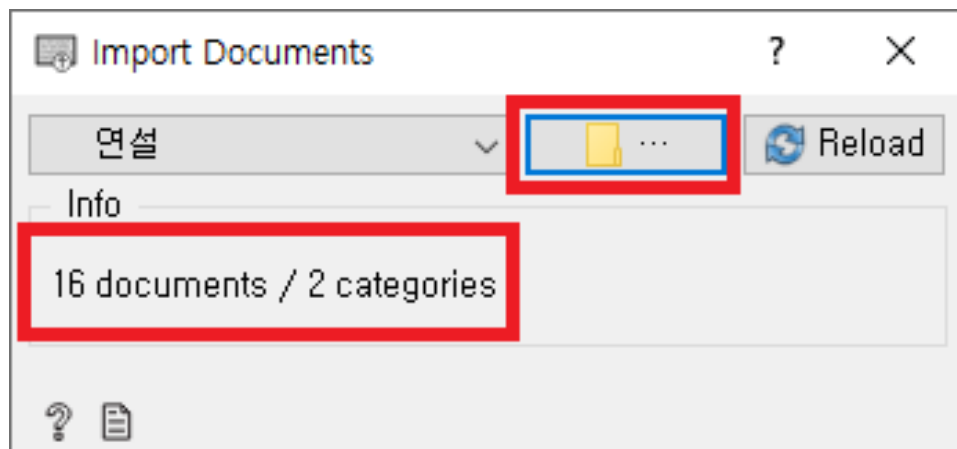
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가



Corpus Viewer

Info

Documents: 16
Preprocessed: False
• Tokens: n/a
• Types: n/a
POS tagged: False
N-grams range: 1-1
Matching: 16/16

Search features

- category
- name
- path
- content

Display features

- category
- name
- path
- content

☐ Show Tokens & Tags

☒ Auto send is on

RegExp Filter:

1	1961-01-09
2	1961-01-20
3	1961-04-27
4	1962-01-20
5	1962-06-11
6	1963-07-26
7	1963-10-26
8	1963-11-22
9	1942-07-04
10	1946-05-19
11	1947-03-10
12	1949-02-21
13	1950-05-26
14	1955-10-28
15	1959-06-20
16	1960-04-21

category: after1961
name: 1961-01-09
path: C:/Users/user/Desktop/연설\after1961\1961-01-09.txt
content: I have welcomed this opportunity to address this historic body, and, through you, the people of Massachusetts to whom I am so deeply indebted for a lifetime of friendship and trust.

For fourteen years I have placed my confidence in the citizens of Massachusetts--and they have generously responded by placing their confidence in me.

Now, on the Friday after next, I am to assume new and broader responsibilities. But I am not here to bid farewell to Massachusetts.

For forty-three years--whether I was in London, Washington, the South Pacific, or elsewhere--this has been my home; and, God willing, wherever I serve this shall remain my home.

It was here my grandparents were born--it is here I hope my grandchildren will be born.

I speak neither from false provincial pride nor artful political flattery. For no man about to enter high office in this country can ever be unmindful of the contribution this state has made to our national greatness.

Its leaders have shaped our destiny long before the great republic was born. Its principles have guided our footsteps in times of crisis as well as in times of calm. Its democratic institutions--including this historic

텍스트 마이닝 프로세스

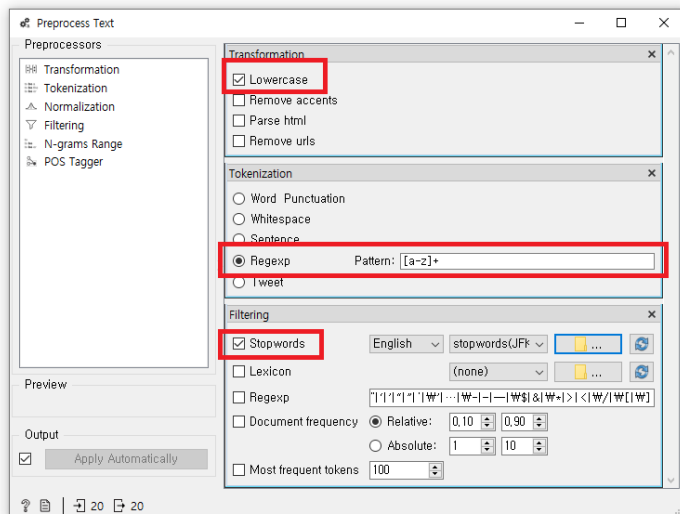
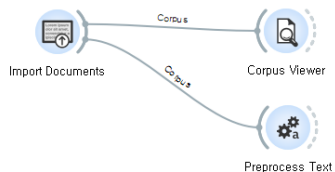
데이터 수집

전처리

모델생성

학습/분류

평가



<데이터 전처리>

1. 소문자
2. 알파벳만 활용 [a-z]+
3. 제외 문자

텍스트 마이닝 프로세스

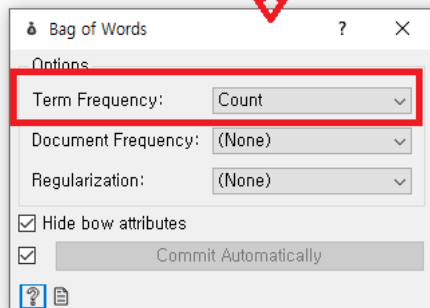
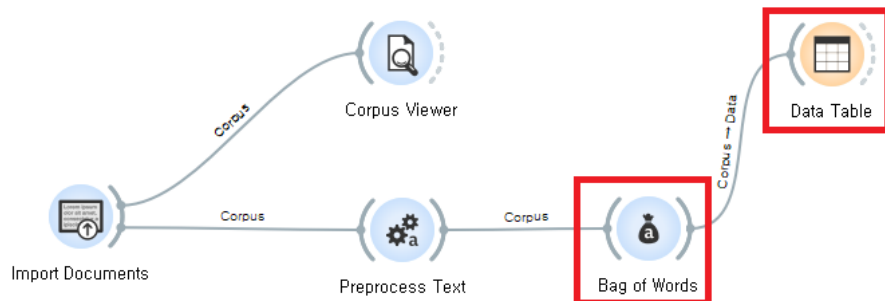
데이터 수집

전처리

모델생성

학습/분류

평가



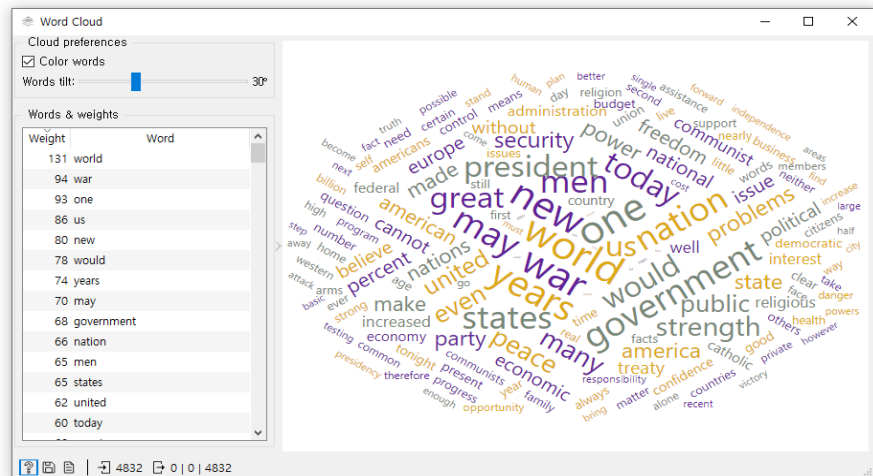
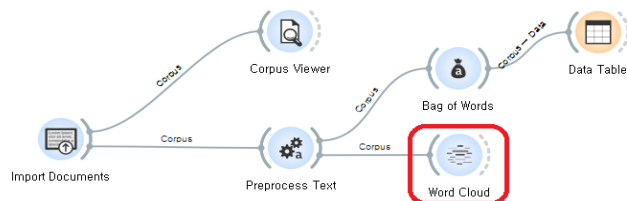
텍스트 마이닝 프로세스

전처리

모델 생성

학습/분류

평가



텍스트 마이닝 프로세스

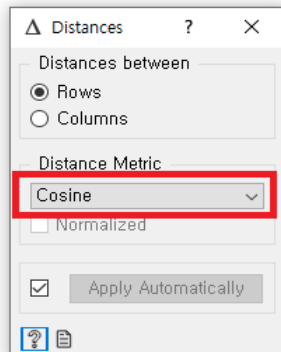
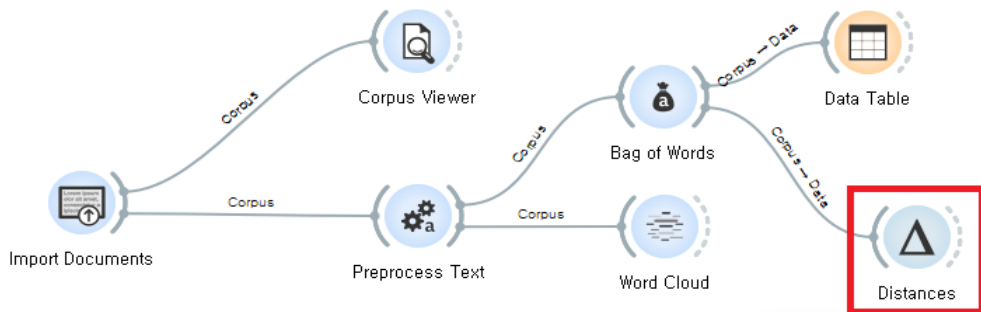
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

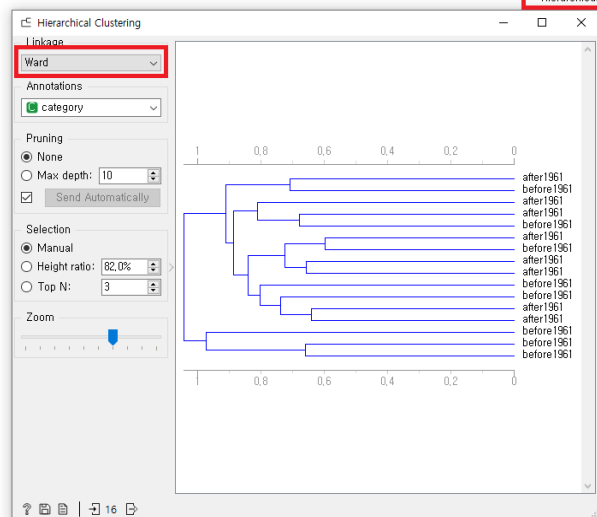
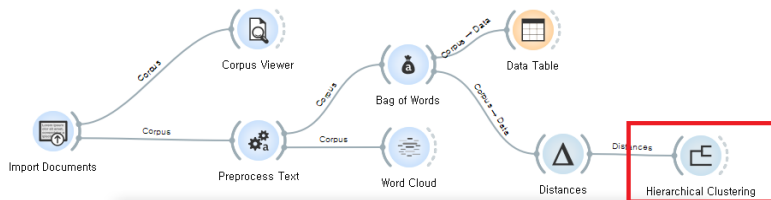
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

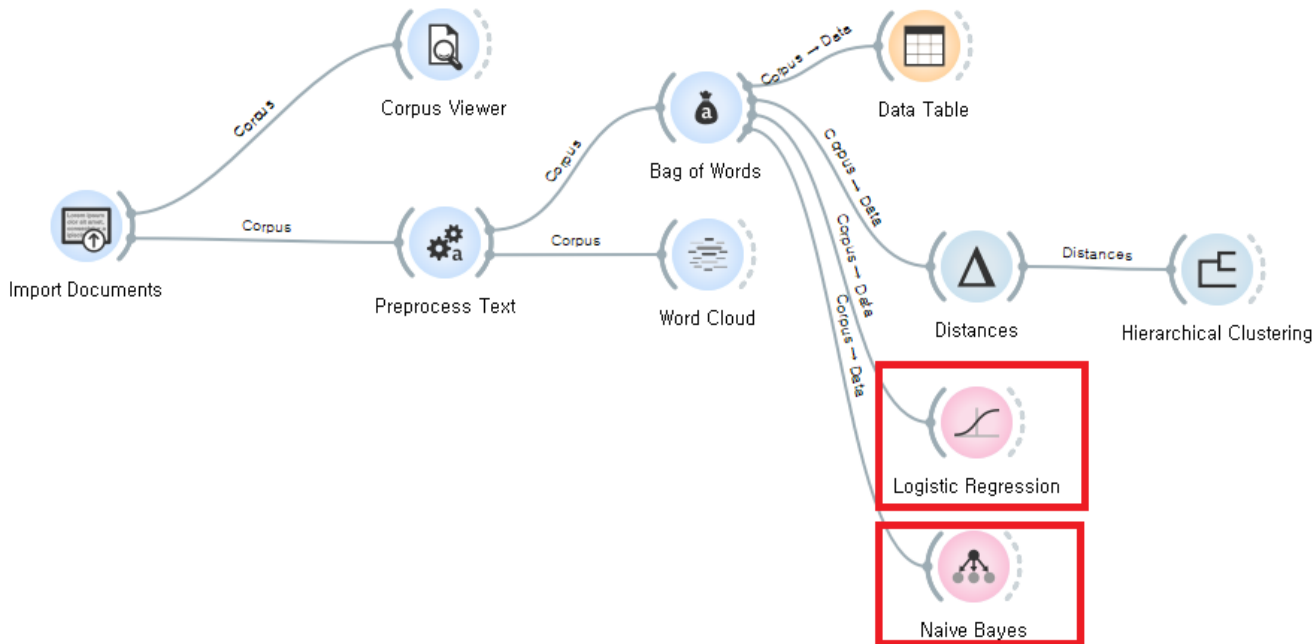
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

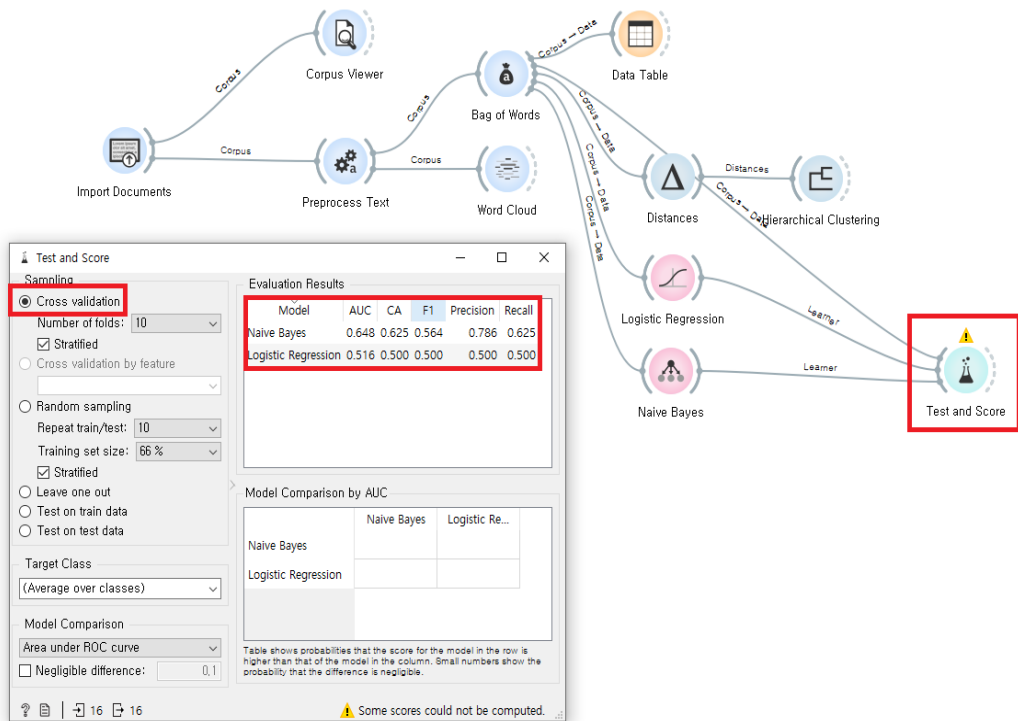
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

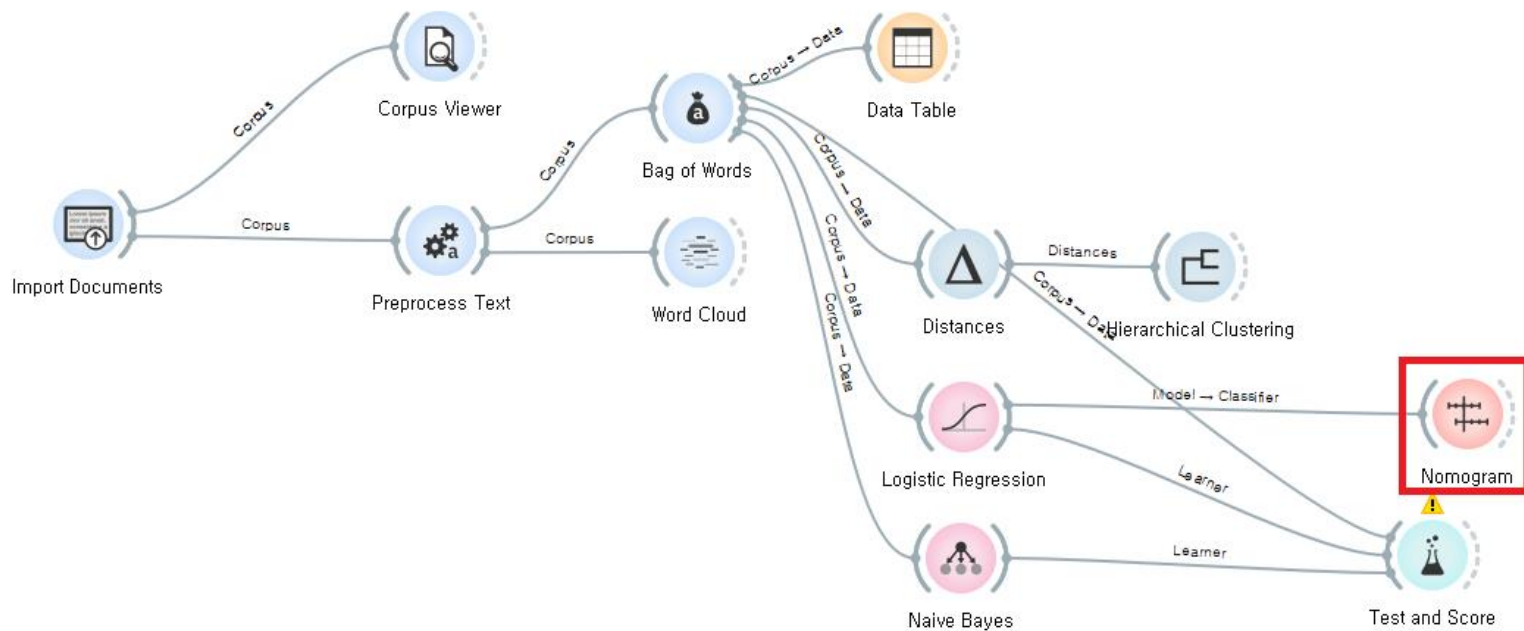
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

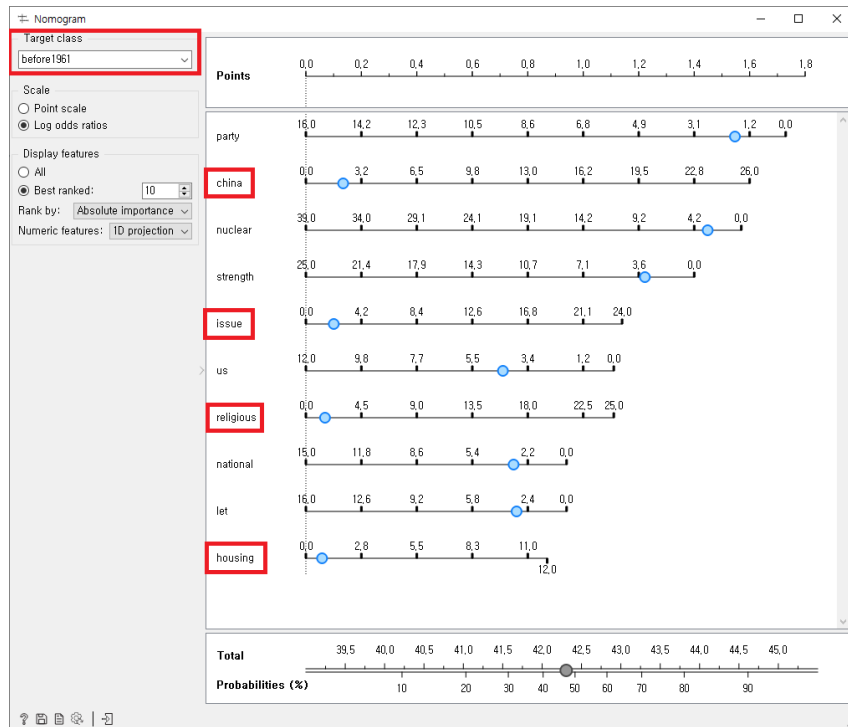
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

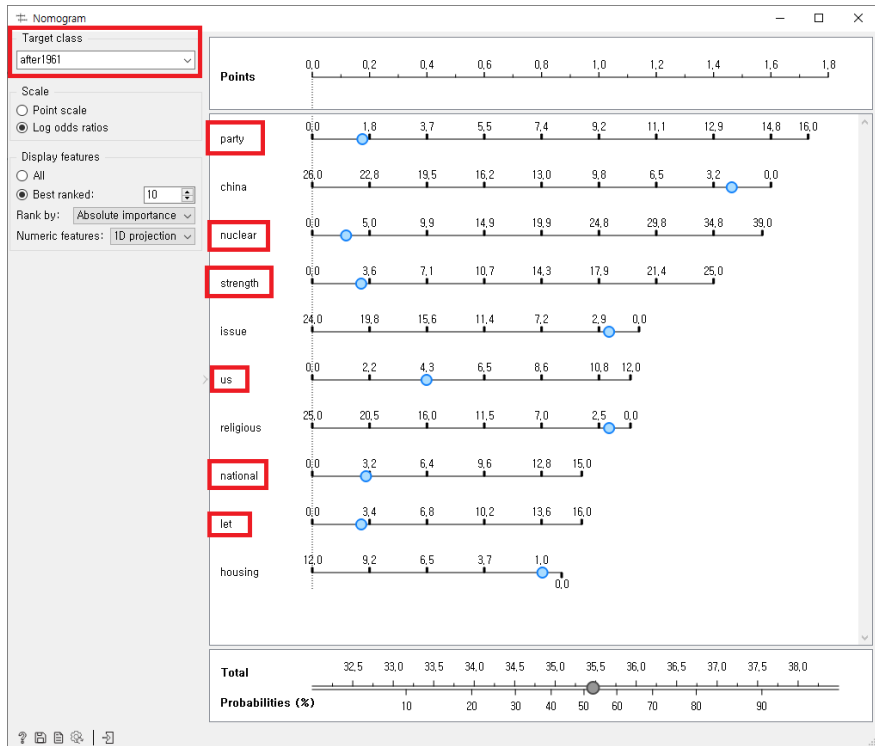
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

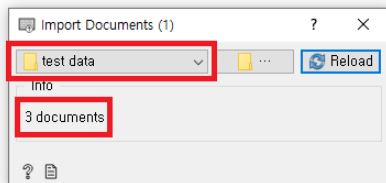
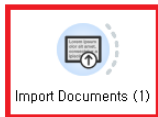
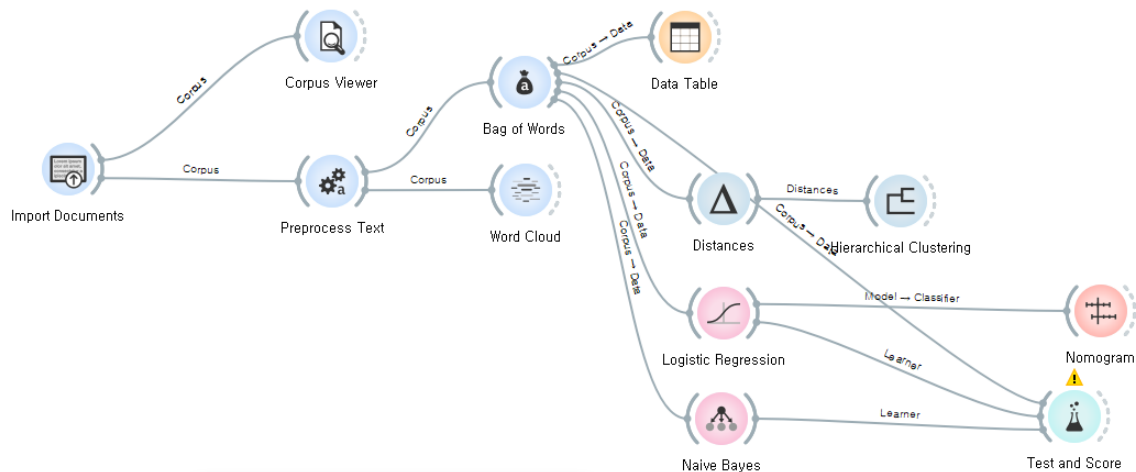
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

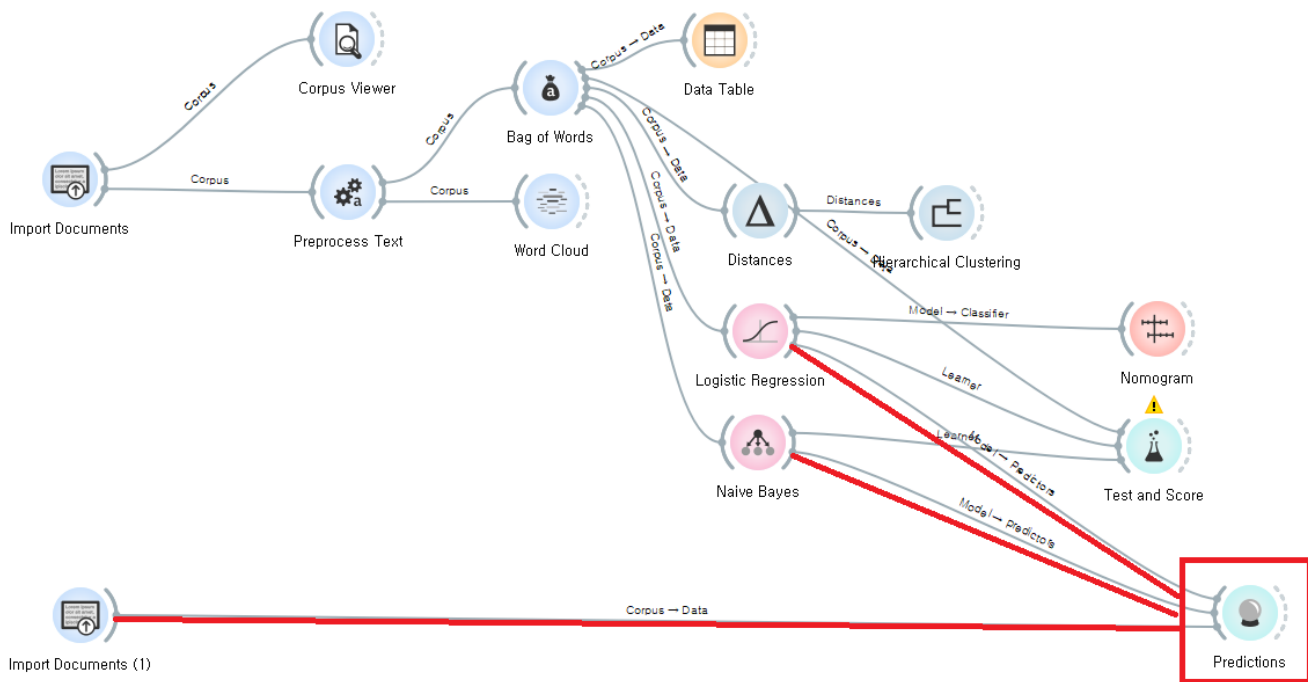
데이터 수집

전처리

모델생성

학습/분류

평가



텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

Predictions

Show probabilities for

after1961
before1961

	Logistic Regression	Naive Bayes	name	path	content
1	after1961	before1961	test1(1945-10-...	C:/Users/user/...	These are the ...
2	before1961	before1961	test2(1958-03-...	C:/Users/user/...	I have just ...
3	after1961	after1961	test3(1962-11-...	C:/Users/user/...	Dr. Kinsolving, ...

Restore Original Order

3 3

텍스트 마이닝(Text mining) 실습과제

텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

Corpus 위젯을 활용해 각자 분석을 원하는
Corpus를 업로드

텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

1. 업로드한 데이터를 바탕으로 Preprocess 과정을 거침
2. 전처리 된 Corpus를 wordcloud 위젯 등을 활용해 시각화

텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

Bag of words 위젯을 활용한 결과에 따라

Words별 distance를 계산하고

이에 따라 군집화를 하고 모델을 생성

텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

분류하고자 하는 새로운 Corpus를 불러옵니다.

기존에 학습한 모델과 연결시켜

분류작업을 시행합니다..

텍스트 마이닝 프로세스

데이터 수집

전처리

모델생성

학습/분류

평가

Test and Score 위젯을 활용해

분류결과를 확인해보고

Predictions 위젯이 새로운 데이터를 어떻게 예측하는지
확인합니다.

질문 있나요?

hsryu13@hongik.ac.kr

