

텍스트 마이닝

홍익 대학교
Hyun-Sun Ryu

데이터의 종류

정형 데이터



고객정보 데이터
매출 데이터
회계 데이터
재고 데이터

비정형 데이터



텍스트 마이닝(Text mining)

텍스트 마이닝이란

1. 비정형화된 데이터를 처리하는 것으로 뉴스 기사, 동화, 리뷰 사이트 등의 내용에서 단어를 추출하거나 주제별로 정리하거나 이메일 스팸 분류 등을 할 수 있음.
2. 텍스트 데이터에서 가치와 의미가 있는 정보를 찾아내는 기법

텍스트 마이닝 용어정리

코퍼스
corpus

말뭉치 - 관련된 문서들의 집합

토큰
token

기호에 의해서 나누어진 기본 단위로
문장을 구분 말뭉치의 가장 작은 단위

파싱 parsing
토큰화 tokenization

텍스트의 단어, 절을 분리하는 작업

텍스트 마이닝 용어정리

어간 추출

stemming

단어의 어간을 추출하는 과정(eating->eat)

불용어

Stop words

의미 없는 단어(the, and, of 등) (common words)

용어-문서 행렬

Term-Document Matrix

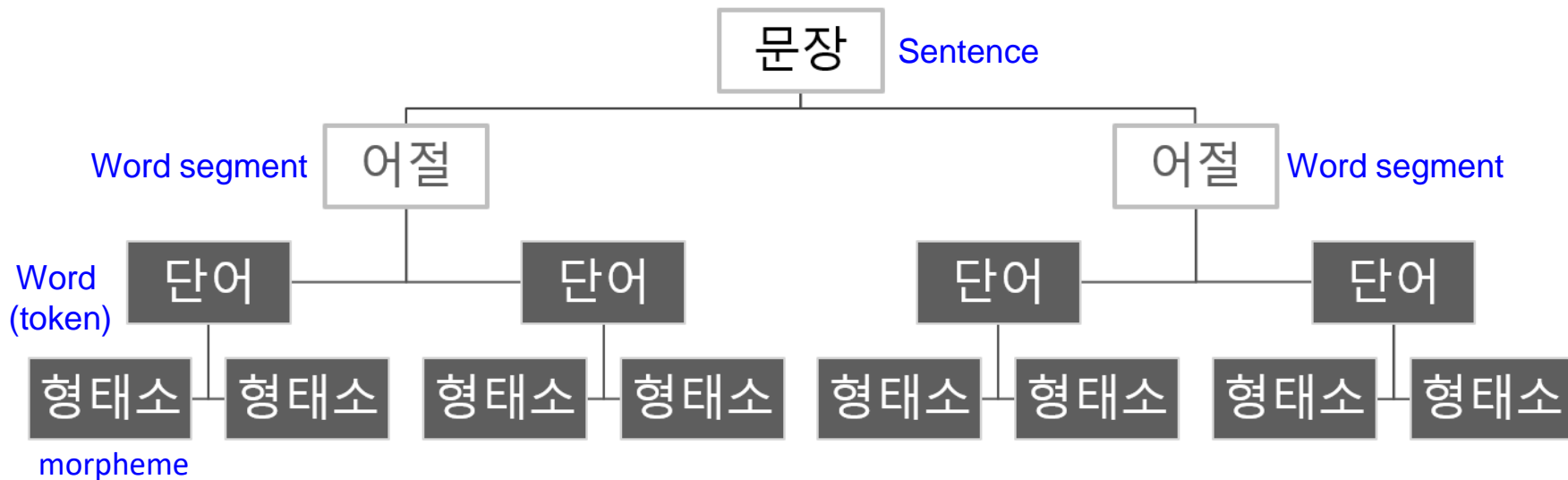
문서와 문서에 있는 단어를 행렬로 정리

용어-문서 행렬(Term-Document Matrix)

parsing, stopwords 처리, stemming 후
아래와 같은 matrix로 정리

	문서1	문서2	문서3
Algebras	1	0	0
And	2	0	1
Categories	1	0	0
Computation	1	1	1
Computational	0	0	1
⋮	⋮	⋮	⋮
of	0	1	0
Parallel	0	0	1
Semantics	1	1	0

텍스트 마이닝 요구 기능



형태소: 더 이상 분석할 수 없는 가장 작은 말의 단위

예: '먹었다', '먹다', '먹을', '먹이다' 에서 '먹'을 제외한 나머지를 **형식형태소**, '먹'을 **실질형태소**라고 함

데이터 전처리(Text Preprocessing)

텍스트를 처리하기 전에 텍스트를 다듬는 작업

- 단어의 토큰화(tokenizing)
- 정규화(Normalization)

문서에서 단어를 추출한 후 같은 의미의 단어를 묶어주는 것을 정규화, 아무런 의미없는 단어를 제거하는 정제(cleaning)도 수행함.

영어를 정규화할때 대소문자 가리지 않고 소문자로 통일.

- 불용어(Stopword)

직접 데이터를 보고 의미없는 단어를 제거. 대표적으로 전치사, 대명사, 접속어가 있음.

- 정규 표현식(regular expression)

정규 표현식으로 특정 규칙에 해당되는 단어

단어의 토큰화

What a wonderful world!

She is a walking dictionary.

→ "what", "a", "wonderful", "world", "!", "she", "is", "a", "walking", "dictionary", "!"

마침표(.), 콤마(,), 세미콜론(:), 물음표(?), 느낌표(!) 등과 같은 구두점(punctuation)과
특수문자는 단어의 의미 없음

KT&G, ph.D., 10.03, She's, walking dictionary → 이런 단어는 어떻게 처리할 것인가?

→ 토큰화를 위한 다양한 알고리즘이 나와있음. 적합한 알고리즘을 선택하여 처리

텍스트 마이닝(Text mining)

텍스트 마이닝

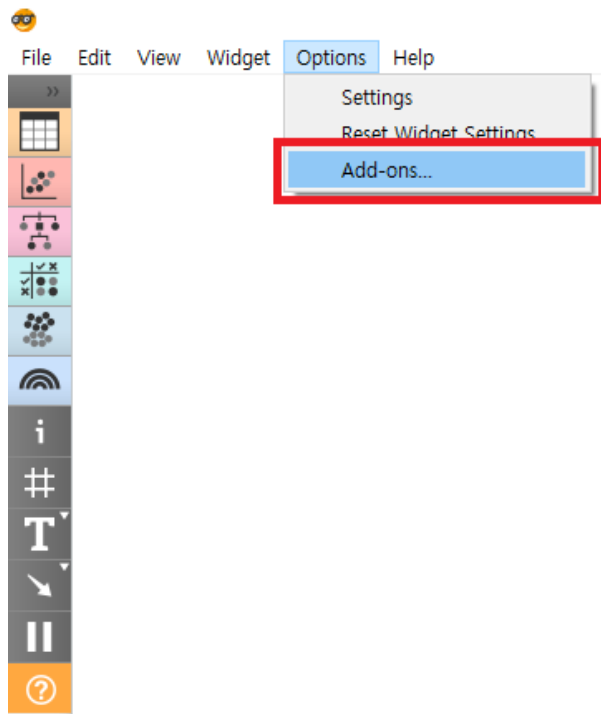
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

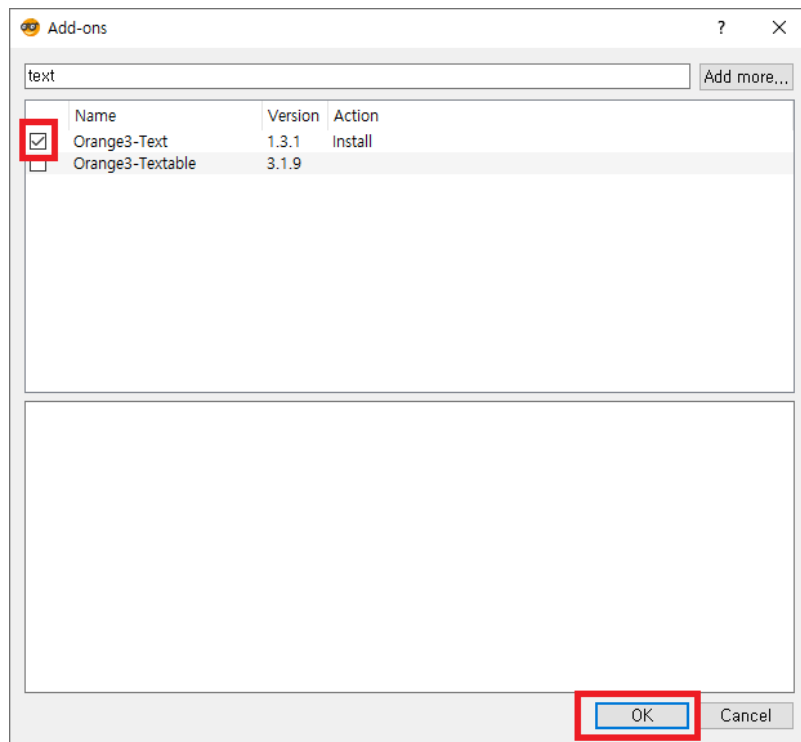
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

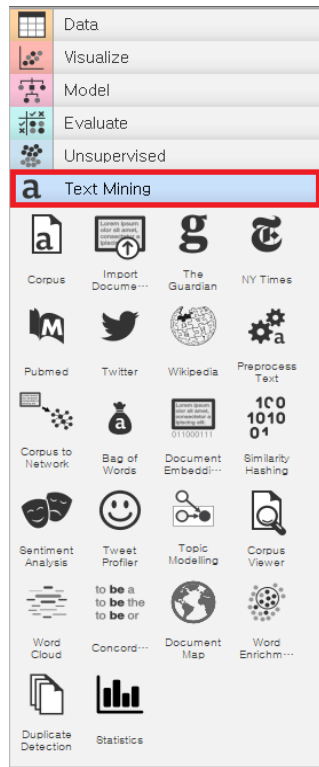
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

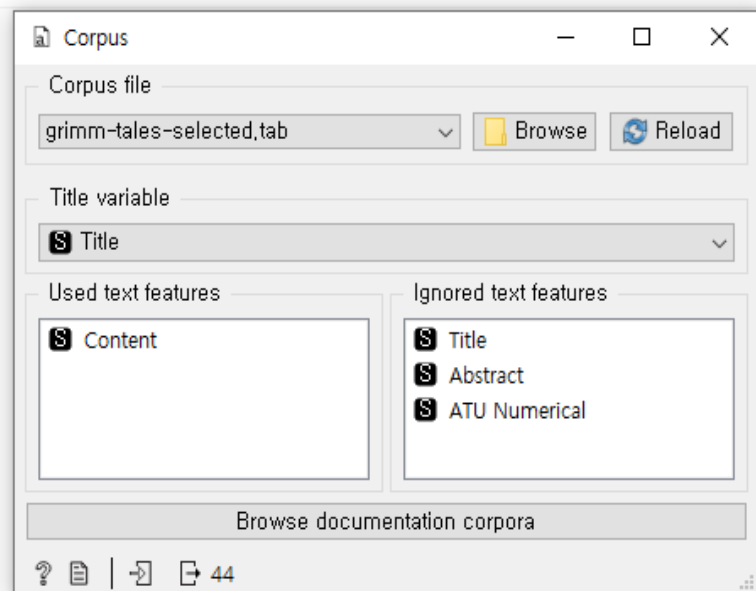
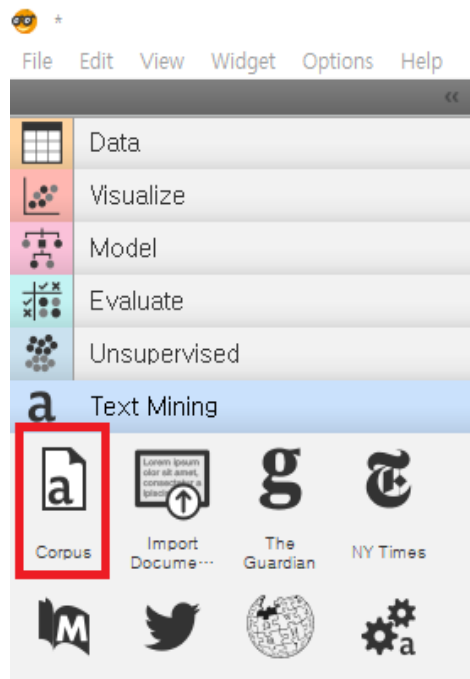
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

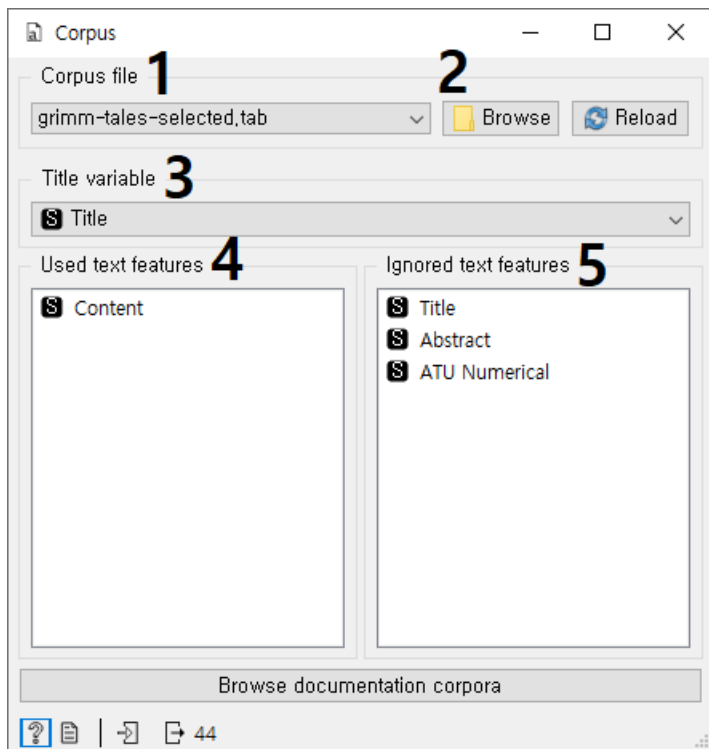
전처리

문서 요약

문서 분류

문서 군집

특징 추출



① Corpus file	데이터 파일을 탐색하거나 샘플 파일을 로드합니다.
② Browse	데이터 파일을 찾습니다.
③ Title variable	Corpus viewer에서 문서 제목으로 표시되는 변수를 선택합니다.
④ Used text features	텍스트 분석에 사용될 feature입니다.
⑤ Ignored text features	텍스트 분석에서 사용되지 않는 feature입니다.

텍스트 마이닝

Aarne-Thompson-Uther (아르네-톰슨-아서)Index (ATU Index)

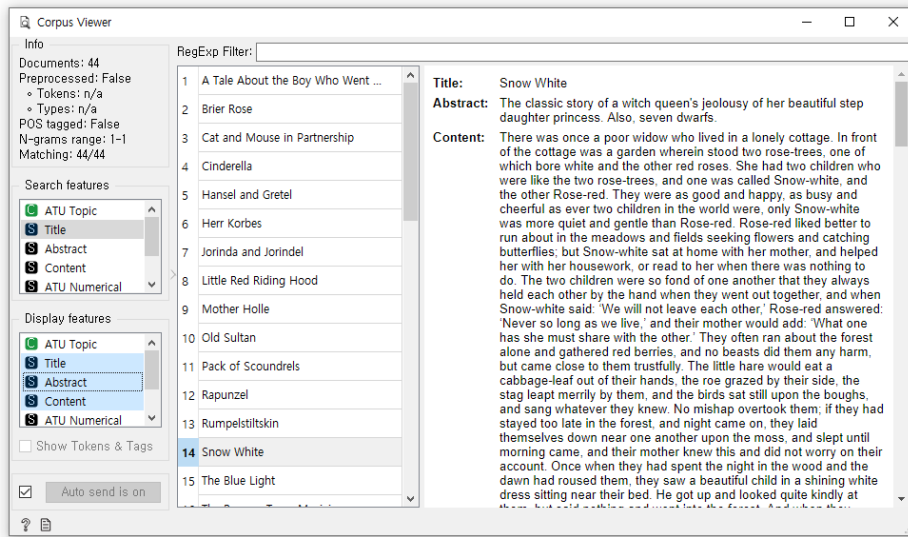
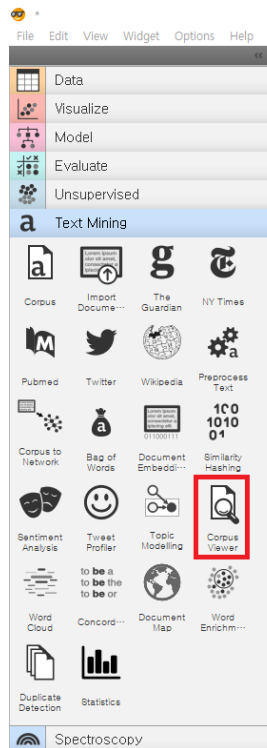
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Word Cloud	말뭉치에서 워드 클라우드를 생성한다.	Corpus	Topics, Corpus, Selected Words, Word Counts

- Word Cloud 위젯을 사용하면 말뭉치의 토큰을 표시하며, 말뭉치 features가 위젯의 입력에 있을 때 말뭉치의 단어 빈도나 평균 단어 수를 나타냄.
- 단어는 **위젯의 빈도에 따라 나열**되며 위젯은 클라우드 단어에서 선택한 토큰을 포함하는 문서를 출력

텍스트 마이닝

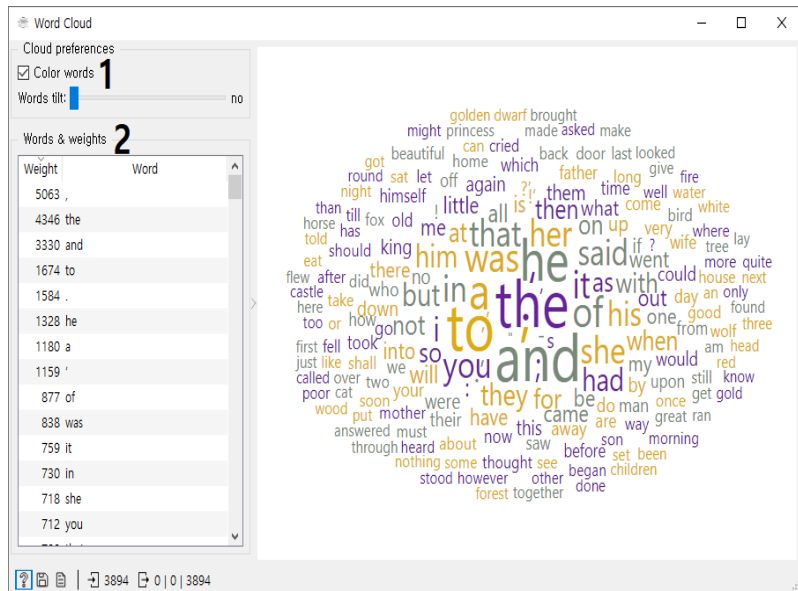
전처리

문서 요약

문서 분류

문서 군집

특징 추출



Plot을 조정할 수 있습니다.

① Cloud Preference

Color words를 활성화 한 경우 단어에는 임의의 색상이 할당되며 비활성화 한 경우 단어는 모두 검정색으로 처리됩니다.

Words tilt에서는 단어 기울기를 조정할 수 있습니다.

② Words & Weights

Words & weights는 말뭉치 또는 항목의 빈도별로 단어를 정렬하여 표시합니다. 단어를 클릭하면 클라우드에서 동일한 단어가 선택되고 일치하는 문서가 출력됩니다. ctrl키를 사용해서 두 개 이상의 단어를 선택하고 선택한 단어 중 임의의 단어와 일치하는 문서가 출력에 나타납니다.

텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Preprocess Text	텍스트 전처리 파이프라인을 구성한다.	Corpus	Corpus

- Preprocess Text 위젯을 사용하면 텍스트를 더 작은 단위로 분할하여 필터링하고 **정규화**를 실행하며 n-gram을 만들고 일부 언어 레이블이 있는 토큰에 태그를 지정
- 분석 단계는 순차적으로 적용되며 순서를 변경할 수 있음.

텍스트 마이닝

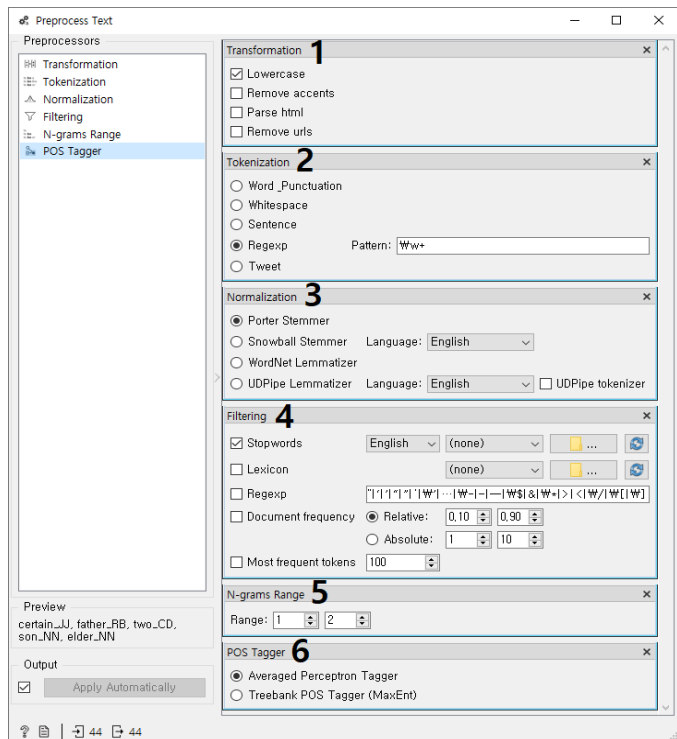
전처리

문서 요약

문서 분류

문서 군집

특징 추출



① Transformation

입력 데이터를 변환합니다. 기본적으로 소문자 변환을 적용합니다.

- **Lowercase**: 모든 텍스트를 소문자로 바꿉니다.
- **Remove accents**: 텍스트의 모든 악센트가 제거됩니다.
- **Parse html**: html태그를 감지하고 텍스트만 구문 분석합니다.
- **Remove urls**: 텍스트에서 url이 제거됩니다.

②Tokenization

텍스트를 더 작은 구성 요소(단어, 문장, 빅그램)로 나눕니다.

- **Word&punctuation**: 텍스트를 단어별로 나누고 구두점 기호를 유지합니다.
- **Whitespace**: 텍스트를 공백으로만 분할합니다.
- **Sentence**: 전체 문장만 유지한 채 텍스트를 완전히 중지하여 분할합니다.
- **Regexp**: 제공된 정규식별로 텍스트를 분할합니다. 기본적으로 단어 단위로만 분할됩니다.
- **Tweet**: 해시태그, 이모티콘 및 기타 특수 기호를 보관하는 사전 훈련된 트위터 모델에 의해 텍스트를 분할합니다.

③Normalization

데이터 정규화를 위한 작업입니다.

- **Poter stemmer**: 기존 포터 스템머를 적용합니다.
- **Snowball stemmer**: 개선된 포터 스템머를 적용하는데 표준화를 위한 언어를 설정해야 합니다.
- **Wordnet lemmatizer**: 영어의 큰 어휘 데이터베이스를 기반으로 하는 토큰에 인지 동의어 네트워크를 적용합니다.
- **UDPipe**: 데이터를 정규화하기 위해 사전 훈련된 모델을 적용합니다.

텍스트 마이닝

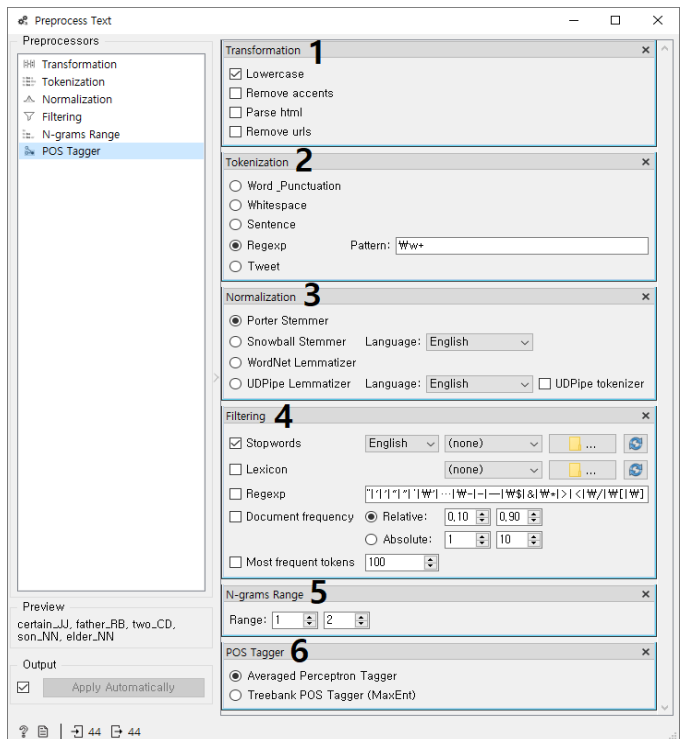
전처리

문서 요약

문서 분류

문서 군집

특징 추출



단어 선택 제거 또는 유지를 합니다.

- **Stopwords:** 텍스트에서 중지 단어(**and**, **or**, **in..**)를 제거합니다.
- **Lexicon:** 파일에 제공된 단어만 보관합니다.
- **Regexp:** 정규식과 일치하는 단어를 제거합니다. 기본값은 구두점을 제거하도록 설정되어 있습니다.
- **Document frequency:** 지정된 문서 수/퍼센트 이상에 나타나는 토큰을 유지합니다. **Absolute**는 지정된 문서 수에 나타나는 토큰만 보관합니다. **Relative**는 지정된 문서 백분율에 나타나는 토큰만 보관합니다.
- **Most frequent tokens:** 지정된 개수의 가장 자주 사용하는 토큰만 유지합니다. 기본값은 가장 자주 사용하는 100개의 토큰입니다.

④Filtering

⑤N-grams
Range

토큰에서 N-gram을 생성합니다. 숫자는 n그램의 범위를 지정하며 기본 값은 1그램과 2그램입니다.

⑥POS Tagger

토큰에 대해 음성 부분 태그를 실행합니다.

- **Averaged Perception Tagger:** Matthew Honnibal의 평균 Perceptron Tagger와 함께 POS태깅을 실행합니다.
- **Treebank POS Tagger (MaxEnt):** 훈련된 Penn Treebank모델로 POS태깅을 실행합니다.

텍스트 마이닝

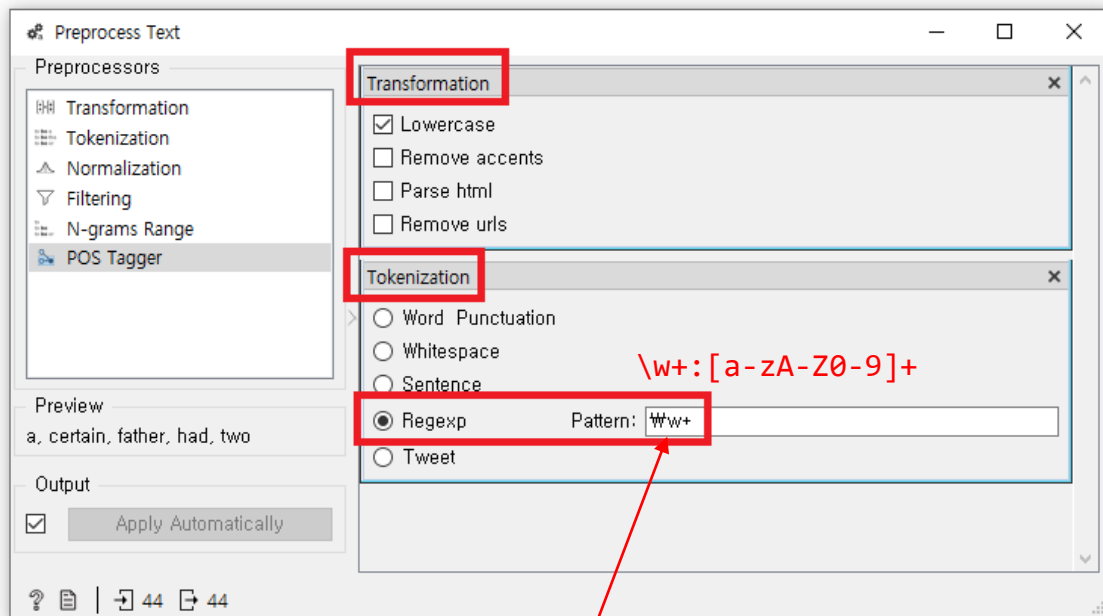
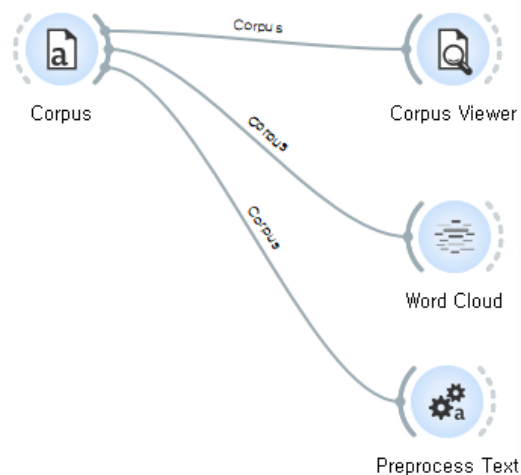
전처리

문서 요약

문서 분류

문서 군집

특징 추출



알파벳이나 숫자 등 한 개 이상이 있는 것들을 남기는 것을 의미
숫자없이 영문자만 남기려고 하면 '[a-z]+' 를 입력

텍스트 마이닝

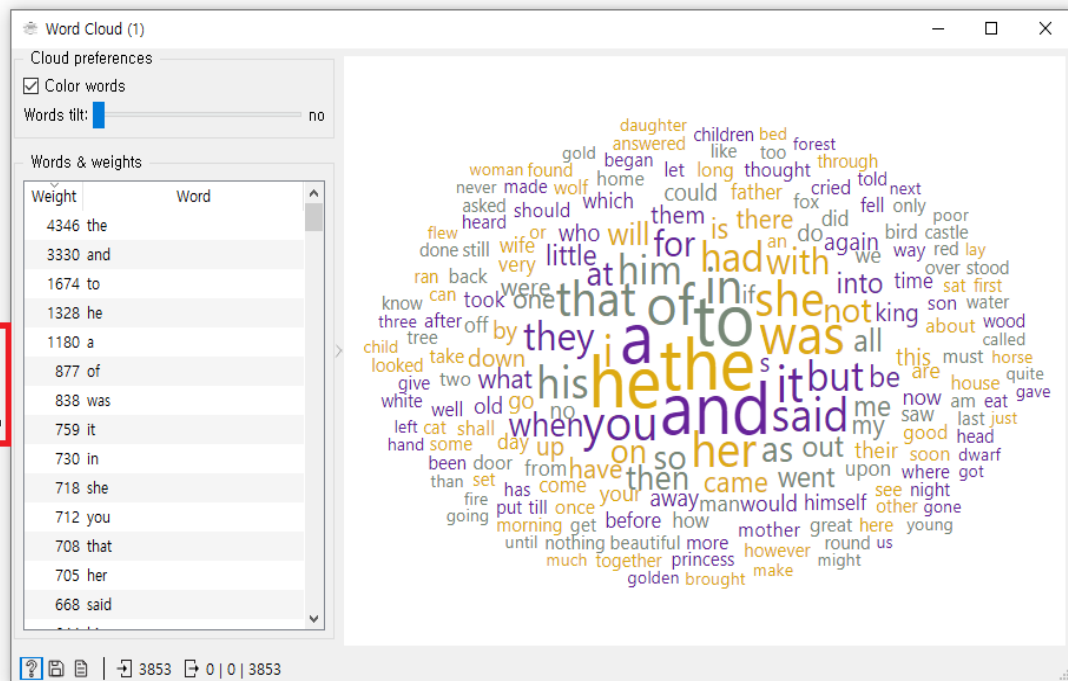
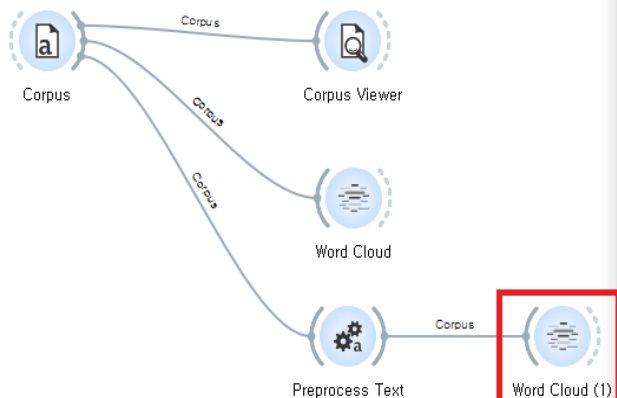
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

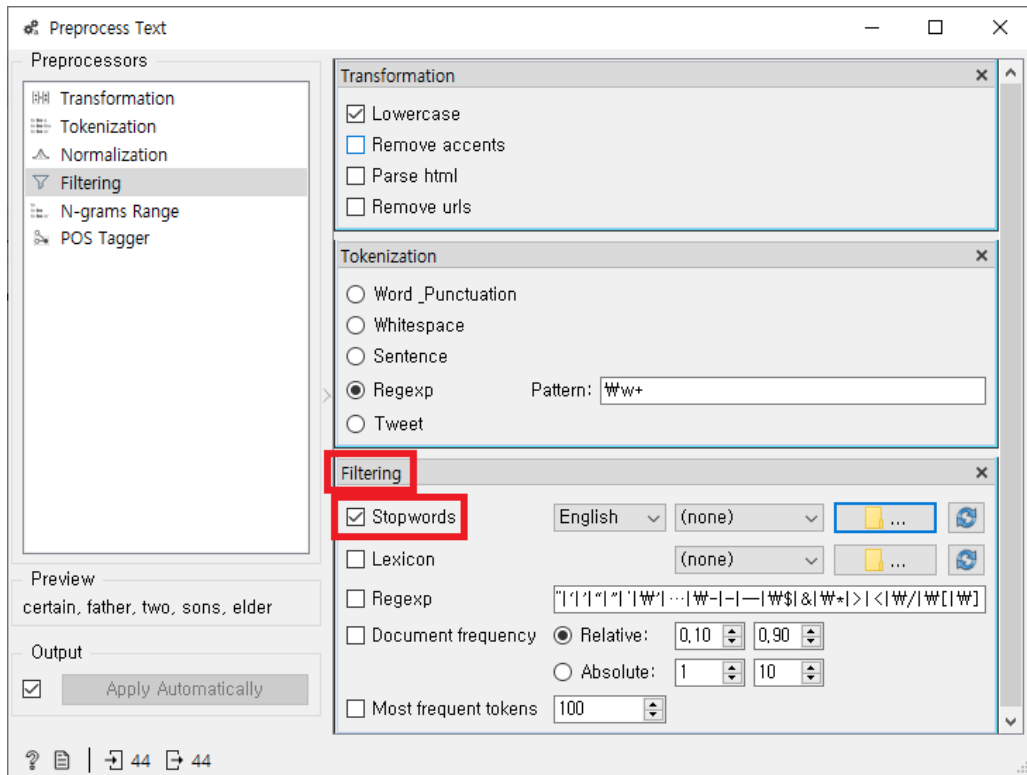
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

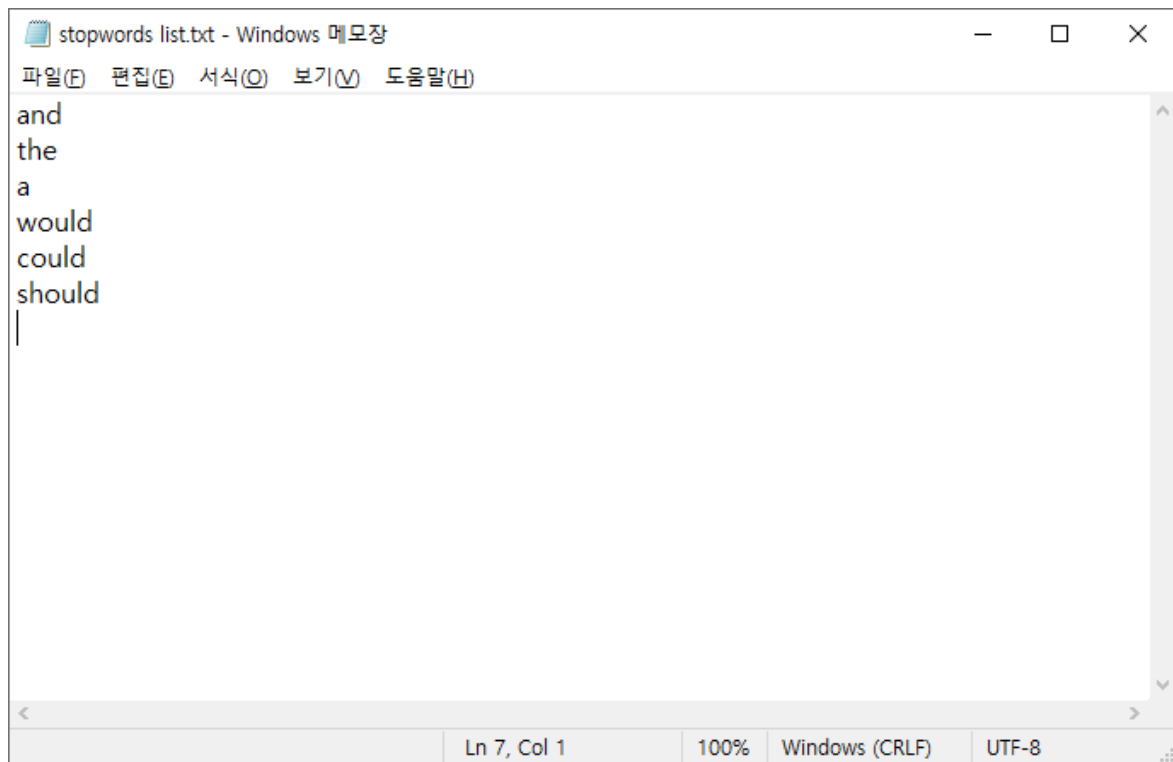
전처리

문서 요약

문서 분류

문서 군집

특징 추출



```
stopwords list.txt - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)
and
the
a
would
could
should
|
Ln 7, Col 1 100% Windows (CRLF) UTF-8
```

텍스트 마이닝

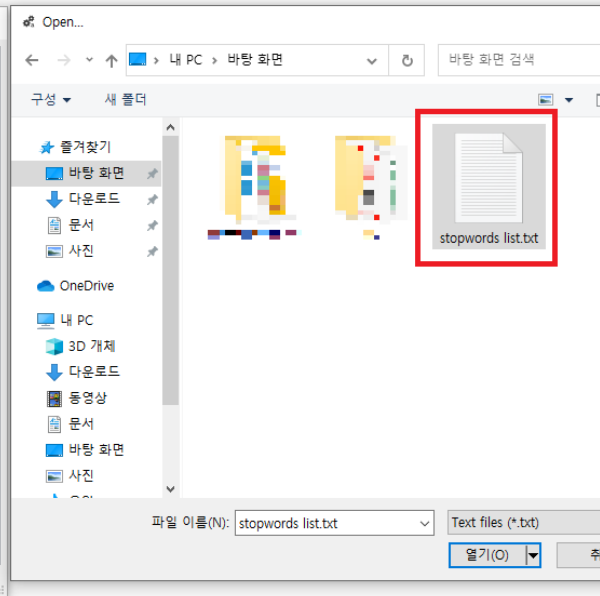
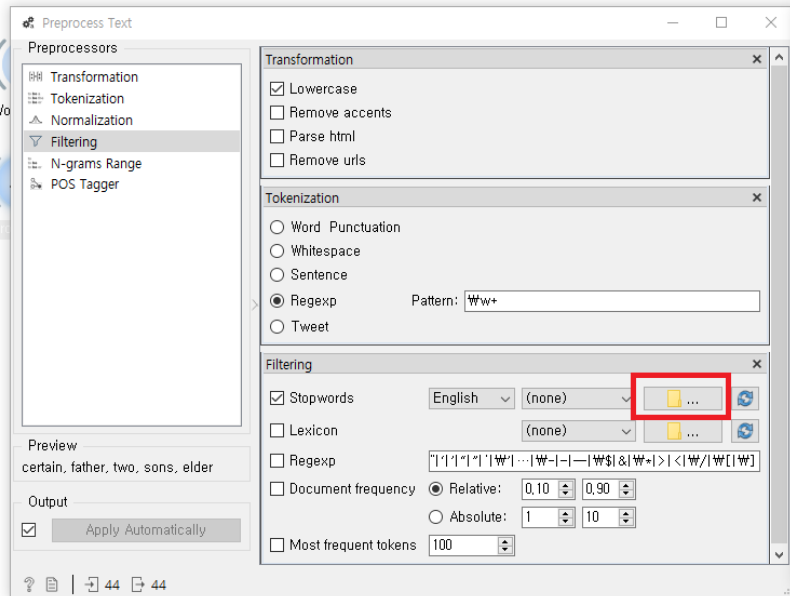
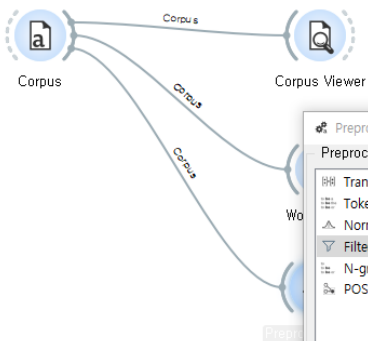
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

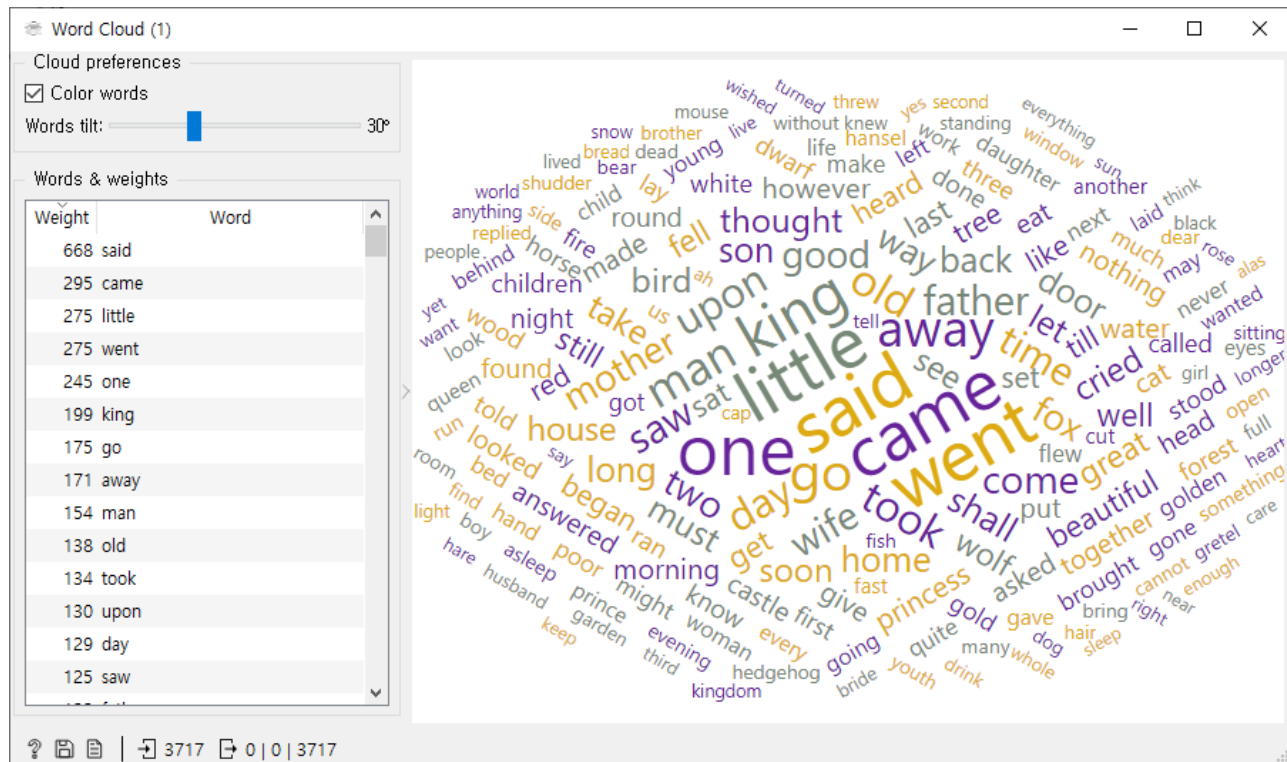
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Topic Modelling	말뭉치의 숨겨진 주제 구조를 밝혀낸다.	Corpus	Corpus, Selected Topic, All Topics

- **Topic Modelling** 위젯을 사용하면 corpus의 단어 군집과 해당 빈도에 기초하여 말뭉치에서 **추상적인 주제**를 발견함.
- 문서는 일반적으로 서로 다른 비율로 여러 개의 주제를 포함하므로 위젯은 문서당 항목 가중치도 보고함.

텍스트 마이닝

문서를 요약하는 토픽 모델링 알고리즘

- 잠재적 의미 분석 (가장 많이 사용) □

- 잠재적 디리클레 할당

- 계층 디리클레 프로세스

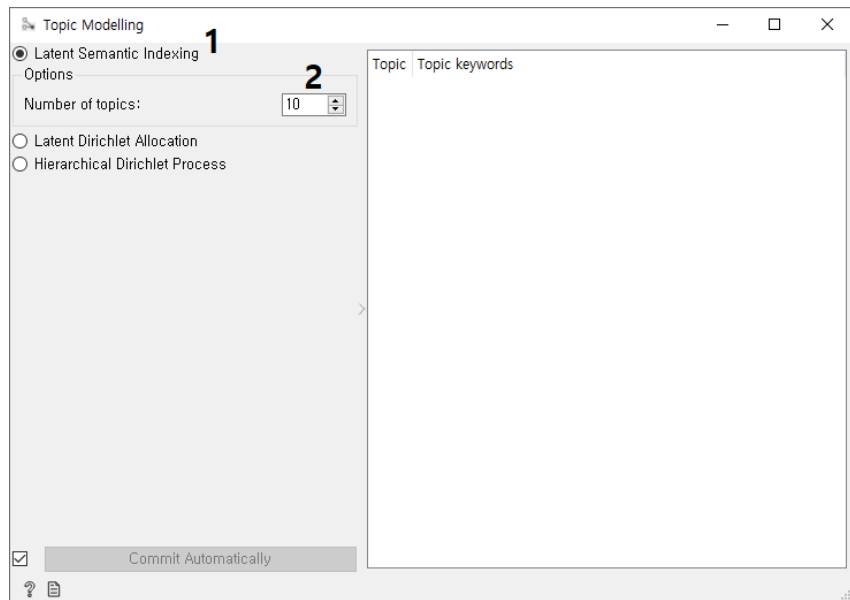
전처리

문서 요약

문서 분류

문서 군집

특징 추출



주제 모델링 알고리즘

①

- Latent Semantic Indexing(음수 및 양수 단어와 주제 가중치를 모두 반환한다)
- Latent Dirichlet Allocation
- Hierarchical Dirichlet Process

알고리즘에 대한 매개변수

②

LSI 및 LDA는 모델링된 항목 수만 허용하며 기본값은 10으로 설정됩니다. 그러나 HDP에는 더 많은 매개변수가 있습니다. 이 알고리즘은 계산적으로 매우 까다롭기 때문에 하위 집합에서 시도하거나 필요한 모든 매개 변수를 미리 설정한 다음 알고리즘을 실행하는 것이 좋습니다.

텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

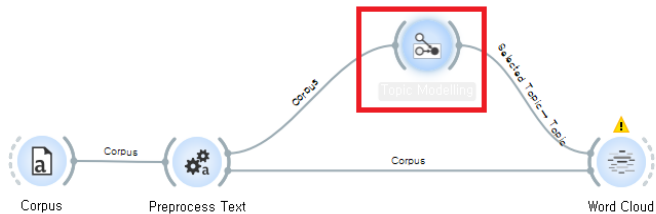
전처리

문서 요약

문서 분류

문서 군집

특징 추출



긍정적인 단어는 **녹색**
부정적인 단어는 **빨간색**

Topic Modelling

☒ Latent Semantic Indexing
Options
Number of topics: 10

☐ Latent Dirichlet Allocation
☐ Hierarchical Dirichlet Process

Commit Automatically

Topic	Topic keywords
1	said, came, went, little, one, king, go, away, man, father
2	little, hansel, king, gretel, children, mother, bird, red, said, forest
3	bird, hansel, gretel, mother, tree, son, beautiful, children, said, for
4	wife, said, fish, came, fisherman, princess, fox, king, horse, old
5	shudder, youth, wife, fire, fish, learn, fisherman, father, home, bo
6	red, hansel, gretel, rose, white, bear, snow, wolf, bird, cap
7	fox, wolf, pick, bird, ashputtel, bride, tree, cap, golden, mother
8	talada, maid, fox, curdken, bride, blow, head, alas, sparrow, sad
9	cap, grandmother, wolf, bear, white, snow, rose, little, children, h
10	hedgehog, hare, field, cap, grandmother, run, wife, king, water, lit

텍스트 마이닝

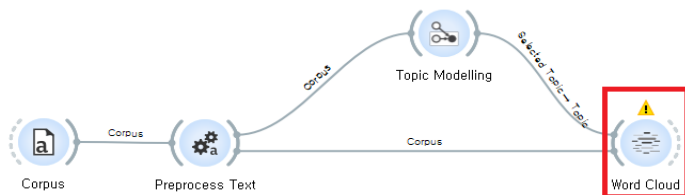
전처리

문서 요약

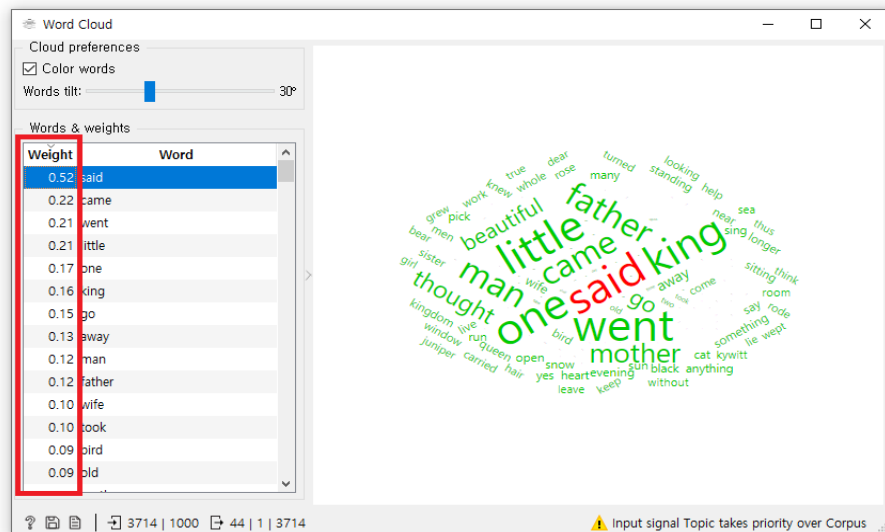
문서 분류

문서 군집

특징 추출



- **양의 가중치**는 해당 단어가 주제를 매우 잘 나타낸다는 것을 뜻하며,
- **음의 가중치**는 해당 단어가 주제를 잘 나타내지 않는다는 것을 의미



텍스트 마이닝

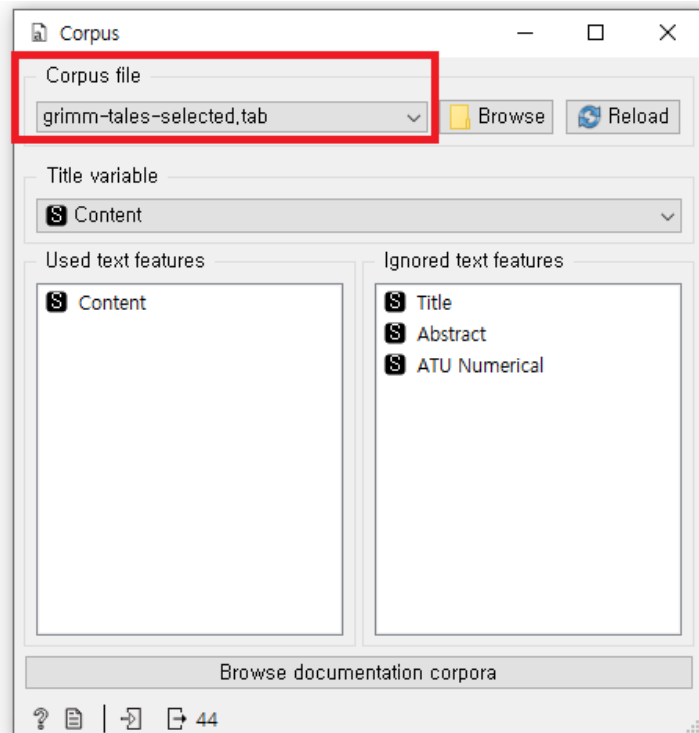
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출

The screenshot displays the 'Corpus Viewer' application window. On the left, a diagram shows a 'Corpus' icon connected to a 'Corpus Viewer' icon. The main window is divided into several sections:

- Info:** Documents: 44, Preprocessed: False, Tokens: n/a, Types: n/a, POS tagged: False, N-grams range: 1-1, Matching: 44/44.
- Search features:** A list of features including 'ATU Topic' (highlighted with a red box), 'Title', 'Abstract', 'Content', 'ATU Numerical', and 'ATU Type'.
- Display features:** A list of features including 'ATU Topic' (highlighted with a red box), 'Title', 'Abstract', 'Content', 'ATU Numerical', and 'ATU Type'.
- RegExn Filter:** A text input field.
- Document List:** A table of documents with their titles and assigned ATU Topics. The table is highlighted with a red box.

Document ID	Document Title	ATU Topic
23	A farmer had a horse th...	Tales of Magic
24	One fine evening a youn...	Tales of Magic
25	A certain king had a ...	Animal Tales
26	There was a man who h...	Tales of Magic
27	The king of a great land ...	Tales of Magic
28	This story was actually ...	Animal Tales
29	Long, long ago, some t...	Tales of Magic
30	There was once a ...	Tales of Magic
31	Once upon a time, a ...	Tales of Magic
32	There was once upon a ...	Tales of Magic
33	Two kings' sons once ...	Animal Tales
34	There was once a queen ...	Animal Tales
35	There was once a man ...	Tales of Magic
36	In a village dwelt a poor ...	Tales of Magic
37	An aged count once live...	Tales of Magic
38	Long before you or I wer...	Tales of Magic
39	FIRST STORY There was ...	Tales of Magic
40	A long time ago there ...	Animal Tales
41	Once in summer-time th...	Animal Tales
42	The wolf had the four wit...	

텍스트 마이닝

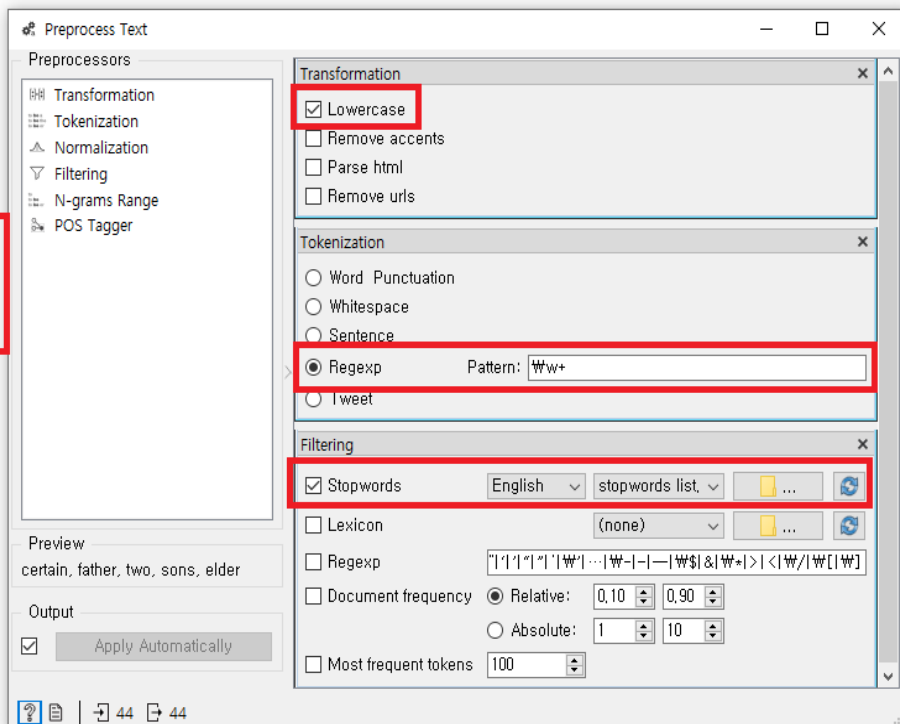
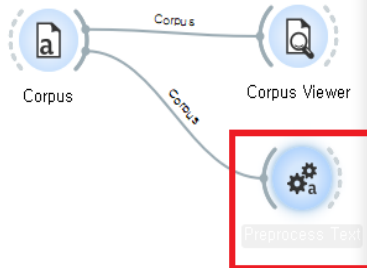
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

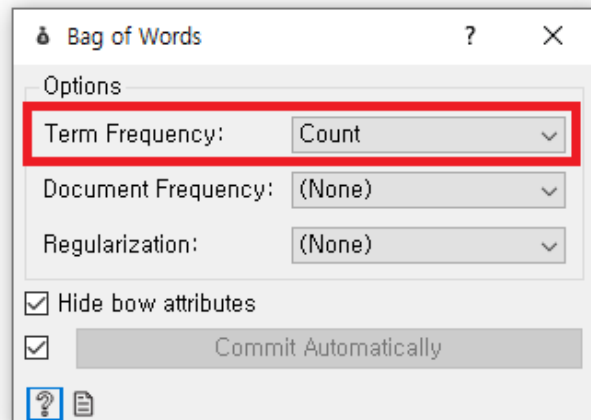
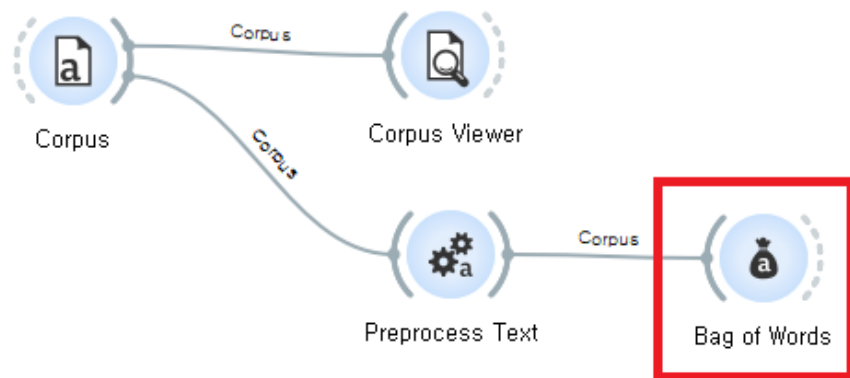
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Logistic Regression	L1(LASSO) 또는 L2(리지) 정규화를 사용한 로지스틱 회귀 분류 알고리즘이다.	Data	Preprocessor Learner, Model, Coefficients

- Logistic Regression 위젯은 데이터에서 로지스틱 회귀 모형을 학습
- 분류 작업에만 사용할 수 있음.

텍스트 마이닝

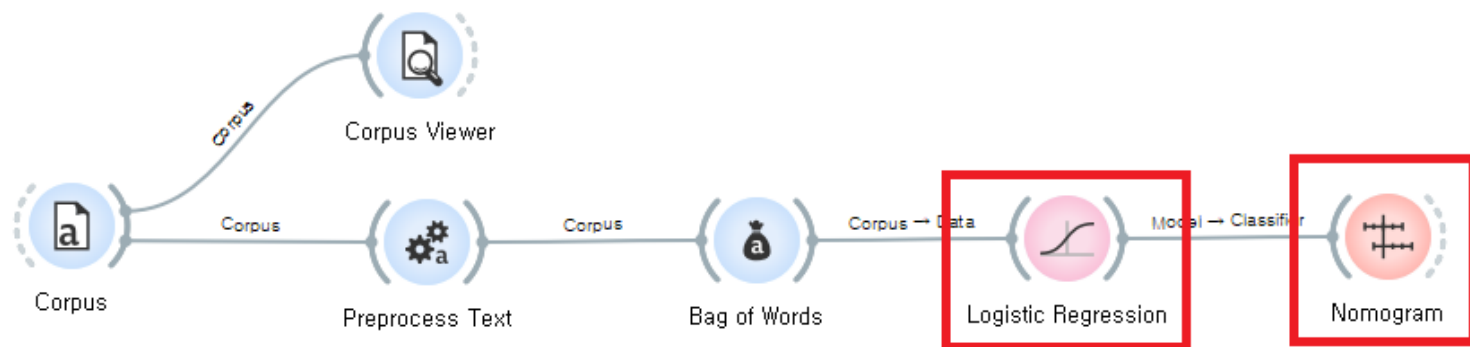
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Nomogram	나이브 베이지안 및 로지스틱 회귀 분류기의 시각화를 위한 노모그램이다.	Classifier	Data features

- Nomogram 위젯은 **일부 분류자 (Naive Bayes, Logistic Regression)의 시각적 표현을 가능**
- 훈련 데이터의 구조와 속성이 클래스 확률에 미치는 영향에 대한 통찰력을 제공
- 분류자의 시각화 외에도 위젯은 클래스 확률을 예측하기 위한 대화형 지원을 제공

텍스트 마이닝

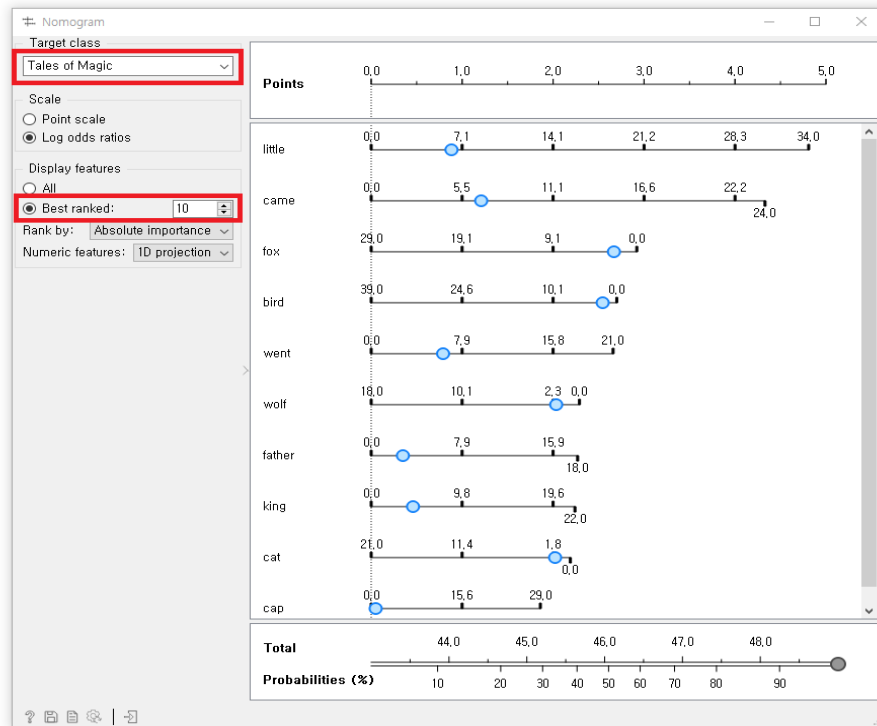
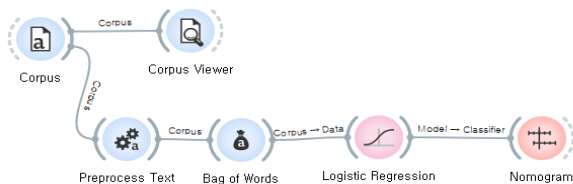
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

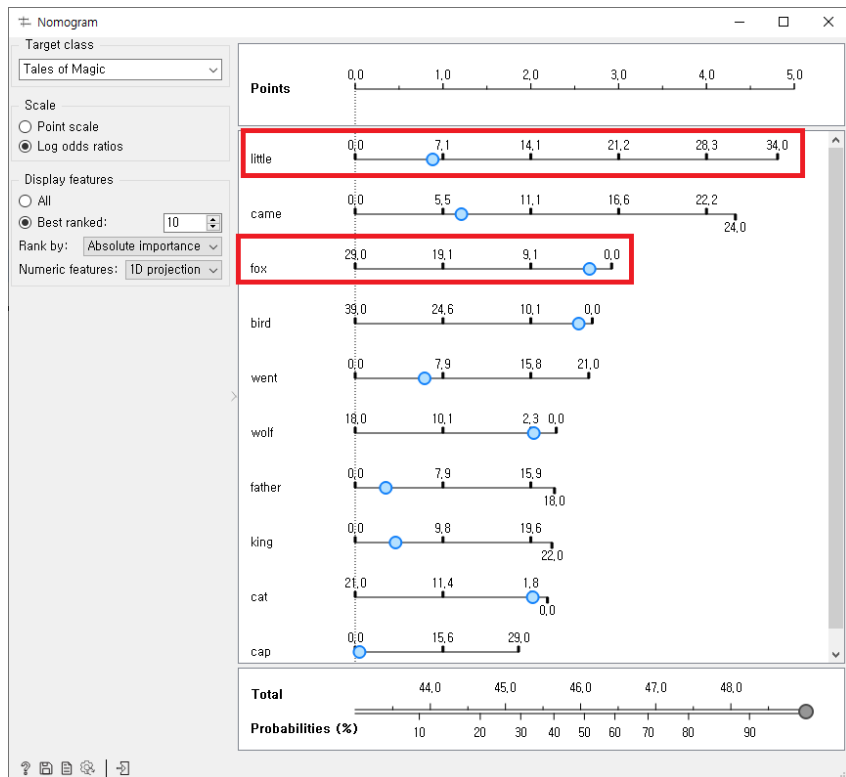
전처리

문서 요약

문서 분류

문서 군집

특징 추출



- 0부터 시작해서 증가하는 단어
→ target을 만드는데 기여하는 단어
- 0이 아닌 숫자부터 감소하는 단어
→ target을 만드는데 방해하는 단어

텍스트 마이닝

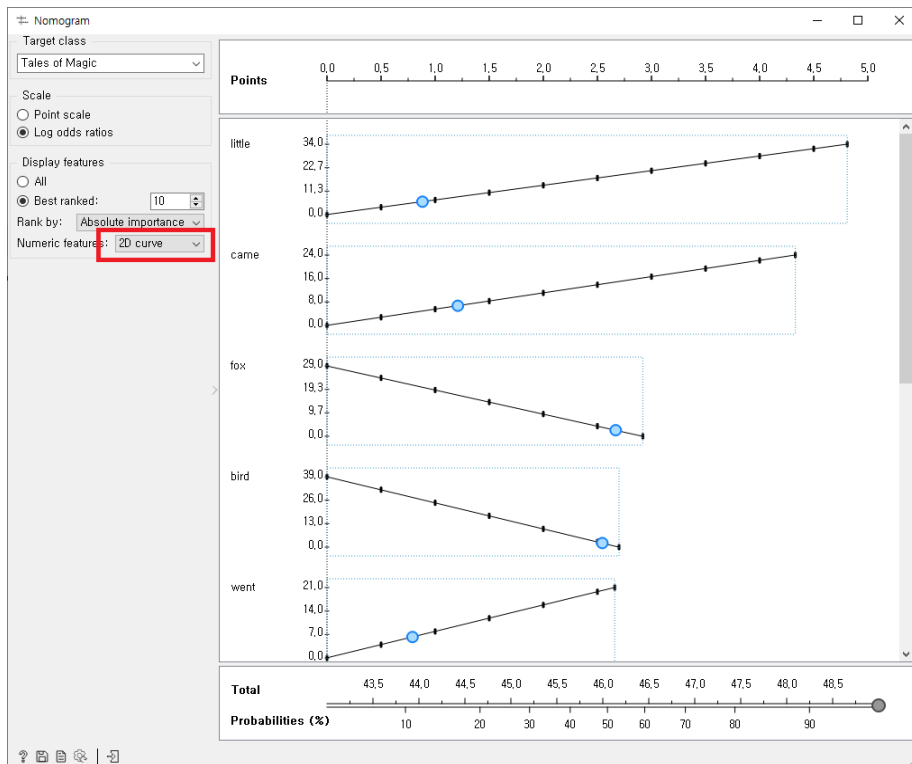
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Naive Bayes	Feature 사이의 독립성을 가정한 Bayes의 정리를 기반으로 한 빠르고 단순한 확률 분류기입니다.	Data	Preprocessor, Learner, Model

- Naïve Bayes 위젯은 데일로부터 Naïve Bayesian 모델을 학습
- **분류 작업에만 사용**할 수 있음.

텍스트 마이닝

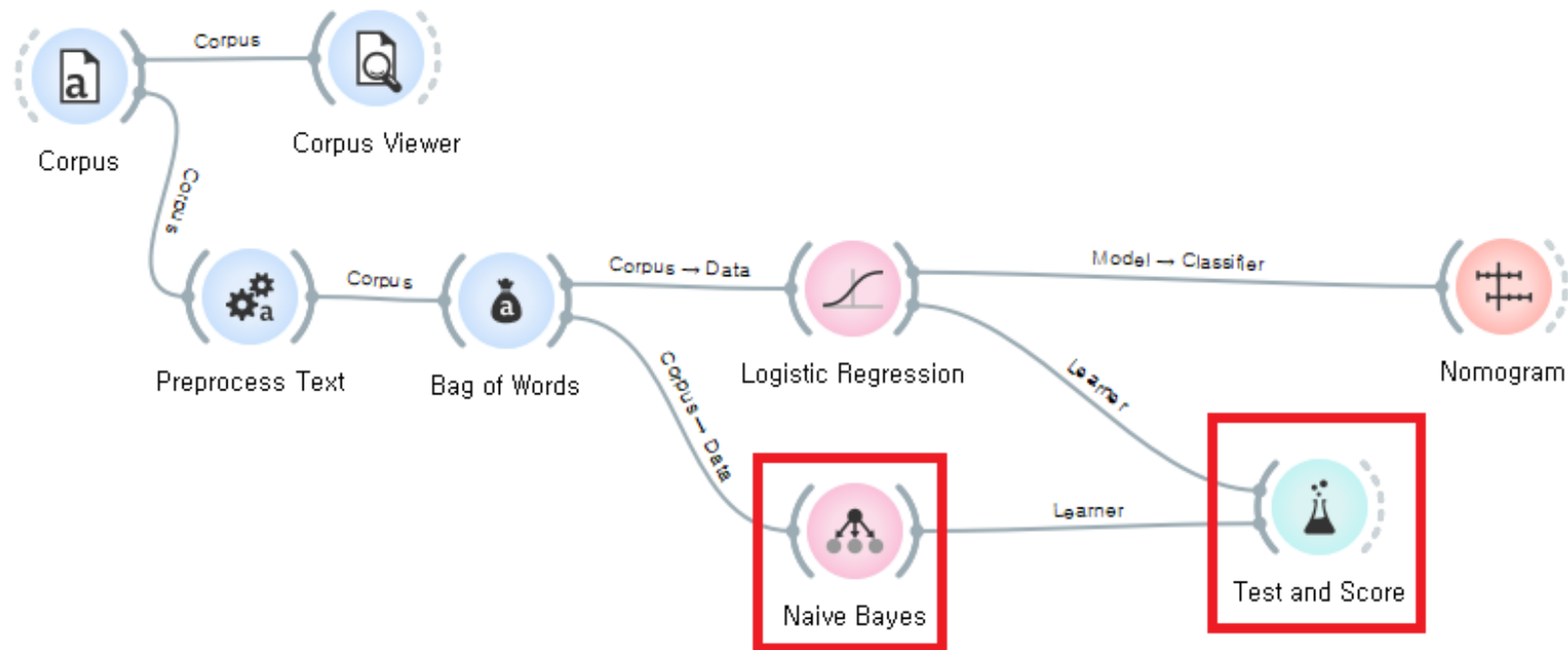
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

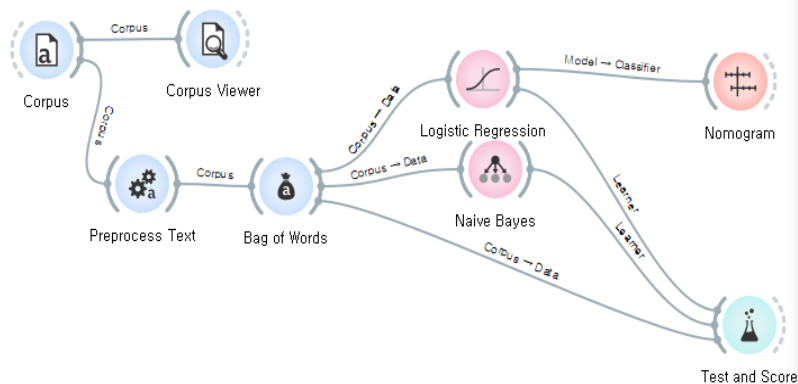
전처리

문서 요약

문서 분류

문서 군집

특징 추출



Test and Score

Sampling

- ☒ Cross validation
 - Number of folds: 10
 - ☒ Stratified
- ☐ Cross validation by feature
 - ATU Type
- ☐ Random sampling
 - Repeat train/test: 10
 - Training set size: 66 %
 - ☒ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

Target Class

(Average over classes)

Model Comparison

Area under ROC curve

☐ Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.968	0.636	0.611	0.807	0.636
Logistic Regression	0.968	0.909	0.910	0.926	0.909

Model Comparison by AUC

	Naive Bayes	Logistic Re...
Naive Bayes		0.500
Logistic Regression	0.500	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

44

텍스트 마이닝

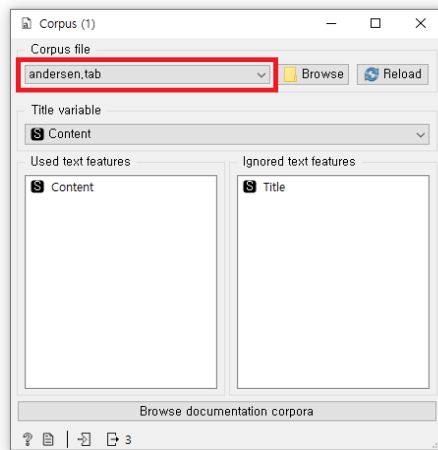
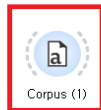
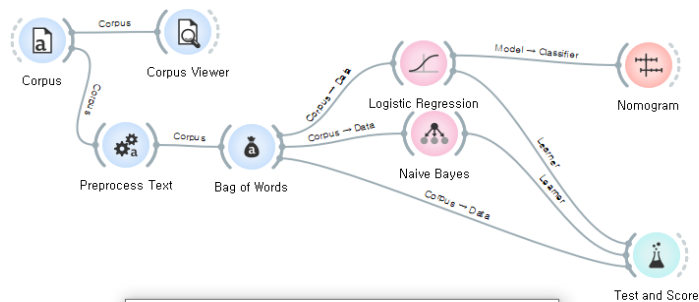
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

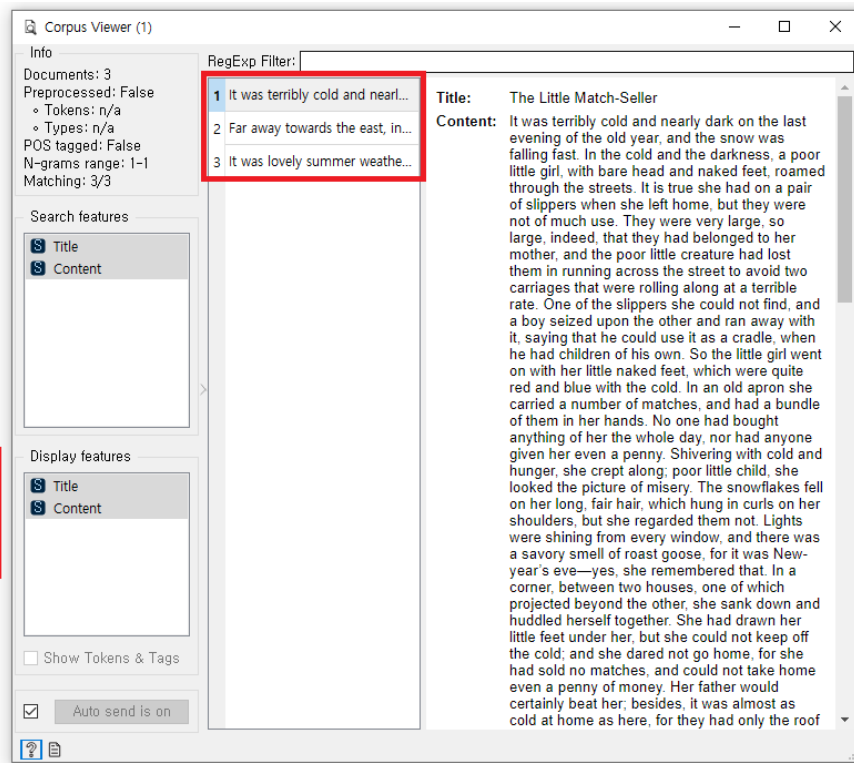
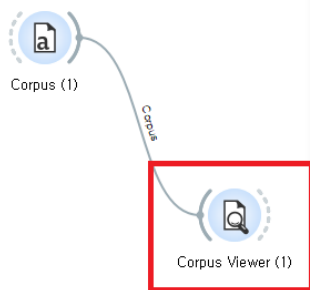
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

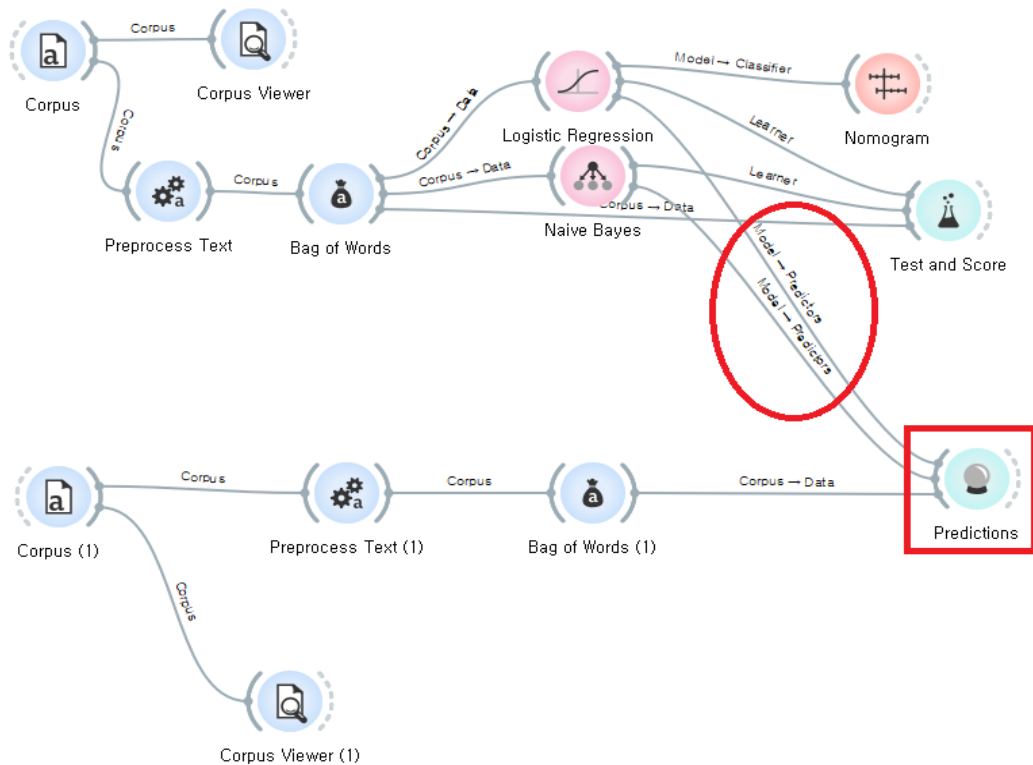
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출

Predictions

Show probabilities for

Animal Tales
Tales of Magic

	Naive Bayes	Logistic Regression
1 Animal Tales	Tales of Magic	
2 Tales of Magic	Tales of Magic	
3 Tales of Magic	Tales of Magic	

Title	Content	{...}
The Little Matc...	It was terribly ...	across=2, ah=1...
The ...	Far away towar...	abilities=1, ...
The Ugly ...	It was lovely ...	able=1, ...

Restore Original Order

? | 3 | 3

텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출

Predictions

Show probabilities for

Animal Tales
Tales of Magic

	Naive Bayes	Logistic Regression	Title	Content	{...}
1	0.00 → Animal Tales	0.77 → Tales of Magic	The Little Matc...	It was terribly ...	across=2, ah=1.
2	1.00 → Tales of Magic	1.00 → Tales of Magic	The ...	Far away towar...	abilities=1, ...
3	1.00 → Tales of Magic	1.00 → Tales of Magic	The Ugly ...	It was lovely ...	able=1, ...

Restore Original Order

텍스트 마이닝



전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Bag of Words	입력 말뭉치에서 단어 주머니를 생성한다. 	Corpus	Corpus

- Bag of Words 위젯을 사용하면 각 데이터 인스턴스(문서)에 대한 단어 수가 있는 말 뭉치를 만들 수 있음.
- 카운트는 절대값, 이진수(포함 또는 포함하지 않음), 준선형(빈도 용어의 대수)로 나타남.
- Bag of Words는 예측 모델링에 사용될 수 있음.

텍스트 마이닝

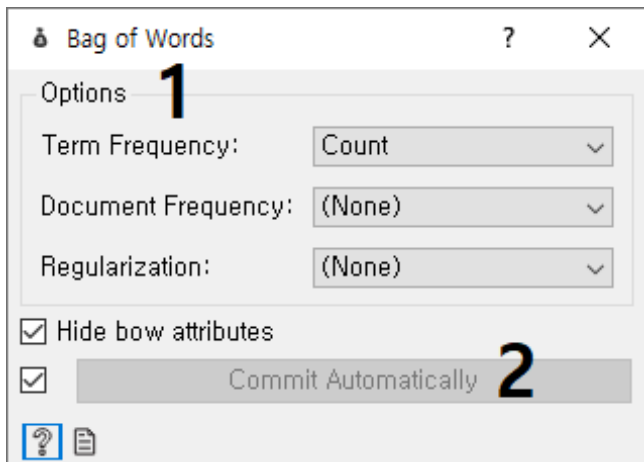
전처리

문서 요약

문서 분류

문서 군집

특징 추출



위젯 매개 변수

①Options

- Term Frequency: Count(문서에서 단어 발생 횟수), Binary(단어가 나타나거나 문서에 나타나지 않음), Sublinear(항 빈도(카운트)의 로그)
- Document Frequency: None(없음), IDF(역 문서 빈도), Smooth IDF(문서 빈도에 하나를 추가하여 영분할을 방지)
- Regularization: None(없음), L1(요소의 합: 벡터 길이를 요소의 합으로 정규화), L2(유클리드: 벡터 길이를 제곱합으로 정규화)

②Commit Automatically

Commit Automatically가 켜져 있으면 변경 내용이 자동으로 전달됩니다.

텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출

Document	the	Man	sat	in	hat	with
The man sat	1	1	1	0	0	0
The man sat in the hat	2	1	1	1	1	0
The man with the hat	2	1	0	0	1	1

- 각각의 단어는 **토큰**이라고 하며 토큰이라고 나누는 것을 **tokenizing**한다고 함.
- 각각의 빈도를 수치상으로 나타내는 작업을 벡터화 한다고 함.
- 이러한 작업이 bag of words 위젯의 기능

텍스트 마이닝

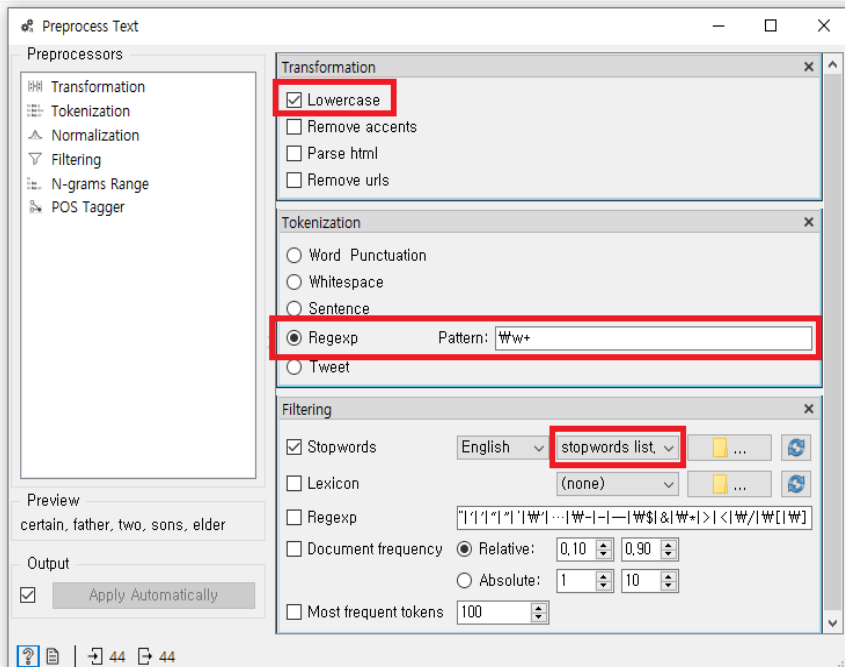
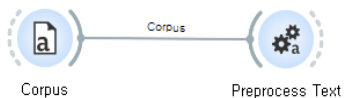
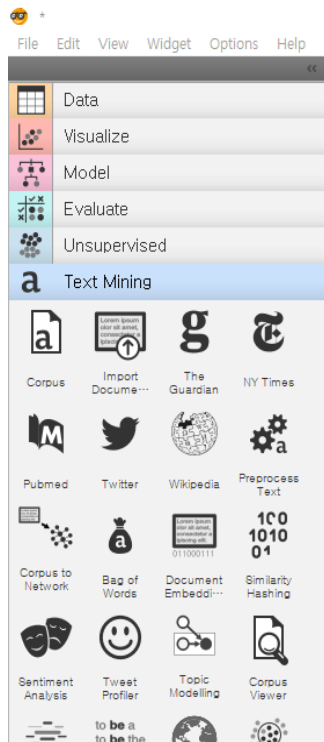
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출



File Edit View Widget Options Help

Data

Visualize

Model

Evaluate

Unsupervised

Text Mining

Corpus

Import Document...

The Guardian

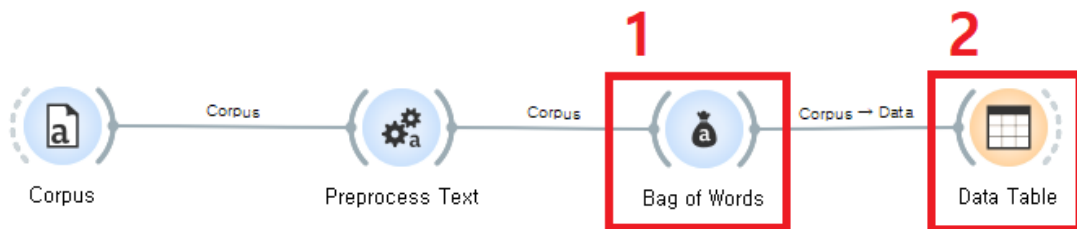
NY Times

M

Twitter

Globe

Gears



텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출

Data Table

Info

44 instances

3177 features (sparse, density 8,25 %)

Target with 2 values

5 meta attributes

Variables

Show variable labels (if present)

Visualize numeric values

Color by instance classes

Selection

Select full rows

Restore Original Order

Send Automatically

44

	bow-feature hidden include skip-normalization title	ATU Topic	Title	Abstract	Content	ATU Numerical	ATU Type	{...}
		True			True			
1	Tales of Magic	A Tale About t...	A simple boy ...	A certain father...	326.0	Supernatural ...	able=1, ...	
2	Tales of Magic	Brier Rose	An offended ...	A king and ...	410.0	Supernatural or...	ale=3, alone=1...	
3	Animal Tales	Cat and Mouse...	A mouse lives ...	A certain cat h...	15.0	Wild Animals	absence=1, ...	
4	Tales of Magic	Cinderella	The familiar sto...	The wife of a ...	510A	Supernatural ...	_my_=1, ...	
5	Tales of Magic	Hansel and ...	A poor ...	Hard by a grea...	327A	Supernatural ...	able=1, ...	
6	Animal Tales	Herr Korbes	A hen and a ...	Once upon a ...	210.0	Domestic ...	aboard=2, ...	
7	Tales of Magic	Jorinda and ...	A witch lures ...	There was once...	405.0	Supernatural or...	_jug=1, alas=2...	
8	Tales of Magic	Little Red Ridin...	A girl known f...	Once upon a ...	333.0	Supernatural ...	able=1, act=1, ...	
9	Tales of Magic	Mother Holle	A widow spoils...	Once upon a ...	480.0	Supernatural ...	according=1, ...	
10	Animal Tales	Old Sultan	A farmer decid...	A shepherd ha...	101.0	Wild Animal an...	accordingly=1, ...	
11	Animal Tales	Pack of ...	A rooster and a...	The rooster sai...	210.0	Domestic ...	able=2, ...	
12	Tales of Magic	Rapunzel	The classic stor...	There were onc...	310.0	Supernatural ...	afraid=1, ...	
13	Tales of Magic	Rumpelstiltskin	A miller's ...	By the side of ...	500.0	Supernatural ...	alas=1, alone=...	
14	Tales of Magic	Snow White	The classic stor...	There was once...	426.0	Supernatural or...	account=1, ...	
15	Tales of Magic	The Blue Light	A wounded ...	There was once...	562.0	Supernatural ...	advice=1, aha=...	
16	Animal Tales	The Bremen ...	A donkey, a do...	An honest ...	130.0	Wild Animal an...	abode=1, ...	
17	Animal Tales	The Crumbs on...	A man tells his ...	One day the ...	236.0	Other Animals ...	anything=2, ...	
18	Animal Tales	The Dog and t...	A merchant ru...	A shepherd's ...	248.0	Other Animals ...	aim=1, aimed=...	
19	Tales of Magic	The Elves and ...	A poor ...	There was once...	503.0	Supernatural ...	always=1, ...	
20	Tales of Magic	The Fisherman ...	A fisherman ...	There was once...	555.0	Supernatural ...	ah=9, alas=3, ...	
21	Animal Tales	The Fox and th...	The fox is ...	It happened th...	105.0	Wild Animal an...	able=1, ah=1, ...	

텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
	쌍별 거리 행렬을 계산한다.	Data	Distances

- Distances 위젯은 데이터 세트의 행 또는 열 사이의 거리를 계산
- 기본적으로 데이터는 개별 features를 동일하게 취급하도록 정규화
- 데이터 인스턴스 사이의 거리를 계산하고 그 결과를 계층적 클러스터링으로 전달하여 데이터 인스턴스 그룹을 찾을 수 있음

텍스트 마이닝

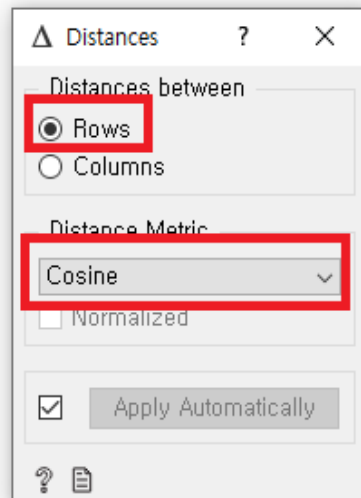
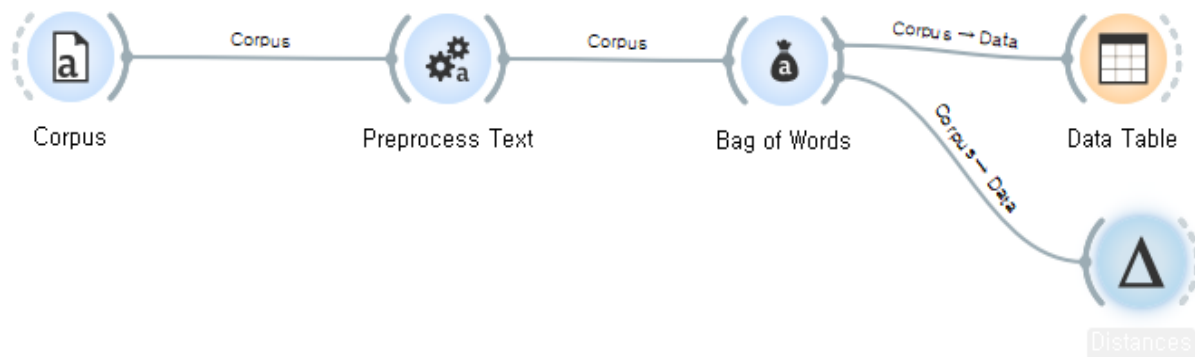
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Hierarchical Clustering	입력된 거리 행렬에서 생성된 계층적 군집화의 덴드로그램을 표시한다.	Distances	Selected Data, Data

- Hierarchical Clustering 위젯은 거리 행렬에서 임의의 개체 유형의 계층적 클러스터링을 계산하고 해당 덴드로그램을 표시

텍스트 마이닝

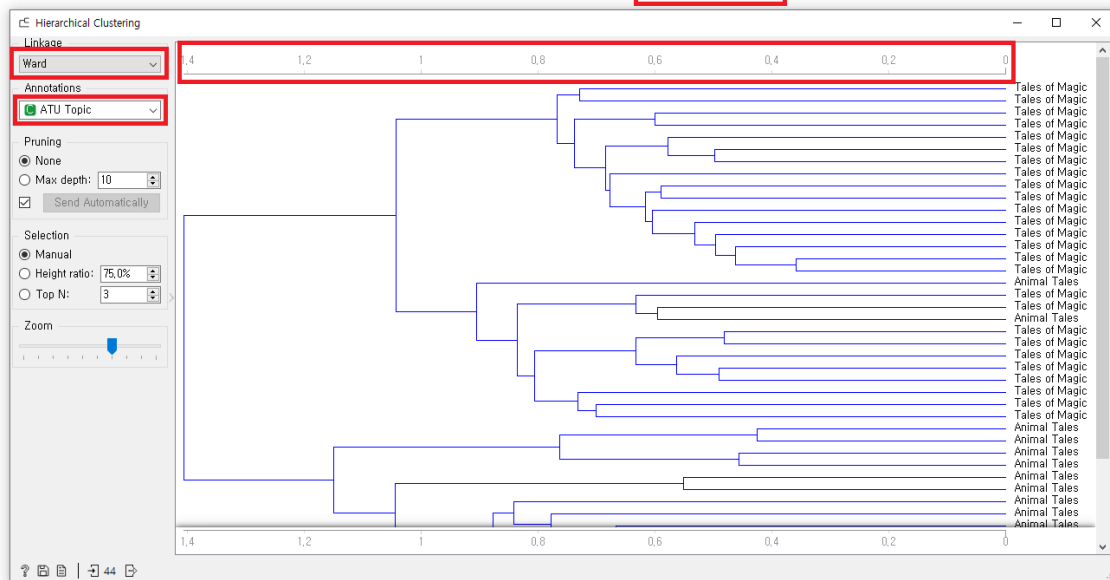
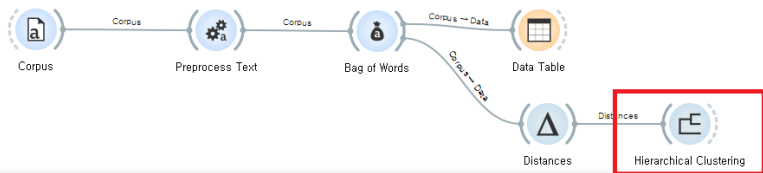
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

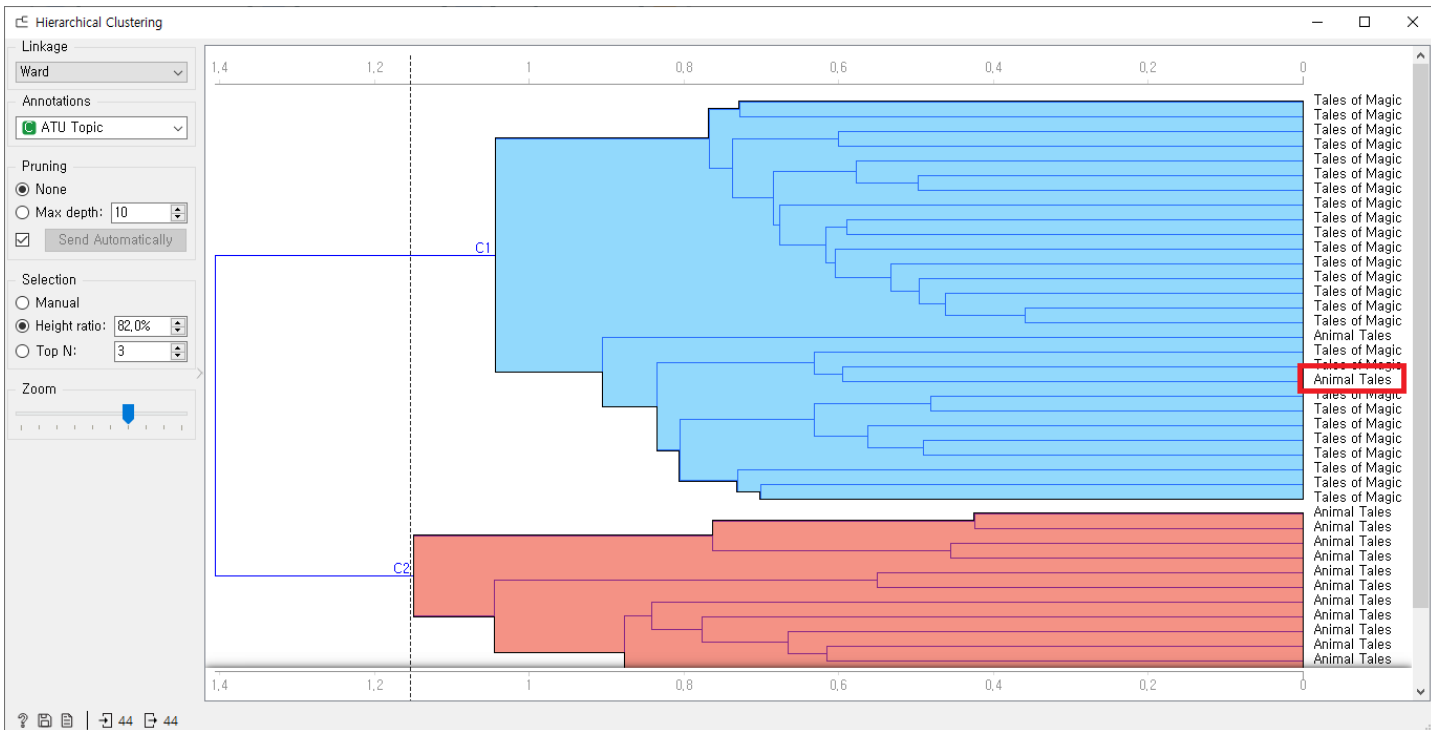
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

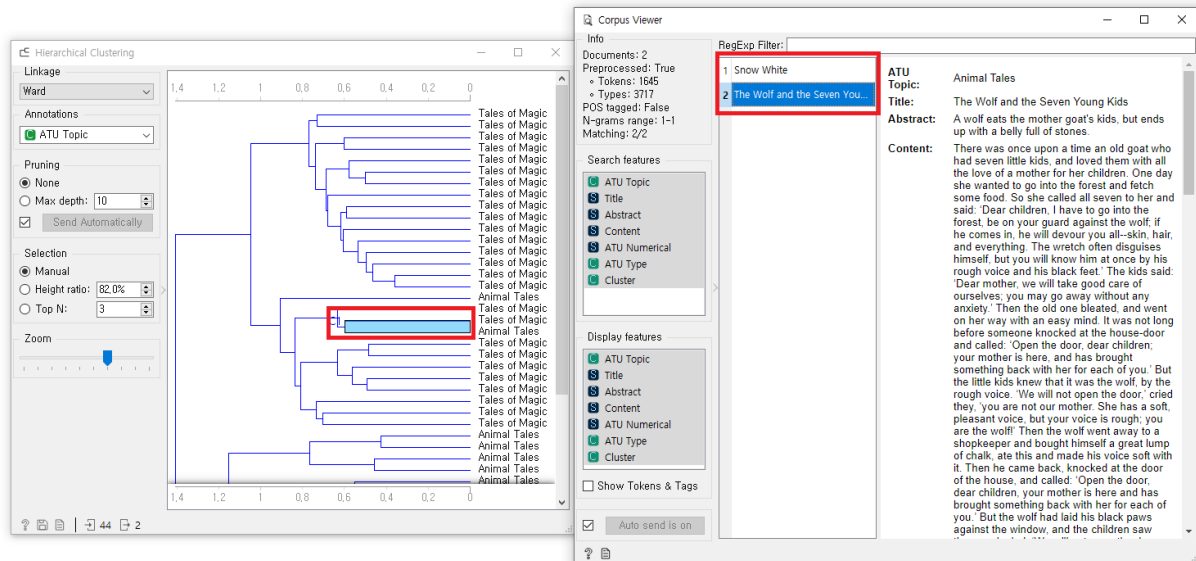
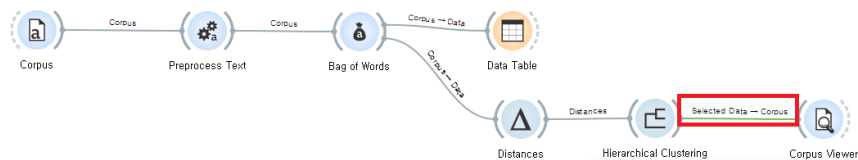
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Sentiment Analysis	텍스트에서 감정을 계산한다.	Corpus	Corpus

- Sentiment Analysis 위젯은 말뭉치의 각 문서에 대한 감성을 예측
- NLTK의 Liu&Hu 및 Vader 정서 모듈과 Data Science Lab의 다국어 정서 어휘를 사용

텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출

Sentiment Analysis

Method

☐ Liu Hu 1 Language: English

☒ Vader 2 Language: English

☐ Multilingual sentiment 3 Language: English

☐ SentiArt Language: English

☐ Custom dictionary 4

Positive: (none) [File icon] [Refresh icon]

Negative: (none) [File icon] [Refresh icon]

☒ Autocommit is on

① Liu hu

어휘 기반 정서 분석 최종 점수는 양의 단어와 음의 단어 합계의 차이로, 문서 길이로 정규화되고 100으로 곱합니다. 최종 점수는 문서의 감정 차이 백분율을 반영합니다.

② Vader

사전 및 규칙 기반 정서 분석입니다.

③ Language

여러 언어에 대한 사전 편집 기반 정서 분석입니다.

④ Custom dictionary

사용자 자신의 긍정 및 부정 감정 사전을 추가합니다. 허용된 소스 유형은 각 단어가 고유한 줄에 있는 .txt 파일입니다. 최종 점수는 류후와 같은 방식으로 계산됩니다.

텍스트 마이닝

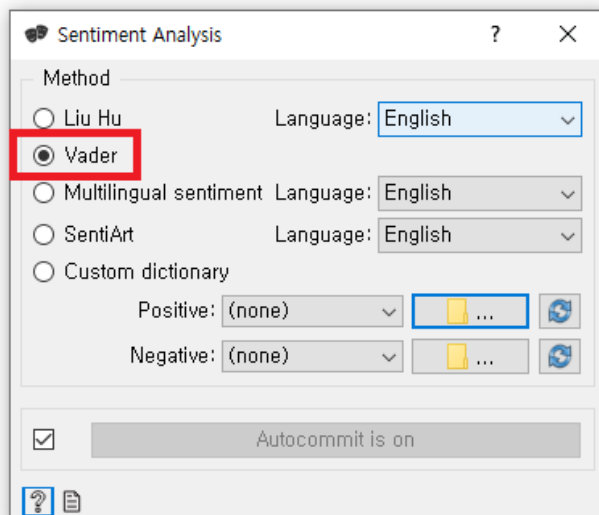
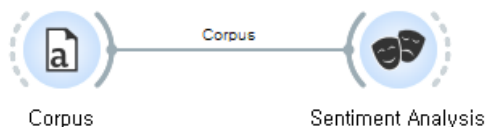
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출



Data Table

Info
44 instances (no missing data)
4 features
Target with 2 values
5 meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

include title	ATU Topic	Title	Abstract	Content True True	ATU Numerical	ATU Type	pos	neg	neu	compound
1	Tales of Magic	A Tale About t...	A simple boy ...	A certain father...	326.0	Supernatural ...	0.067	0.113	0.82	-0.9996
2	Tales of Magic	Brier Rose	An offended ...	A king and ...	410.0	Supernatural or...	0.133	0.055	0.812	0.9993
3	Animal Tales	Cat and Mouse...	A mouse lives ...	A certain cat h...	15.0	Wild Animals	0.101	0.065	0.834	0.9913
4	Tales of Magic	Cinderella	The familiar sto...	The wife of a ...	510A	Supernatural ...	0.07	0.071	0.859	0.9446
5	Tales of Magic	Hansel and ...	A poor ...	Hard by a grea...	327A	Supernatural ...	0.084	0.094	0.823	-0.9615
6	Animal Tales	Herr Korbes	A hen and a ...	Once upon a ...	210.0	Domestic ...	0.015	0.102	0.883	-0.9877
7	Tales of Magic	Jorinda and ...	A witch lures ...	There was once...	405.0	Supernatural or...	0.096	0.085	0.819	0.9788
8	Tales of Magic	Little Red Ridin...	A girl known f...	Once upon a ...	333.0	Supernatural ...	0.11	0.07	0.821	0.9962
9	Tales of Magic	Mother Holle	A widow spoils...	Once upon a ...	480.0	Supernatural ...	0.133	0.087	0.78	0.9976
10	Animal Tales	Old Sultan	A farmer decid...	A shepherd ha...	101.0	Wild Animal an...	0.123	0.101	0.776	0.9697
11	Animal Tales	Pack of ...	A rooster and a...	The rooster sai...	210.0	Domestic ...	0.02	0.089	0.891	-0.9955
12	Tales of Magic	Rapunzel	The classic stor...	There were onc...	310.0	Supernatural ...	0.106	0.113	0.782	-0.9096
13	Tales of Magic	Rumpelstiltskin	A miller's ...	By the side of ...	500.0	Supernatural ...	0.099	0.067	0.835	0.9923
14	Tales of Magic	Snow White	The classic stor...	There was once...	426.0	Supernatural or...	0.108	0.092	0.8	0.996
15	Tales of Magic	The Blue Light	A wounded ...	There was once...	562.0	Supernatural ...	0.082	0.077	0.841	0.9271
16	Animal Tales	The Bremen ...	A donkey, a do...	An honest ...	130.0	Wild Animal an...	0.111	0.137	0.753	-0.9924
17	Animal Tales	The Crumbs on...	A man tells his ...	One day the ...	236.0	Other Animals ...	0.074	0.057	0.869	0.7149
18	Animal Tales	The Dog and t...	A merchant ru...	A shepherd's ...	248.0	Other Animals ...	0.03	0.189	0.781	-0.9998
19	Tales of Magic	The Elves and ...	A poor ...	There was once...	503.0	Supernatural ...	0.168	0.052	0.78	0.9988
20	Tales of Magic	The Fisherman ...	A fisherman ...	There was once...	555.0	Supernatural ...	0.075	0.061	0.864	0.9895
21	Animal Tales	The Fox and th...	The fox is ...	it happened th...	105.0	Wild Animal an...	0.078	0.045	0.878	0.8719
22	Animal Tales	The Fox and th...	A hungry fox ...	The fox once ...	227.0	Other Animals ...	0.134	0.062	0.804	0.9541
23	Animal Tales	The Fox and th...	A farmer will ...	A farmer had a...	47A	Wild Animals	0.125	0.092	0.783	0.9537

텍스트 마이닝


전처리

문서 요약

문서 분류

문서 군집

특징 추출

위젯	설명	입력	출력
 Statistics	문서에 대한 새 통계 변수를 작성한다.	Corpus	Corpus

- **Statistics** 위젯은 문서 통계를 말뭉치에 추가하는 **feature** 생성 위젯
- 표준 통계 측정치와 사용자 정의 변수를 모두 지원
- **Features**를 + 기호와 함께 추가할 수 있으며 x기호로 제거할 수 있음

텍스트 마이닝

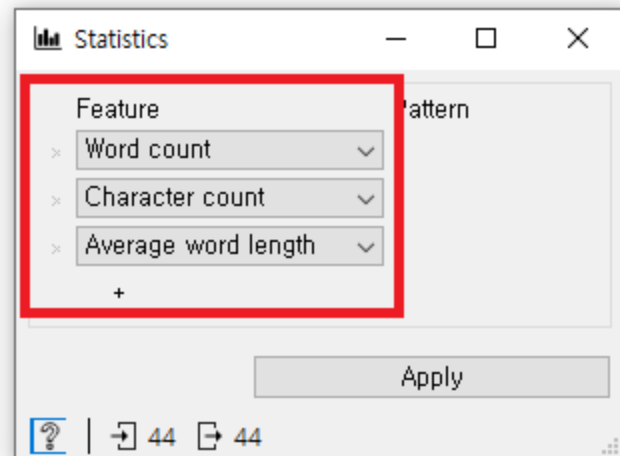
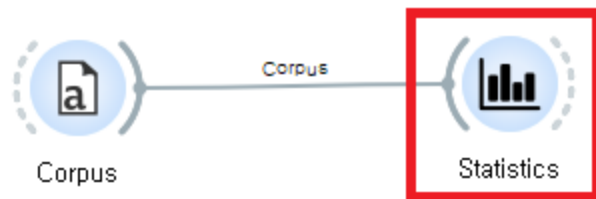
전처리

문서 요약

문서 분류

문서 군집

특징 추출



텍스트 마이닝

전처리

문서 요약

문서 분류

문서 군집

특징 추출



Data Table

Info
44 instances (no missing data)
3 features
Target with 2 values
5 meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☐ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	include title	ATU Topic	Title	Abstract	Content True True	ATU Numerical	ATU Type	Word count	Character count	Average word length
1		Tales of Magic	A Tale About t...	A simple boy ...	A certain father...	326.0	Supernatural ...	3780	14678	3.88307
2		Tales of Magic	Brier Rose	An offended ...	A king and ...	410.0	Supernatural or...	1510	6038	3.99868
3		Animal Tales	Cat and Mouse...	A mouse lives ...	A certain cat h...	15.0	Wild Animals	975	3835	3.93333
4		Tales of Magic	Cinderella	The familiar sto...	The wife of a ...	510A	Supernatural ...	2563	10045	3.91924
5		Tales of Magic	Hansel and ...	A poor ...	Hard by a grea...	327A	Supernatural ...	2929	11781	4.02219
6		Animal Tales	Herr Korbes	A hen and a ...	Once upon a ...	210.0	Domestic ...	352	1410	4.00568
7		Tales of Magic	Jorinda and ...	A witch lures ...	There was once...	405.0	Supernatural or...	1133	4523	3.99206
8		Tales of Magic	Little Red Ridin...	A girl known f...	Once upon a ...	333.0	Supernatural ...	1378	5617	4.0762
9		Tales of Magic	Mother Holle	A widow spoils...	Once upon a ...	480.0	Supernatural ...	1267	4997	3.94396
10		Animal Tales	Old Sultan	A farmer decid...	A shepherd ha...	101.0	Wild Animal an...	870	3360	3.86207
11		Animal Tales	Pack of ...	A rooster and a...	The rooster sai...	210.0	Domestic ...	802	3257	4.0611
12		Tales of Magic	Rapunzel	The classic stor...	There were onc...	310.0	Supernatural ...	1401	5588	3.98858
13		Tales of Magic	Rumpelstiltskin	A miller's ...	By the side of ...	500.0	Supernatural ...	1138	4382	3.85062
14		Tales of Magic	Snow White	The classic stor...	There was once...	426.0	Supernatural or...	2421	9847	4.06733
15		Tales of Magic	The Blue Light	A wounded ...	There was once...	562.0	Supernatural ...	1739	6915	3.97642
16		Animal Tales	The Bremen ...	A donkey, a do...	An honest ...	130.0	Wild Animal an...	1361	5270	3.87215
17		Animal Tales	The Crumbs on...	A man tells his ...	One day the ...	236.0	Other Animals ...	162	636	3.92593

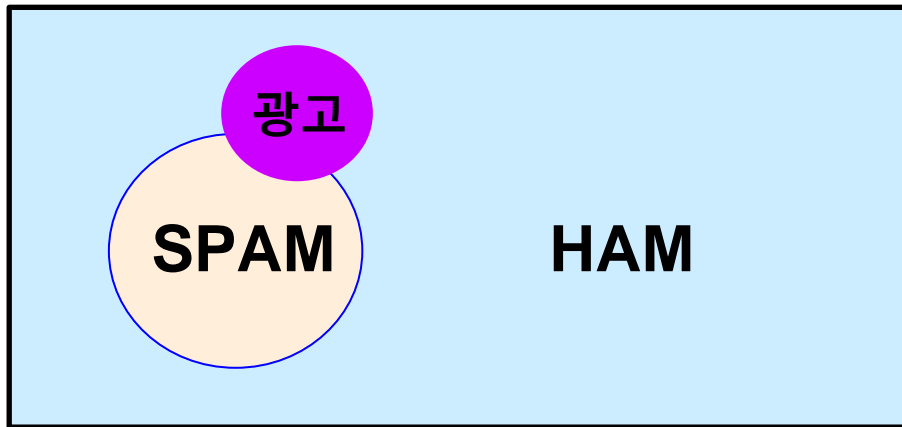
스팸메일 필터

- 질문: 어떤 이메일을 하나 수신했을 때, 이 이메일이 스팸일 확률은?

SPAM or HAM?



받은 편지함



Bayes' Theorem(베이즈 정리) ☆ ☆

- 베이즈 정리는 새로운 정보를 기반으로 이전의 확률을 갱신할 때 사용
- 18세기 토마스 베이즈에 의해 개발
- 조건부 확률의 확장버전

Prior probability(사전확률)을 기반으로 이후 사후확률을 업데이트 하는 것

베이즈 정리

The diagram illustrates Bayes' Theorem with the following components and labels:

- Posterior probability (사후확률)**: Labeled in red, with an arrow pointing to the left side of the equation, $P(B | A)$.
- Conditional probability (조건부확률)**: Labeled in red, with an arrow pointing to the numerator term $P(A | B)$.
- Prior probability (사전확률)**: Labeled in red, with an arrow pointing to the numerator term $P(B)$.
- Prior probability (사전확률)**: Labeled in red, with an arrow pointing to the denominator term $P(A)$.

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)}$$

- 어떻게 사후확률을 찾아낼 것인가?

Frequency table(빈도표) and Likelihood table(우도표)

“광고”			
포함	불포함		
SPAM	4	16	20
HAM	1	79	80
		5	95
		100	

“광고”			
포함	불포함		
SPAM	$\frac{4}{20}$ =0.2	$\frac{16}{20}$ =0.8	20
HAM	$\frac{1}{80}$ =0.0125	$\frac{79}{80}$ =0.9875	80
		5	95

베이즈 정리를 적용:

“광고”라는 단어가 들어간 메일이 스팸일 확률은?

$$\begin{aligned} P(SPAM|\text{광고}) &= \frac{P(\text{광고}|SPAM)P(SPAM)}{P(\text{광고})} \\ &= \frac{\frac{4}{20} \times \frac{20}{100}}{\frac{5}{100}} = 0.80 \text{ 확률이 증가} \end{aligned}$$

질문 있나요?

hsryu13@hongik.ac.kr

