

데이터 과학 기초

07

텍스트 분석

경북대학교 배준현 교수
(joonion@knu.ac.kr)



07. 텍스트 분석

- 비정형 데이터: *Unstructured* Data
 - 미리 정의된 데이터 모델이 없거나, 미리 정의된 방식으로 정리되지 않은 정보
 - 이미지, 텍스트, 사운드, 동영상, 기타 등등
 - 비정형 데이터의 처리: *Embedding*
 - 비정형 데이터의 특징을 추출하여 정형 데이터로 바꾸기
 - 이미지 임베딩: *ImageNet*
 - 텍스트 임베딩: *Bag of Words, Word2Vec*



07. 텍스트 분석

- 자연어 처리: *NLP, Natural Language Processing*
 - 자연어: 사람이 일상 생활에서 사용하는 언어
 - 자연어 처리: 번역, 요약, 분류, 감성 분석, 챗봇, 기타 등등
 - 자연어의 구성: 문서, 문장, 단어
 - 문서: 문장들의 집합. *Document*
 - 문장: 단어들의 집합
 - 단어: 텍스트 분석의 기본 단위. *Word = Term*



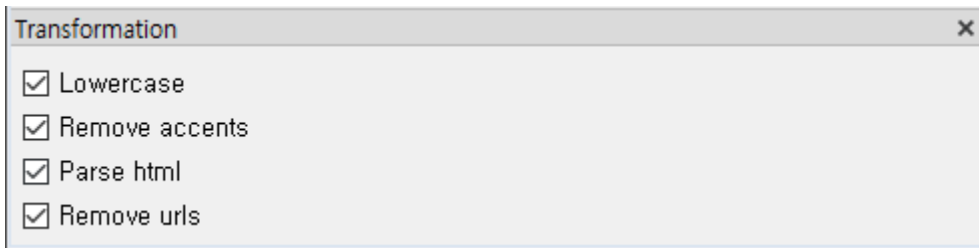
07. 텍스트 분석

- 텍스트 전처리: Text Preprocessing
 - 코퍼스(*Corpus*): 말뭉치. 텍스트 분석을 위한 데이터셋.
 - 텍스트 분석을 위한 전처리 과정:
 - 변환: Transformation
 - 토큰화: Tokenization
 - 정규화: Normalization
 - 필터링: Filtering
 - N-그램: N-grams Range
 - 품사 태깅: POS Tagger

07. 텍스트 분석

■ 변환: *Transformation*

- Text *Cleansing*: 텍스트에서 불필요한 요소를 제거하고 수정.

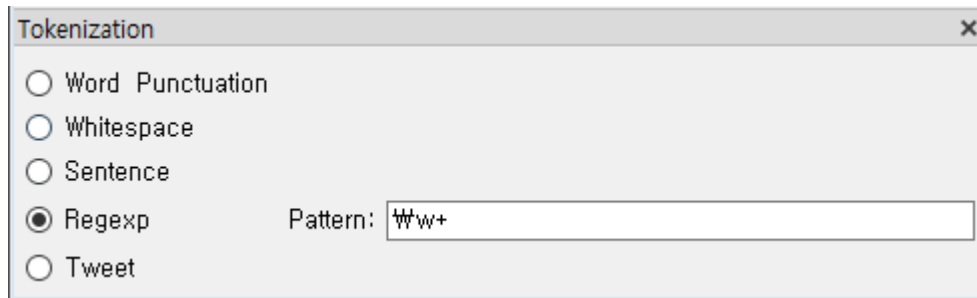




07. 텍스트 분석

■ 토큰화: *Tokenization*

- 문서에서 문장을 분리, 또는, 문장에서 단어를 분리
- 정규 표현식: *Regular Expression*
 - 규칙을 기반으로 토큰을 분리할 수 있음

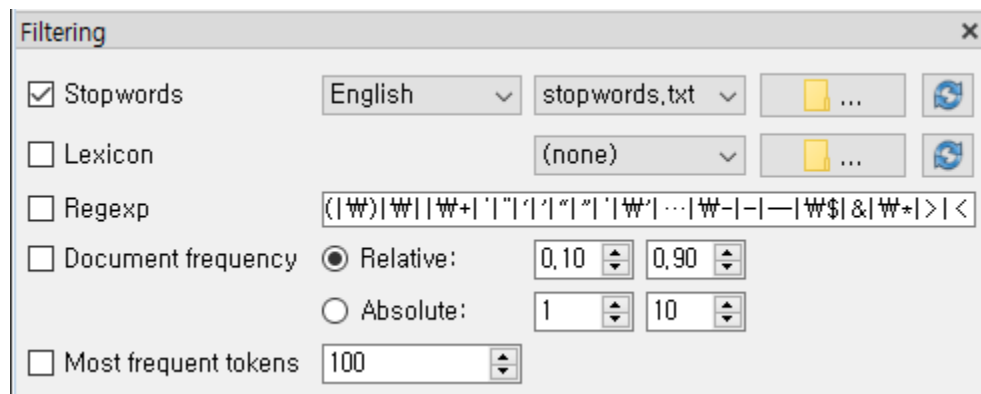




07. 텍스트 분석

■ 필터링: *Filtering*

- 불용어 (*Stopwords*) 처리: 분석에 필요하지 않은 단어를 제거
- 출현 빈도수가 높은 단어만 선택

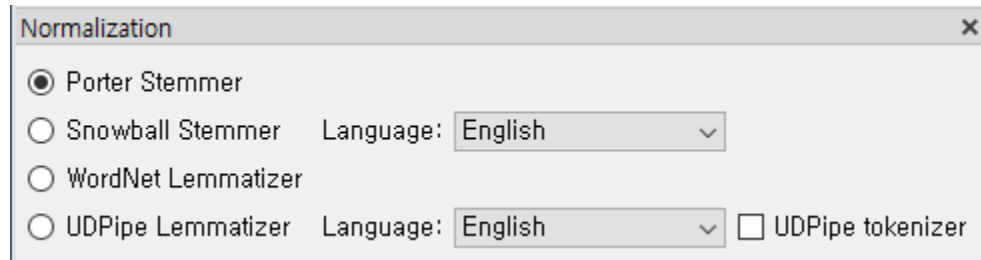




07. 텍스트 분석

■ 정규화: *Normalization*

- 형태소 분석: 같은 의미지만 서로 다른 단어들을 하나의 단어로 일반화
 - 어간 추출: *Stemming*
 - 표제어 추출: *Lemmatization*





07. 텍스트 분석

- N -그램: N -grams Range
 - 분석 결과의 단위를 연속적인 N 개의 토큰으로 구성

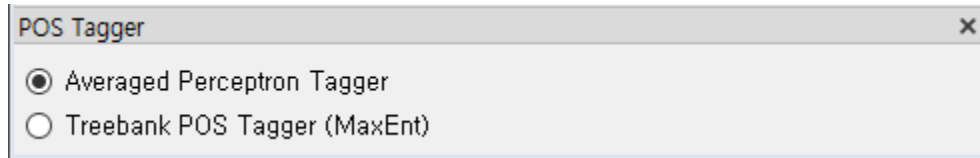
N-grams Range

Range: 1 2

How are you? Fine, thank you. And you?

07. 텍스트 분석

- 품사 태깅: POS(Part-Of-Speech) Tagger
 - 각 단어의 품사를 태그하여 단어의 의미 파악

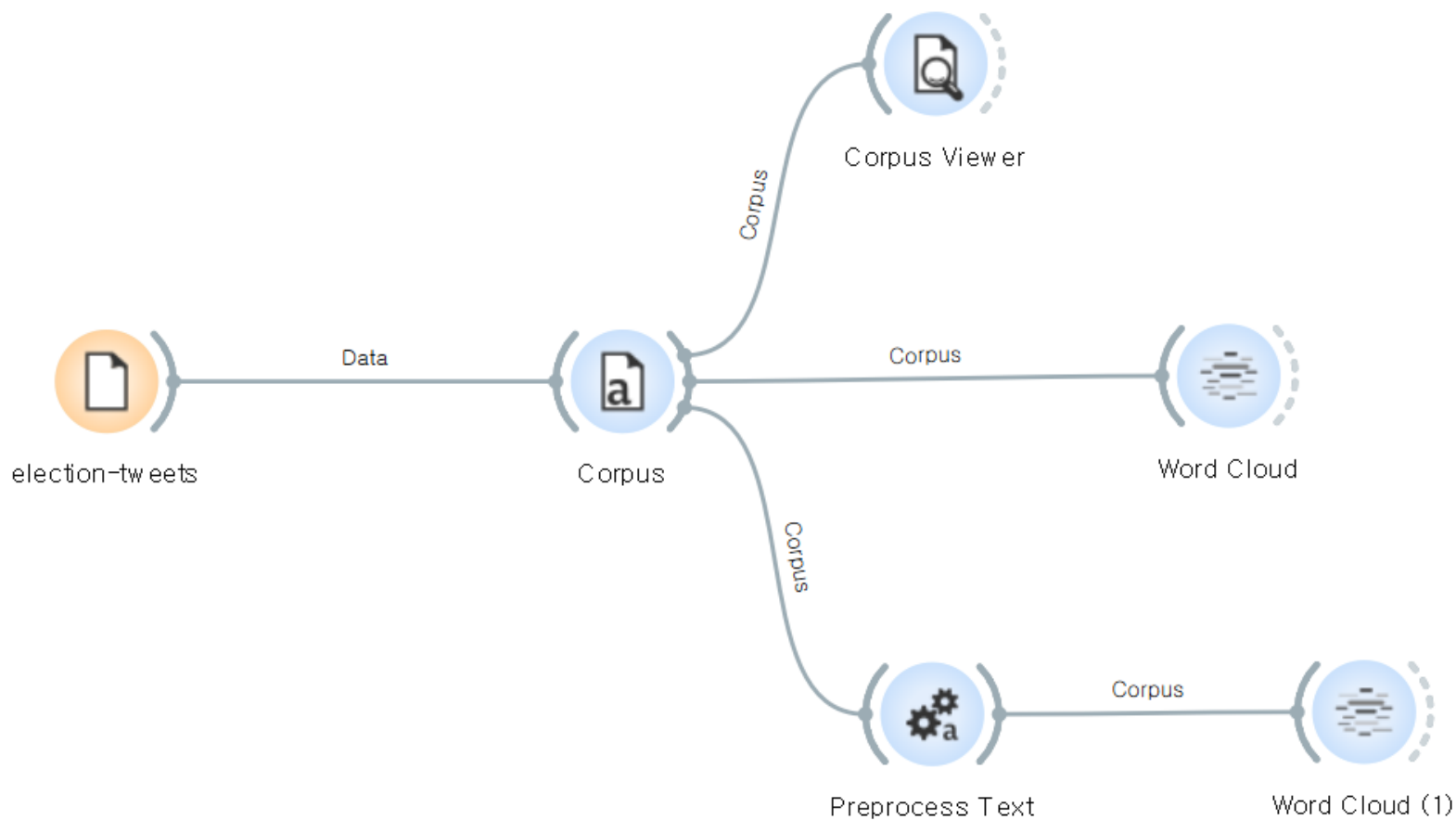


I believe I can *fly*. Because I am a *fly*.



07. 텍스트 분석

■ Orange: Text Mining





07. 텍스트 분석

election-tweets

Source

☒ File: election-tweets-2016.tab

☐ URL:

Info

6444 instance(s)
7 feature(s) (28.5% missing values)
Classification: categorical class with 2 values (no missing values)
4 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	Is retweet	categorical	feature	False, True
2	Time	datetime	feature	
3	Language	categorical	feature	da, en, es, et, fi, fr, tl, und
4	Retweet count	numeric	feature	
5	Favorite count	numeric	feature	
6	Longitude	numeric	feature	
7	Latitude	numeric	feature	
8	Author	categorical	target	HillaryClinton,realDonaldTrump
9	Source URL	text	meta	
10	Content	text	meta	
11	Original author	text	meta	
12	Place	text	meta	

? | 6444

Corpus

Corpus file

reuters-r8-test.tab

Title variable

Original author

Used text features

Content

Ignored text features

Source URL
 Original author
 Place

? | 6444 6444



07. 텍스트 분석

Corpus Viewer

Info

Documents: 6444
Preprocessed: False
Tokens: n/a
Types: n/a
POS tagged: False
N-grams range: 1-1
Matching: 6444/6444

Search features

- ☒ Is retweet
- ☐ Time
- ☐ Language
- ☐ Retweet count
- ☐ Favorite count
- ☐ Longitude
- ☐ Latitude
- ☐ Author
- ☐ Source URL
- ☐ Content
- ☐ Original author
- ☐ Place

Display features

- ☒ Time
- ☒ Language
- ☐ Retweet count
- ☐ Favorite count
- ☐ Longitude
- ☐ Latitude
- ☐ Author
- ☐ Source URL
- ☐ Content
- ☐ Original author
- ☐ Place

☐ Show Tokens & Tags

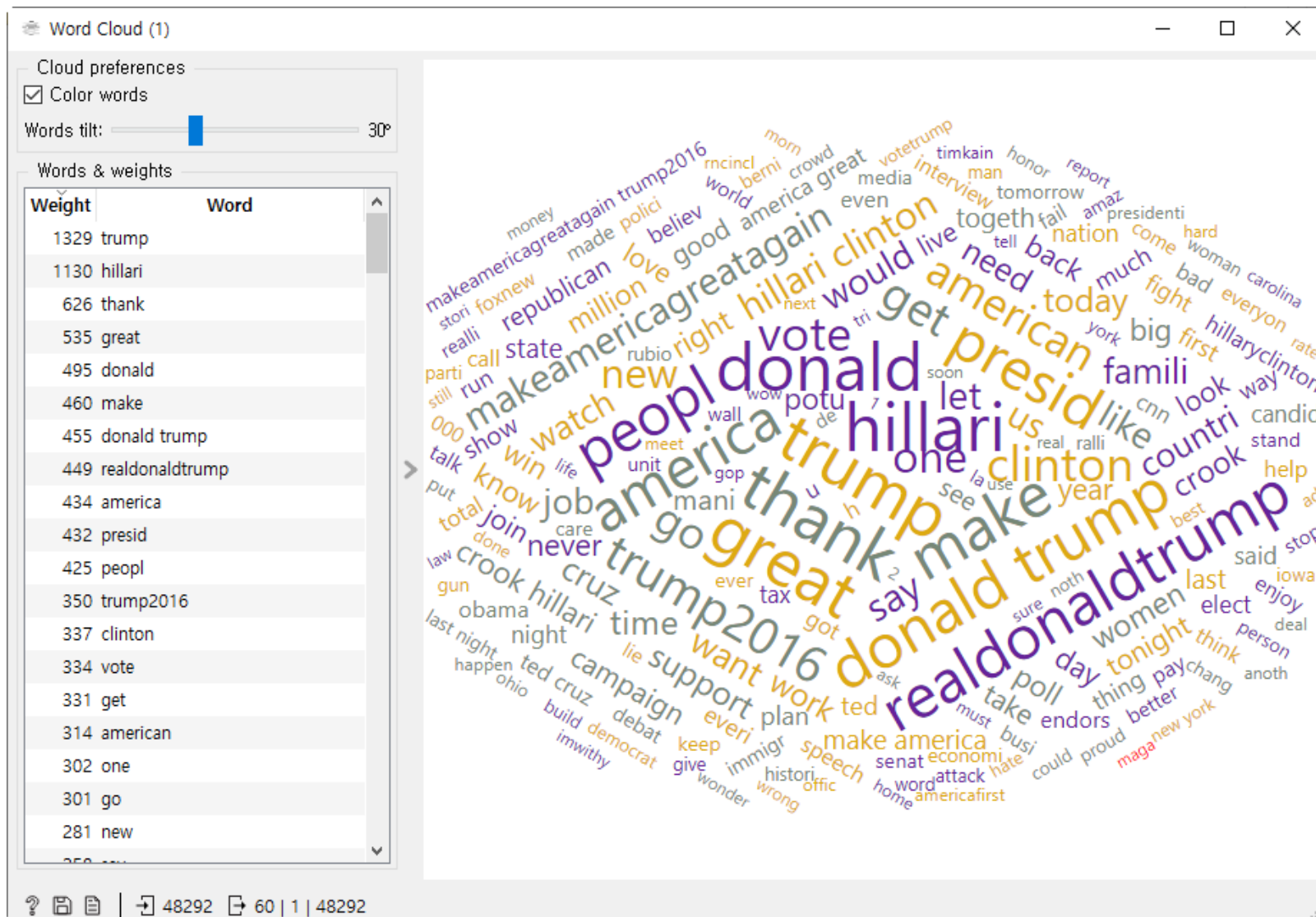
☒ Auto send is on

RegExp Filter:

1	https://studio.twitter.com (1)
2	http://twitter.com (1)
3	https://about.twitter.com/...
4	https://studio.twitter.com (2)
5	https://about.twitter.com/...
6	http://twitter.com/download/...
7	https://about.twitter.com/...
8	https://about.twitter.com/...
9	http://twitter.com/download/...
10	http://twitter.com (2)
11	http://twitter.com (3)
12	http://twitter.com/download/...
13	http://twitter.com/download/...
14	http://twitter.com/download/...
15	https://about.twitter.com/...
16	http://twitter.com/download/...
17	http://twitter.com/download/...
18	https://about.twitter.com/...
19	https://about.twitter.com/...
20	https://about.twitter.com/...
21	https://about.twitter.com/...
22	https://about.twitter.com/...
23	https://studio.twitter.com (3)
24	https://about.twitter.com/...

Time: 2016-09-27 21:35:28
Language: en
Retweet count: 1303.0
Favorite count: 2849.0
Longitude: ?
Latitude: ?
Author: HillaryClinton
Source URL: https://about.twitter.com/products/tweetdeck
Content: This election is too important to sit out. Go to https://t.co/tTgeqxNqYm and make sure you're registered. #NationalVoterRegistrationDay -H
Original author: ?
Place: ?

6444 1 | 6443



07. 텍스트 분석

- 단어 가방: *Bag of Words*, BoW
 - 문서(*Document*)에 포함되어 있는 단어(*Word*)의 빈도 수로 특징 추출
 - 순서나 문맥을 무시하고 가방에 단어를 담기 때문에 처리가 간단함

D.1: John likes to watch movies. Mary likes movies too.

D.2: Mary also likes to watch football games.

BoW.1: John:1, likes:2, to:1, watch:1, movies:2, Mary:1, too:1

BoW.2: Mary:1, also:1, likes:1, to:1, watch:1, football:1, games:1



07. 텍스트 분석

- 단어 빈도: *TF, Term Frequency*
 - 특정 단어가 문서 내에서 얼마나 자주 등장하는가?

BoW.1: John:1, likes:2, to:1, watch:1, movies:2, Mary:1, too:1

BoW.2: Mary:1, also:1, likes:1, to:1, watch:1, football:1, games:1

	John	likes	to	watch	movies	Mary	too	also	football	games
D.1	1	2	1	1	2	1	1	0	0	0
D.2	0	1	1	1	0	1	0	1	1	1

07. 텍스트 분석

- 문서-단어 행렬: *DTM, Document-Term Matrix*
 - 각 문서별로 해당 단어가 몇 번 나타나는가를 표시한 행렬
 - Term Frequency:
 - *Count*: 단어가 나타나는 횟수로 만드는 경우
 - *Binary*: 단순히 단어가 포함되어 있는가의 여부
 - Sublinear: 단어-빈도(TF)의 로그값

07. 텍스트 분석

- 문서 빈도: *DF, Document Frequency*
 - 단어의 빈도수만으로 벡터화를 할 경우, 단어의 중요성을 간과하게 됨.
 - 예) The anatomy of a large-scale hypertextual web search engine.
 - 특별한 단어에 **가중치**를 주는 방법: 단어의 **희소성**을 고려
 - 개별 문서에 자주 나타나는 단어는 높은 가중치를 주되,
 - 모든 문서에 전반적으로 자주 나타나는 단어에 대해서는 페널티를 부여.
 - Document Frequency:
 - *IDF: Inverse Document Frequency*
 - Smooth IDF: Divide-by-Zero 방지를 위해 DF에 1을 더한 값을 사용

07. 텍스트 분석

- *TF-IDF*: Term Frequency - Inverse Document Frequency
 - 텍스트 마이닝과 정보 검색에서 많이 이용하는 가중치 부여 방법
 - TF-IDF는 단어 빈도(TF)와 역 문서 빈도(IDF)를 곱한 값
 - 전체 문서를 D , 문서를 d , 단어를 t 라 할 때
 - $TFIDF(t, d, D) = TF(t, d) * IDF(t, D)$
 - $TF(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\max\{f(w, d) : w \in d\}}$
 - $IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$



07. 텍스트 분석

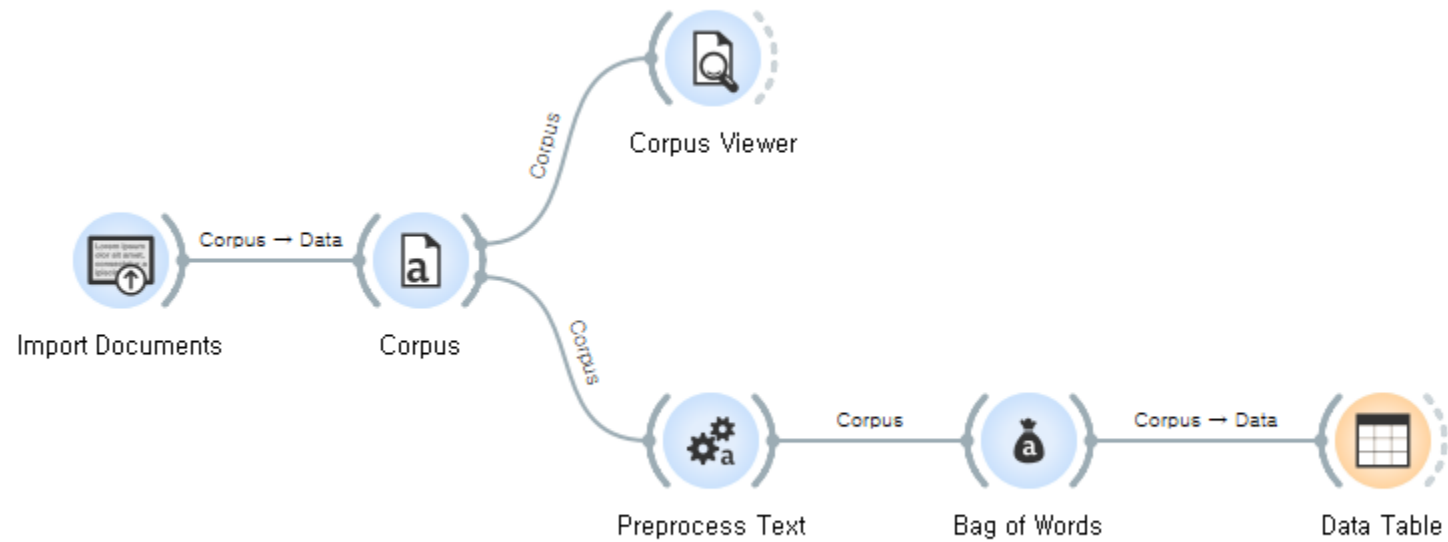
■ 단어 가방의 한계:

- 희소 행렬: *Sparse Matrix*
 - 많은 문서에서 많은 단어를 추출하면 단어 집합의 크기가 커지고, 각 문서에 포함된 단어의 수는 일정하므로, 매우 큰 희소 행렬이 됨.
 - 희소 행렬의 연산에 적절한 행렬 처리가 필요함
- 문맥과 의미: *Contexts* and *Semantics*
 - BoW는 단어의 순서를 고려하지 않으므로 문맥적 의미를 무시함.
 - 단어와 문장의 순서를 고려하여 문맥과 의미를 반영할 필요가 있음.



07. 텍스트 분석

■ Orange: Bag of Words





07. 텍스트 분석

Corpus Viewer

Info

Documents: 2
Preprocessed: False
• Tokens: n/a
• Types: n/a
POS tagged: False
N-grams range: 1-1
Matching: 2/2

Search features

- ☒ name
- ☒ path
- ☒ content

Display features

- ☒ name
- ☒ path
- ☒ content

☐ Show Tokens & Tags

☒ Auto send is on

RegExp Filter:

Document	name	content
1 Document 1	1	John likes to watch movies. Mary likes movies too.
2 Document 2		

? | 2 | 1 | 1



07. 텍스트 분석

Bag of Words
?
×

Options

Term Frequency: Count

Document Frequency: (None)

Regularization: (None)

☐ Hide bow attributes

☒ Commit Automatically

?
📄
|
↩ 2
↪ 2

{...}

john=1, likes=2, mary=1, movies=2, to=1, too=1, watch=1
also=1, football=1, games=1, likes=1, mary=1, to=1, watch=1

{...}

john, likes, mary, movies, to, too, watch
also, football, games, likes, mary, to, watch

{...}

john=0.693147, movies=1.38629, too=0.693147
also=0.693147, football=0.693147, games=0.693147

07. 텍스트 분석

- 텍스트 분류: Text *Classification*
 - 주어진 문서를 특정 카테고리로 분류하는 기법
 - 예) 뉴스 기사 자동 분류, 스팸 메일 필터링.

07. 텍스트 분석

- Dataset: reuters-r8-**train**.tab
 - 원본: UCI M/L Repository
 - Reuters-21578 Text Classification Collection
 - <https://archive-beta.ics.uci.edu/ml/datasets/137>



The screenshot shows the UCI M/L Repository entry for the Reuters-21578 Text Categorization Collection. The header is blue with a database icon, the title 'Reuters-21578 Text Categorization Collection', and the date 'Donated 1997-09-26'. Below this, it shows '19115 views' and '0 citations'. There are 'Download' and 'Cite' buttons. The 'General Information' section is expanded, showing an 'Abstract' that states: 'This is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories.'

Reuters-21578 Text Categorization Collection
Donated 1997-09-26

19115 views 0 citations

Download Cite

General Information [\[edit\]](#)

Abstract
This is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories.

07. 텍스트 분석

■ 데이터 탐색:

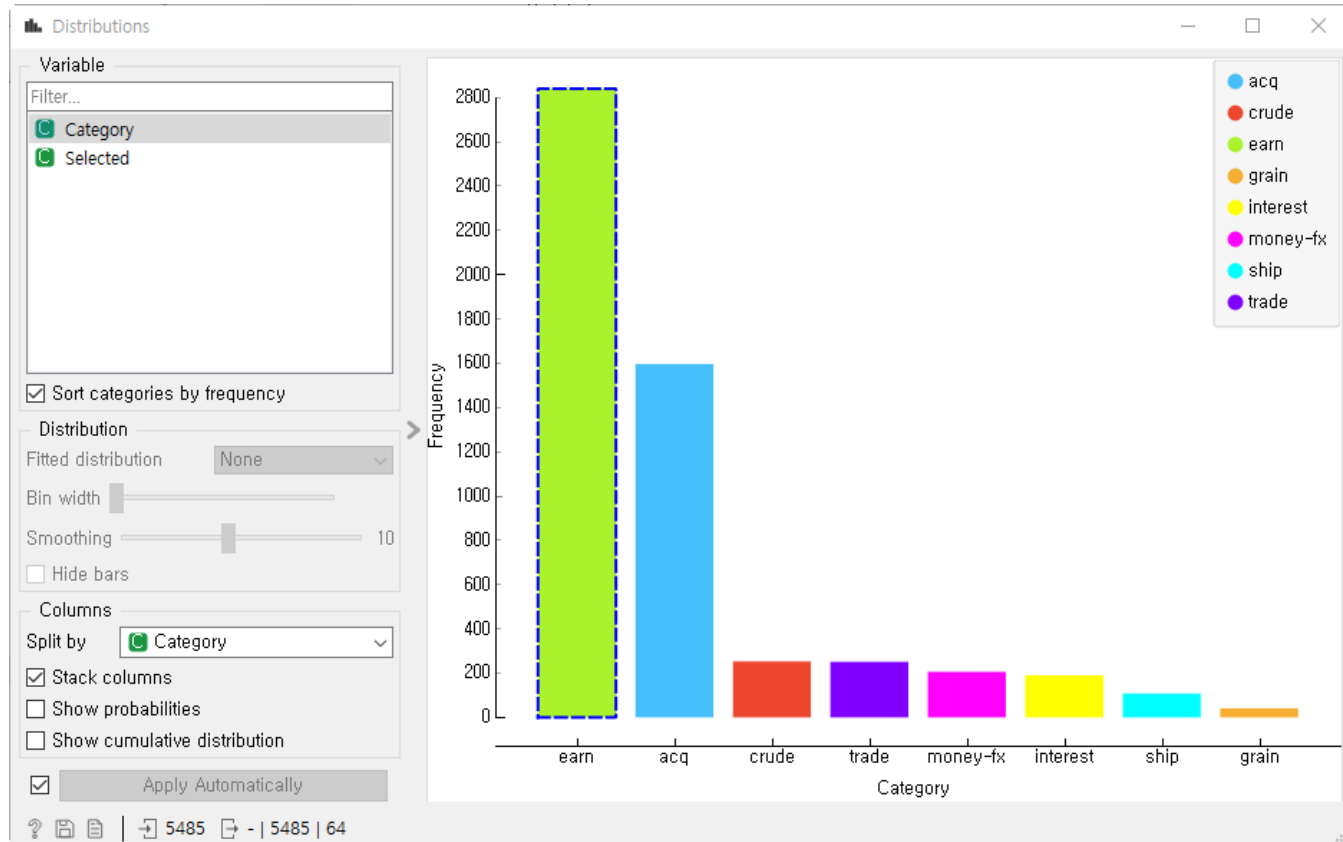
- Reuters-R8: 로이터의 뉴스기사 모음. R8은 전체 중 8개의 섹션만 모음.
- About the preprocessing:
 - Substitute TAB, NEWLINE and RETURN characters by SPACE.
 - Keep only letters: punctuation, numbers are removed.
 - Turn all letters to lowercase.
 - Substitute multiple SPACES by a single SPACE.
 - The title/subject of each document is simply added in the beginning of the document's text.



07. 텍스트 분석

■ 데이터 탐색:

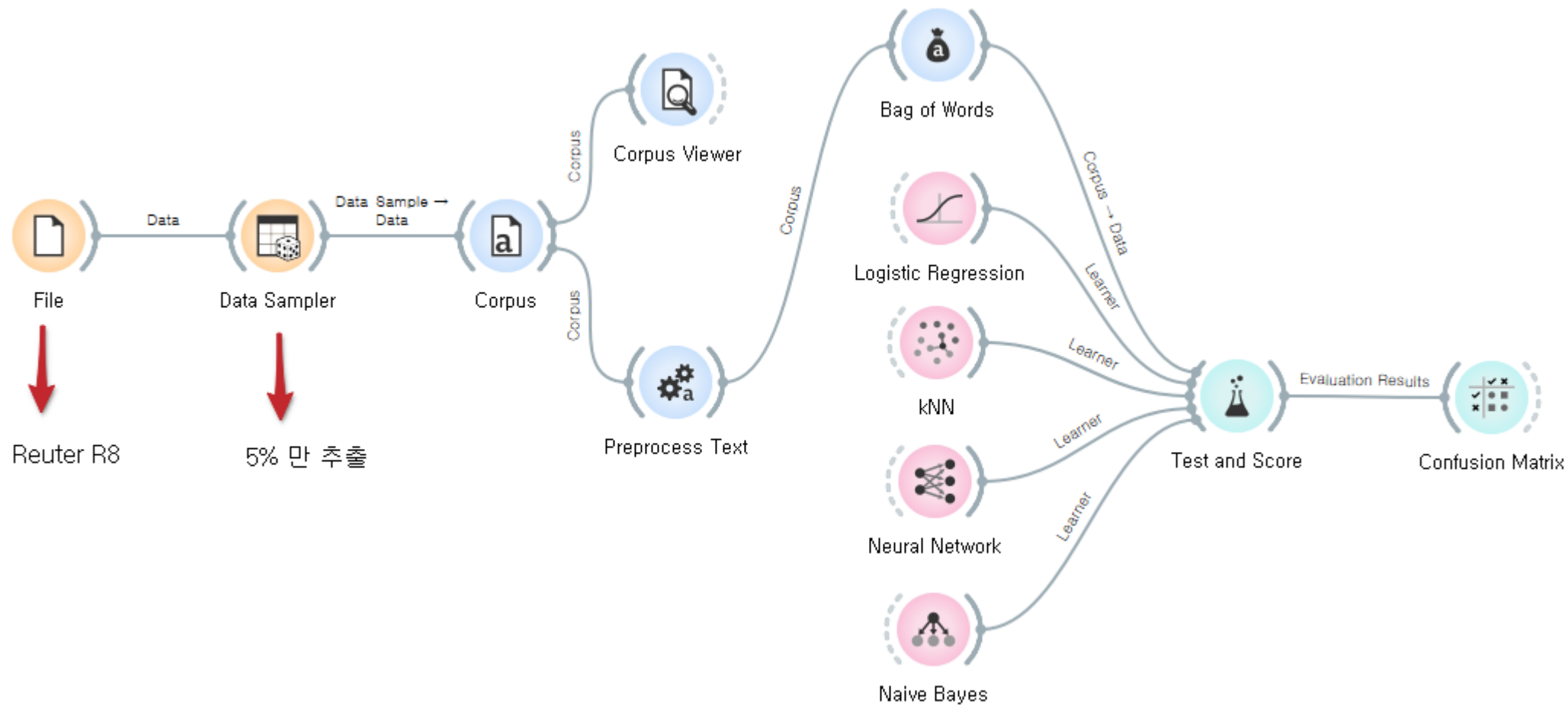
- 총 5,485개의 기사: 총 8개의 섹션
 - earn, acq, crude, trade, money-fx, interest, ship, gain





07. 텍스트 분석

로이터 기사 섹션 분류기:





07. 텍스트 분석

File

Source

☒ File: reuters-r8-train.tab

☐ URL:

Info

5485 instance(s)
0 feature(s) (no missing values)
Classification: categorical class with 8 values (no missing values)
1 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	Category	categorical	target	acq, crude, earn, grain, interest, money-fx, ship, trade
2	Text	text	meta	

? | 5485

Data Sampler

Sampling Type

☒ Fixed proportion of data:
5 %

☐ Fixed sample size
Instances:
☐ Sample with replacement

☐ Cross validation
Number of subsets:
Unused subset:

☐ Bootstrap

Options

☒ Replicable (deterministic) sampling
☒ Stratify sample (when possible)

? | 5485 | 275 | 5210



07. 텍스트 분석

Corpus Viewer

Info
Documents: 275
Preprocessed: False
• Tokens: n/a
• Types: n/a
POS tagged: False
N-grams range: 1-1
Matching: 275/275

Search features
Category
Text

Display features
Category
Text

☐ Show Tokens & Tags
☒ Auto send is on

RegExp Filter:

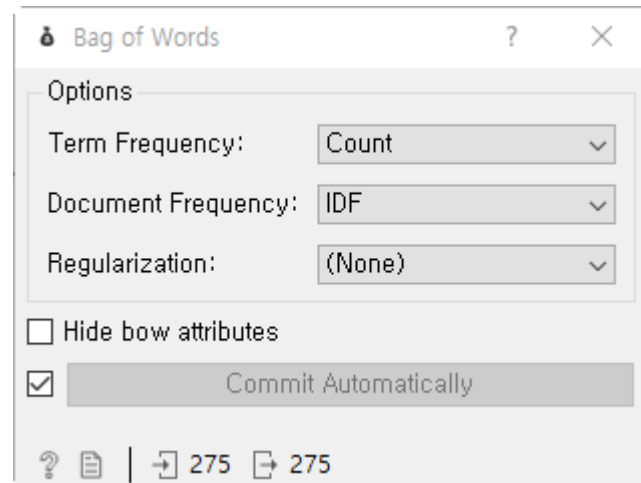
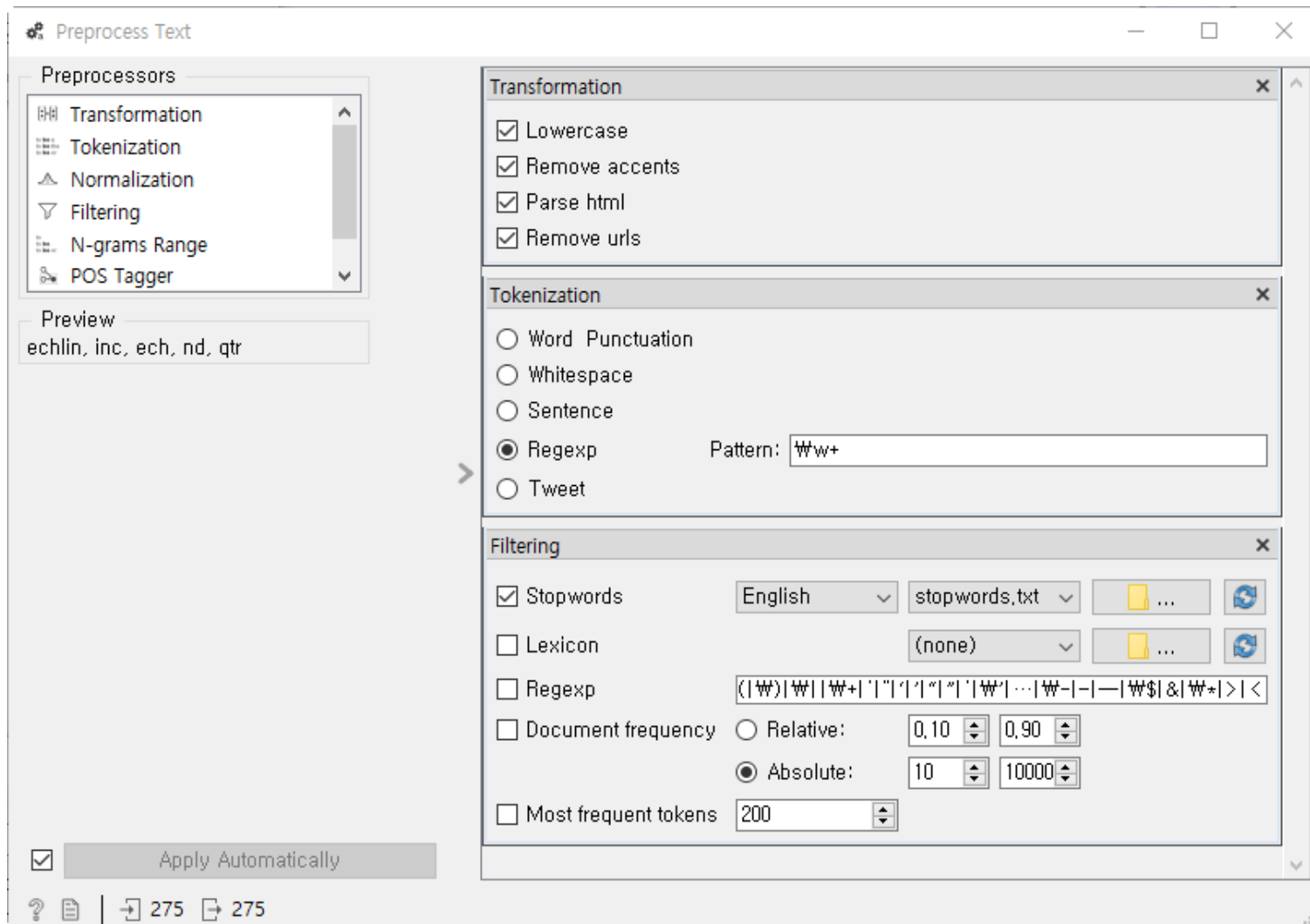
1	echlin inc ech nd qtr feb net shr cts vs cts net vs...
2	renault chrysler in accord for chrysler to buy am...
3	ussr crop weather summary usda noaa gradual ...
4	hoogovens concludes takeover of philips cirex ...
5	teck to increase stake in trilogy resource trilogy ...
6	regal petroleum ltd rplo year shr loss nine cts ne...
7	schwab completes purchase of schwab and co ...
8	dart darta makes offer for supermarkets sgl dart ...
9	carver corp cavr th qtr net shr cts vs cts net vs ...
10	usair u rejects twa twa takeover bid usair group ...
11	convest energy partners ltd cep th qtr loss shr lo...
12	general partners in gencorp gy proxy fight gener...
13	ivaco inc year net shr dlrs vs dlrs net vs revs ...
14	phh group phh regular qtlly dividend qtlly div cts ...
15	intelligent business ibcc st qtr jan shr three cts v...
16	videotron buys into exhibit company groupe ...
17	gulf arab ministers discuss economic cooperatio...
18	bristol myers bmy reviewing scimed merger ...
19	opac wants dlr oil price opac official opac

Category: money-fx
Text: gulf arab ministers discuss economic cooperation finance and economy ministers of the gulf cooperation council gcc opened a two day meeting to discuss further economic integration officials said they said issues to be discussed by the ministers from bahrain kuwait oman qatar saudi arabia and the united arab emirates uae would include a recommendation by central bank governors on a common currency exchange rate the governors agreed in january on a denominator on which to base currencies of the six states any decision will be forwarded for final approval to a gcc summit meeting due in saudi arabia late this year the six states have different currency systems saudi arabia bahrain qatar and the uae are linked in theory to the international monetary fund s basket of currencies the special drawing right sdr but in practice to the dollar oman links its currency formally to the dollar while kuwait pegs its dinar to a trade weighted basket devised by itself the denominator chosen by central bank governors has not been disclosed but some bankers expect the currencies to be linked to the sdr or a trade weighted basket opening the meeting ahmed al tayer the uae s minister of state for finance and industry said implementation of joint economic agreements is increasingly linking the interests of gcc citizens together the general assembly of the gulf investment corporation met in abu dhabi earlier under the chairmanship of bahrain s finance and national economy minister ibrahim abdul karim the corporation was formed to contribute to joint economic and investment projects in the gcc officials said the corporation s assets rose to billion dollars last year from billion at the end of reuter

? | 275 | 1 | 274



07. 텍스트 분석





07. 텍스트 분석

Logistic Regression

Name: Logistic Regression

Regularization type: Lasso (L1)

Strength: Weak Strong

C=1

☐ Balance class distribution

☒ Apply Automatically

kNN

Name: kNN

Neighbors

Number of neighbors: 7

Metric: Euclidean

Weight: Uniform

☒ Apply Automatically

Neural Network

Name: Neural Network

Neurons in hidden layers: 100

Activation: ReLu

Solver: SGD

Regularization, $\alpha=0.0001$:

Maximal number of iterations: 200

☒ Replicable training

Cancel ☐ Apply

Naive Bayes

Name: Naive Bayes

☒ Apply Automatically



07. 텍스트 분석

Test and Score

Sampling

- ☒ Cross validation
 - Number of folds: 10
 - ☐ Stratified
- ☐ Cross validation by feature
- ☐ 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
kNN	0.782	0.516	0.352	0.267	0.516
Neural Network	0.921	0.771	0.738	0.757	0.771
Naive Bayes	0.679	0.196	0.258	0.519	0.196
Logistic Regression	0.962	0.836	0.817	0.810	0.836

Model Comparison by AUC

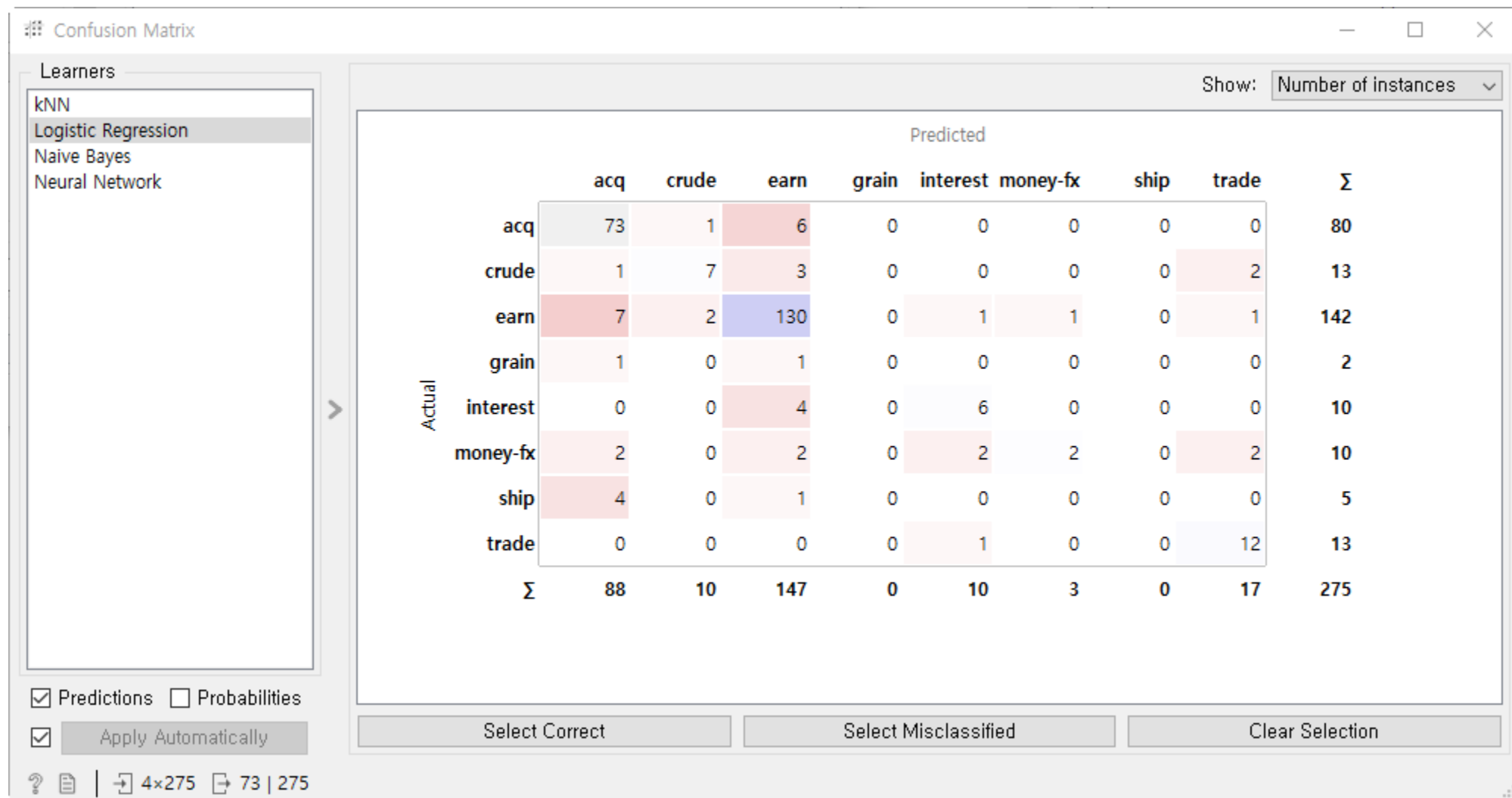
	kNN	Neural Net...	Naive Bayes	Logistic Re...
kNN		0.002	0.978	0.000
Neural Network	0.998		1.000	0.015
Naive Bayes	0.022	0.000		0.000
Logistic Regression	1.000	0.985	1.000	

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.

275 | - | 275 | 4x275



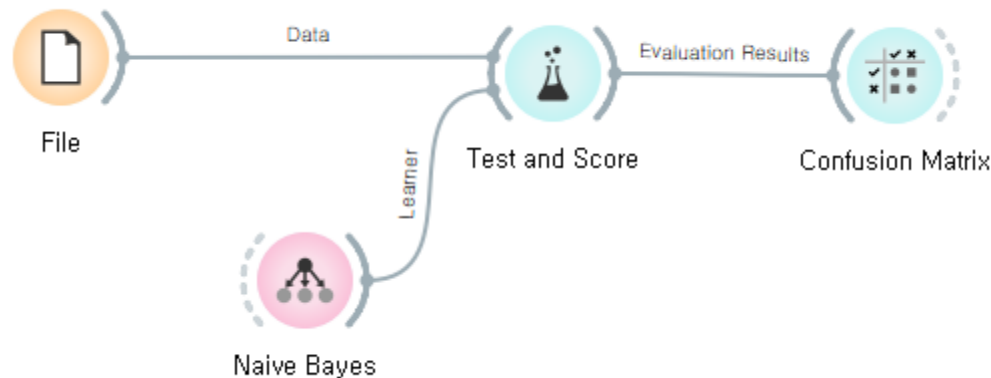
07. 텍스트 분석





07. 텍스트 분석

- 나이브 베이지안 분류기: *Naïve Bayesian* Classifier
 - 특징변수(*feature*)가 서로 독립사건이라는 **순진한** 가정하에
 - **베이즈 정리**를 적용하여 **확률적**으로 목적변수(*target*)를 추론하는 분류기
 - 예) **스팸 메일** 분류기, 암 진단



07. 텍스트 분석

■ 확률: *Probability*

- 시행, 사건, 확률: *trial, event, and probability*
 - 시행 = 동전 던지기, 사건의 집합: $S = \{H, T\}$
 - 확률: $P(H)$ = 동전 던지기를 했을 때, 앞면(H)이 나올 확률
 - 사건: 상호배타적이고 포괄적(*mutually exclusive* and *exhaustive*)

07. 텍스트 분석

■ 결합 확률: *Joint* Probability

- $P(A \cap B)$: 두 사건 A 와 B 가 동시에 일어날 확률
- 독립 사건: *Independent* Events
 - 사건 A 가 일어나는 것과 무관하게 사건 B 가 일어나는 경우
 - 예) A : 동전의 앞면이 나오는 사건, B : 주사위의 짝수가 나오는 사건
- 종속 사건: *Dependent* Events
 - 두 사건 A 와 B 가 독립사건이 아닌 경우
 - 예) A : 주사위의 홀수가 나오는 사건, B : 주사위의 소수가 나오는 사건
- 두 사건 A, B 가 독립사건일 경우:
 - $P(A \cap B) = P(A) \times P(B)$

07. 텍스트 분석

- **조건부 확률: *Conditional* Probability**
 - $P(A|B)$: 사건 B 가 발생한 경우, 사건 A 가 발생할 확률
 - 두 사건 A 와 B 가 **독립사건**이라면,
 - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
 - $P(B|A) = \frac{P(A \cap B)}{P(A)}$
 - 따라서, $P(A|B) \times P(B) = P(B|A) \times P(A)$



07. 텍스트 분석

■ 베이즈 정리: *Bayes Theorem*

사후 확률
(*posterior* probability)

우도
(*likelihood*)

사전 확률
(*prior* probability)

증거
(*evidence*)

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



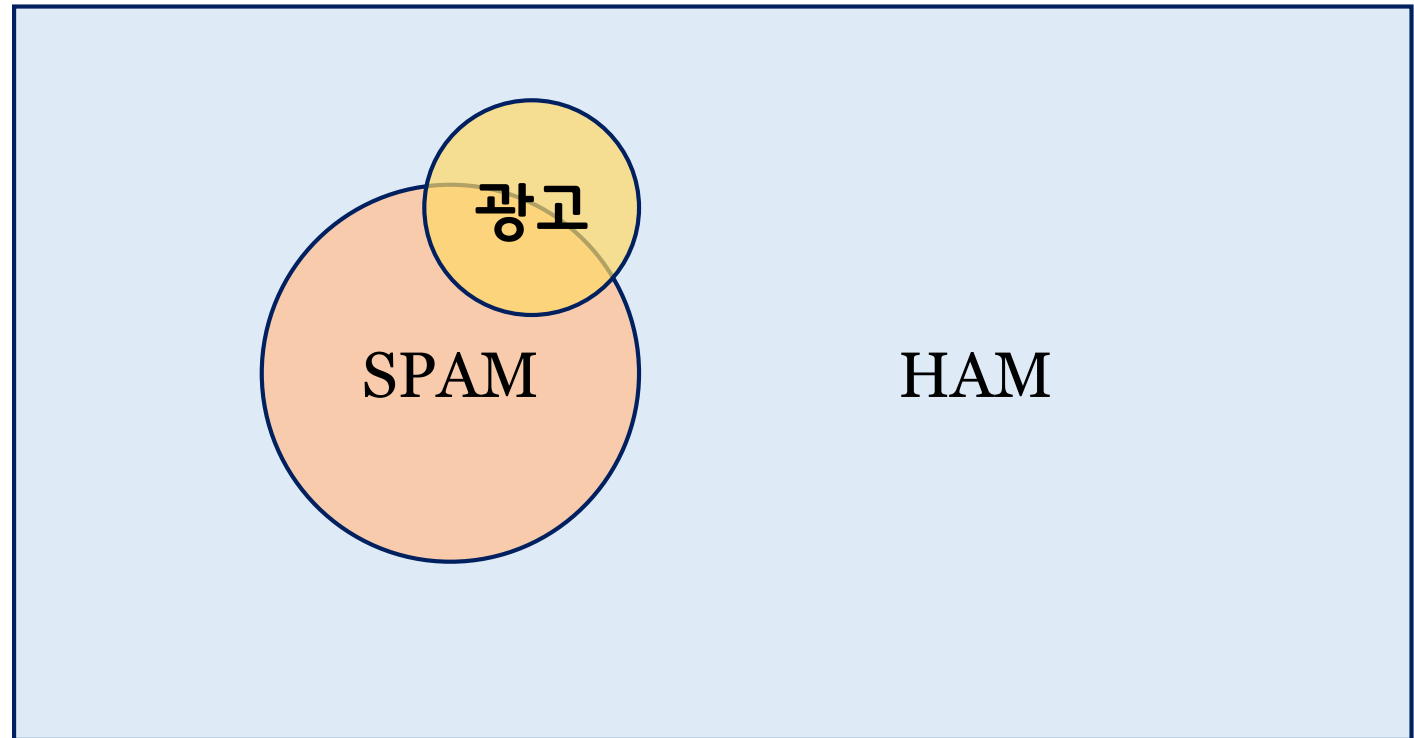
07. 텍스트 분석

■ 스팸 메일 분류기 만들기:

SPAM? or *HAM?*



받은 편지함





07. 텍스트 분석

■ 과거 데이터로부터의 학습:

- 빈도표와 우도표: *frequency* table and *likelihood* table

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

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

07. 텍스트 분석

■ 베이즈 정리의 적용:

- “광고”라는 단어가 들어간 메일이 스팸일 확률은?

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



07. 텍스트 분석

■ 나이브 베이지안 알고리즘:

- 매우 순진한 가정: 사전 확률을 구성하는 사건은 모두 독립사건이다.
- 만약, 내가 받은 이메일에 다음과 같은 단어들이 포함되었다면?
 - 광고(W_1), 자기야(W_2), 사랑해(W_3), 폭탄 세일(W_4)

$$P(SPAM|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4|SPAM)P(SPAM)}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

$$= \frac{P(W_1|SPAM)P(\neg W_2|SPAM)P(\neg W_3|SPAM)P(W_4|SPAM)P(SPAM)}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

07. 텍스트 분석

■ 나이브 베이지안 분류기:


- 사후 확률을 통해서 해당 사건이 발생할 것인지 여부를 예측
 - 예) 스팸 메일일 확률이 90%, 암 환자일 확률이 98%
- 순진하고, 때로는 잘못된 가정을 했음에도 불구하고 우수한 성능을 보임







07. 텍스트 분석

- Dataset: SMS Spam Collection
 - 원본: UCI M/L Repository
 - <https://archive-beta.ics.uci.edu/ml/datasets/228>

The screenshot shows the dataset page for 'SMS Spam Collection' on the UCI M/L Repository. The header is blue with a database icon, the title 'SMS Spam Collection', and the date 'Donated 2012-06-22'. Below this, it shows '41 views' and '2 citations'. There are 'Download' and 'Cite' buttons. The 'General Information' section is expanded, showing an 'Abstract' which states: 'The SMS Spam Collection is a public set of SMS labeled messages that have been collected for mobile phone spam research.'

 **SMS Spam Collection**
Donated 2012-06-22

 41 views  2 citations

 Download  Cite

General Information [\[edit\]](#)

Abstract
The SMS Spam Collection is a public set of SMS labeled messages that have been collected for mobile phone spam research.

07. 텍스트 분석

■ R에서 스팸 필터 만들기:

1. 데이터 불러오기

```
# read the sms data into the sms data frame
sms_raw <- read.csv("sms_spam_clean.csv",
                    stringsAsFactors = FALSE,
                    encoding = "UTF-8")

str(sms_raw)
sms_raw$type <- as.factor(sms_raw$type)
str(sms_raw$type)
table(sms_raw$type)
```



07. 텍스트 분석

2. 워드 클라우드 보기

```
# word cloud visualization
library(wordcloud)
wordcloud(sms_corpus_clean, min.freq = 50, random.order = FALSE)

spam <- subset(sms_raw, type == "spam")
ham   <- subset(sms_raw, type == "ham")

wordcloud(spam$text, max.words = 40, scale = c(3, 0.5))
wordcloud(ham$text, max.words = 40, scale = c(3, 0.5))
```



07. 텍스트 분석

3. 텍스트 전처리

```
# build a corpus using the text mining (tm) package
```

```
library(tm)
```

```
sms_corpus <- VCorpus(VectorSource(sms_raw$text))
```

```
sms_corpus
```

```
# clean up the corpus using tm_map()
```

```
sms_corpus_clean <- tm_map(sms_corpus, content_transformer(tolower))
```

```
sms_corpus_clean <- tm_map(sms_corpus_clean, removeNumbers) # remove numbers
```

```
sms_corpus_clean <- tm_map(sms_corpus_clean, removeWords, stopwords()) # remove stop words
```

```
sms_corpus_clean <- tm_map(sms_corpus_clean, removePunctuation) # remove punctuation
```


07. 텍스트 분석

```
# illustration of word stemming
library(SnowballC)
wordStem(c("learn", "learned", "learning", "learns"))

sms_corpus_clean <- tm_map(sms_corpus_clean, stemDocument)
sms_corpus_clean <- tm_map(sms_corpus_clean, stripWhitespaces) # eliminate whitespaces

# examine the final clean corpus
lapply(sms_corpus[1:3], as.character)
lapply(sms_corpus_clean[1:3], as.character)

# create a document-term sparse matrix
sms_dtm <- DocumentTermMatrix(sms_corpus_clean)
```



07. 텍스트 분석

```
> lapply(sms_corpus[1:3], as.character)
$`1`
[1] "Hope you are having a good week. Just checking in"
$`2`
[1] "K..give back my thanks."
$`3`
[1] "Am also doing in cbe only. But have to pay."

> lapply(sms_corpus_clean[1:3], as.character)
$`1`
[1] "hope good week just check"
$`2`
[1] "kgive back thank"
$`3`
[1] "also cbe pay"
```



07. 텍스트 분석

4. 훈련용/시험용 데이터 준비

```
# creating training and test datasets
```

```
sms_dtm_train <- sms_dtm[1:4169, ]
```

```
sms_dtm_test  <- sms_dtm[4170:5558, ]
```

```
# also save the labels
```

```
sms_train_labels <- sms_raw[1:4169, ]$type
```

```
sms_test_labels  <- sms_raw[4170:5558, ]$type
```

```
# check that the proportion of spam is similar
```

```
prop.table(table(sms_train_labels))
```

```
prop.table(table(sms_test_labels))
```

```
sms_dtm_freq_train <- removeSparseTerms(sms_dtm_train, 0.999)
```

```
sms_dtm_freq_train
```



07. 텍스트 분석

```
# save frequently-appearing terms to a character vector
sms_freq_words <- findFreqTerms(sms_dtm_train, 5)
str(sms_freq_words)

# create DTMs with only the frequent terms
sms_dtm_freq_train <- sms_dtm_train[ , sms_freq_words]
sms_dtm_freq_test <- sms_dtm_test[ , sms_freq_words]

# convert counts to a factor
convert_counts <- function(x) {
  x <- ifelse(x > 0, "Yes", "No")
}

# apply() convert_counts() to columns of train/test data
sms_train <- apply(sms_dtm_freq_train, MARGIN = 2, convert_counts)
sms_test <- apply(sms_dtm_freq_test, MARGIN = 2, convert_counts)
```



07. 텍스트 분석

5. 나이브 베이지안 분류기 적용

```
library(e1071)
sms_classifier <- naiveBayes(sms_train, sms_train_labels)
sms_test_pred <- predict(sms_classifier, sms_test)
```

6. 혼동 행렬 출력하기

```
library(gmodels)
CrossTable(sms_test_pred, sms_test_labels,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```



07. 텍스트 분석

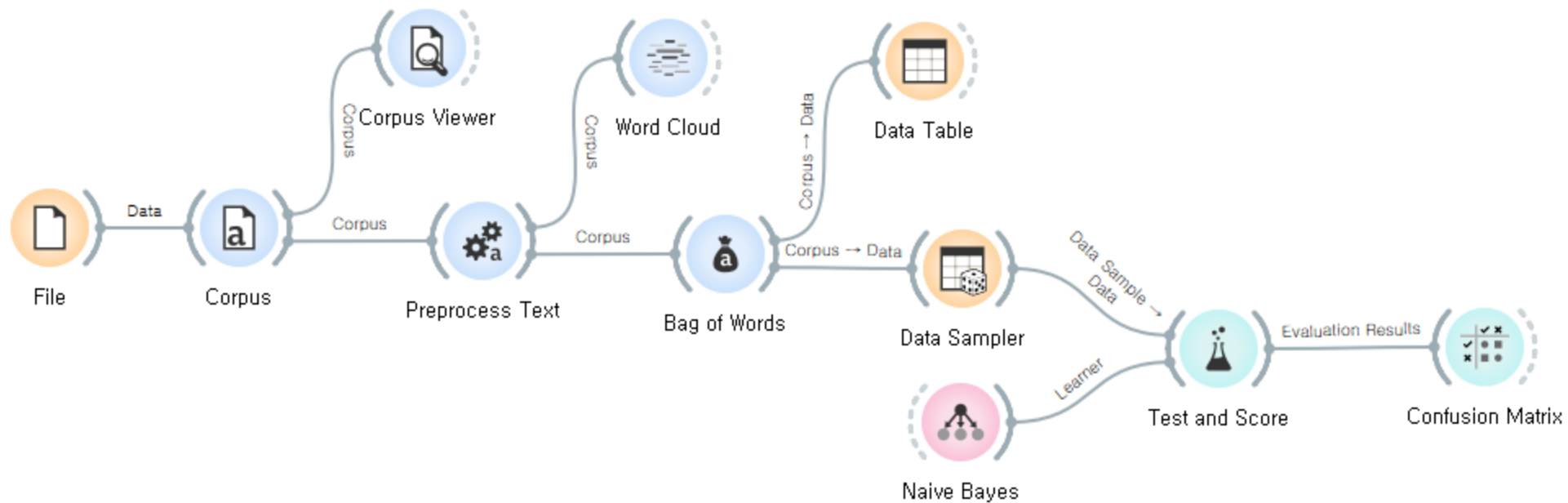
Total Observations in Table: 1389

predicted	actual		Row Total
	ham	spam	
ham	1200 0.864	30 0.022	1230
spam	6 0.004	153 0.110	159
Column Total	1206	183	1389



07. 텍스트 분석

■ Orange에서 스팸 필터 만들기:





07. 텍스트 분석

Test and Score

Sampling

- ☒ Cross validation
 - Number of folds: 10
 - ☐ Stratified
- ☐ Cross validation by feature
- ☐ 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	0.983	0.954	0.955	0.957	0.954

Model Comparison by AUC

	Naive Bayes
Naive Bayes	

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.

612 | - | 612 | 1x612

Confusion Matrix

Learners

Naive Bayes

Show: Number of instances

		Predicted		Σ
		ham	spam	
Actual	ham	527	18	545
	spam	10	57	67
Σ		537	75	612

☒ Predictions ☐ Probabilities

☒ Apply Automatically

Select Correct Select Misclassified Clear Selection

1x612 18 | 612

07. 텍스트 분석

■ 오피니언 마이닝과 감성 분석:

Opinion Mining & Sentiment Analysis

- 문서의 감정을 긍정/중립/부정 등으로 파악하기 위한 방법
- 소셜 미디어, 온라인 리뷰, 영화 댓글 분석 등에 다양하게 활용

Sentiment Analysis



Positive



Negative



Neutral

"I am happy with this water bottle."



Positive

"This is a bad investment."



Negative

"I am going to walk today."



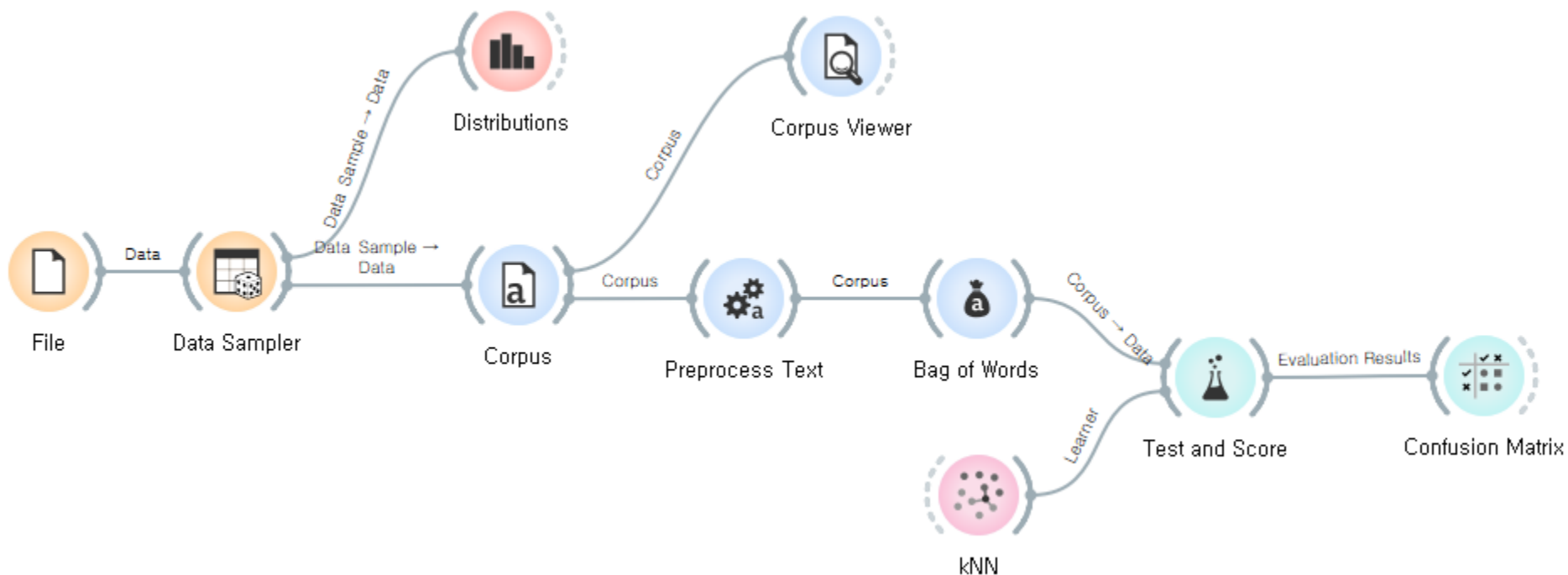
Neutral



07. 텍스트 분석

■ 감성 분석을 위한 지도 학습:

- 학습 데이터에 긍정/부정/중립 등의 라벨이 포함되어 있음
- 기존의 텍스트 기반 분류 알고리즘과 동일한 방법으로 분석 가능함





07. 텍스트 분석

- Dataset: **캐글 IMDB 영화 리뷰 데이터셋**
 - Bag of Words Meets Bags of Popcorn
 - <https://www.kaggle.com/c/word2vec-nlp-tutorial/data>
 - 총 50000개의 IMDB 영화 리뷰에 대한 데이터셋
 - 3개의 변수: id, sentiment(target), review(text)
 - sentiment: 1 for positive, 0 for negative
 - labeledTrainData.tsv



07. 텍스트 분석

File

Source

☒ File: labeledTrainData.tsv ... Reload

☐ URL: ...

Info

25000 instance(s)
2 feature(s) (no missing values)
Data has no target variable.
1 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	id	numeric	feature	
2	sentiment	categorical	target	0, 1
3	review	text	meta	

Reset Apply

Browse documentation datasets

? | 25k

Data Sampler

Sampling Type

☒ Fixed proportion of data:

1 %

☐ Fixed sample size

Instances: 1

☐ Sample with replacement

☐ Cross validation

Number of subsets: 10

Unused subset: 1

☐ Bootstrap

Options

☒ Replicable (deterministic) sampling

☐ Stratify sample (when possible)

Sample Data

? | 25k | 250 | 24.8k

Corpus

Corpus file

election-tweets-2016.tab Browse Reload

Title variable

(no title)

Used text features

review

Ignored text features

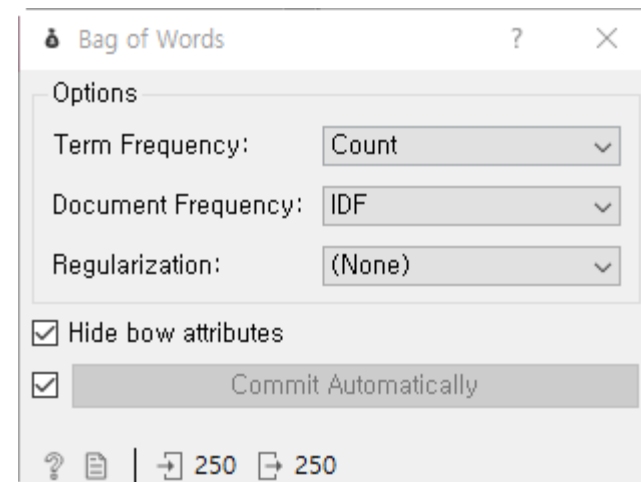
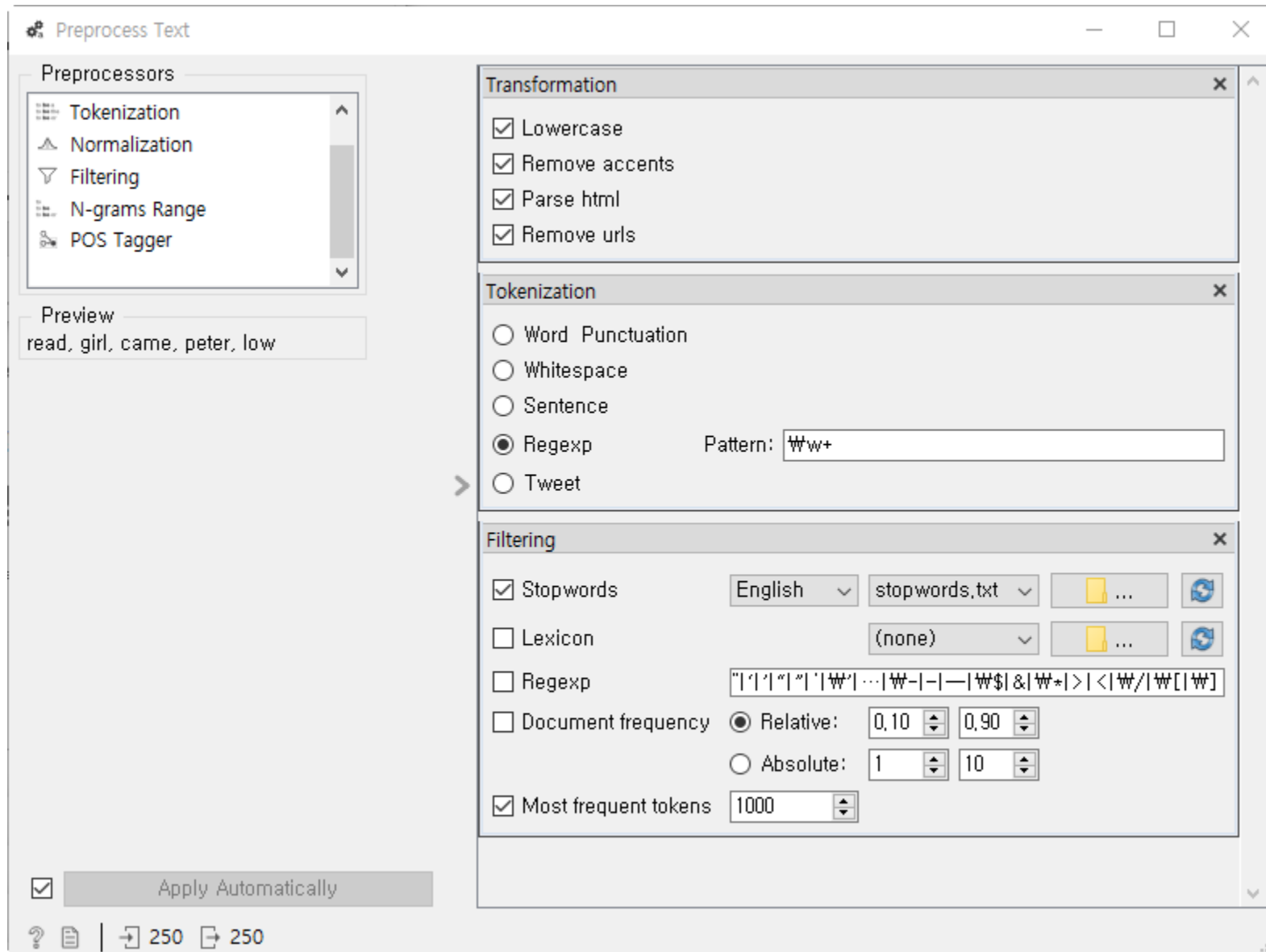
Text

Browse documentation corpora

? | 250 | 250



07. 텍스트 분석





07. 텍스트 분석

Test and Score

Sampling

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

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.709	0.660	0.653	0.665	0.660

Model Comparison by AUC

	kNN
kNN	

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.

250 | - | 250 | 1×250

Confusion Matrix

Learners

kNN

Show: Number of instances

		Predicted		Σ
		0	1	
Actual	0	104	28	132
	1	57	61	118
Σ		161	89	250

☒ Predictions ☐ Probabilities

☒ Apply Automatically

Select Correct Select Misclassified Clear Selection

1×250 104 | 250

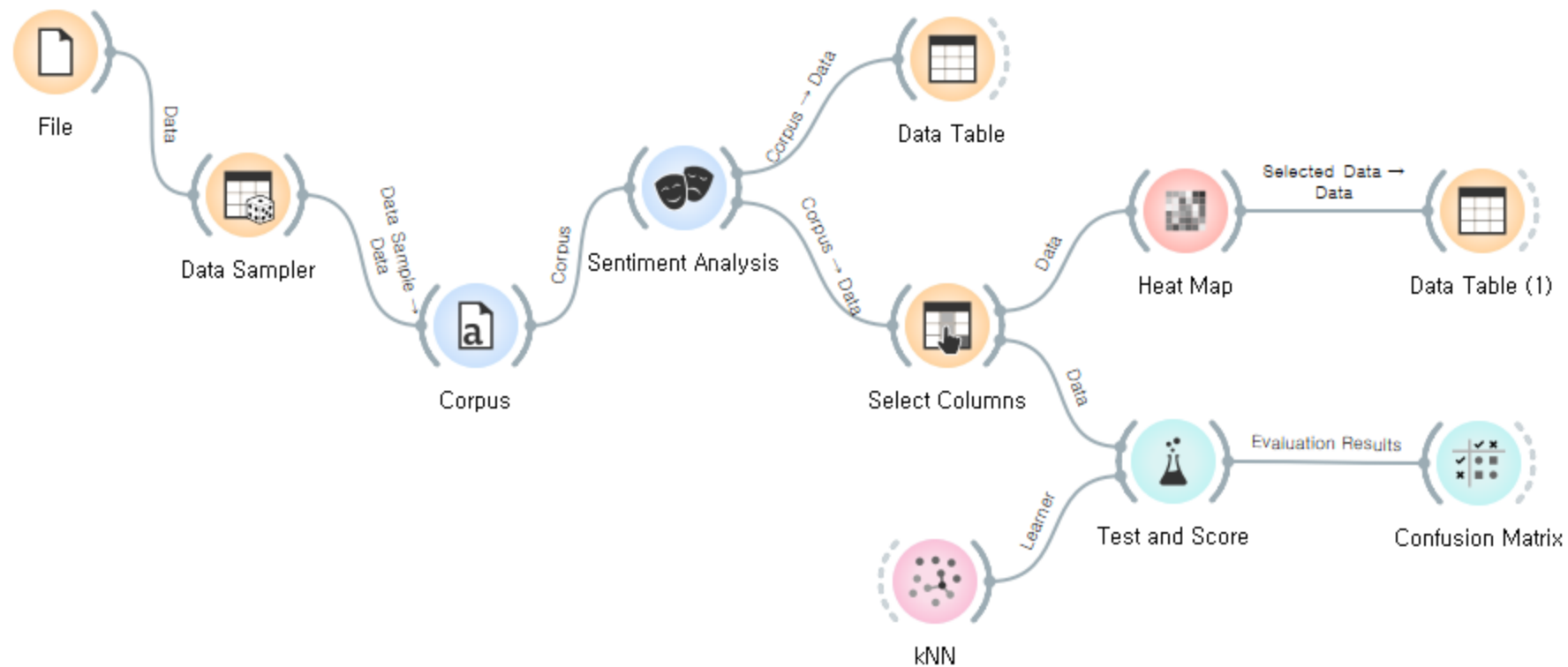


07. 텍스트 분석

- 감성 분석을 위한 비지도 학습:
 - 어휘 사전(*Lexicon*)을 기반으로 텍스트의 감성을 분석
 - 감성 지수(극성 점수): *Polarity Score*
 - 부정(-1)에서 긍정(+1)까지, 감성의 정도를 의미하는 점수
 - 단어의 위치, 주변 단어, 문맥, 품사(POS) 등을 고려해서 점수 부여
 - 감성 분석을 위한 라이브러리:
 - Liu-Hu: NLTK에 포함된 Lexicon 기반의 감성 분석 모듈
 - *VADER*: 소셜 미디어의 텍스트에 특화된 감성 분석 모듈
 - SentiArt: VSM(Vector Space Model) 기반의 감성 분석 모듈

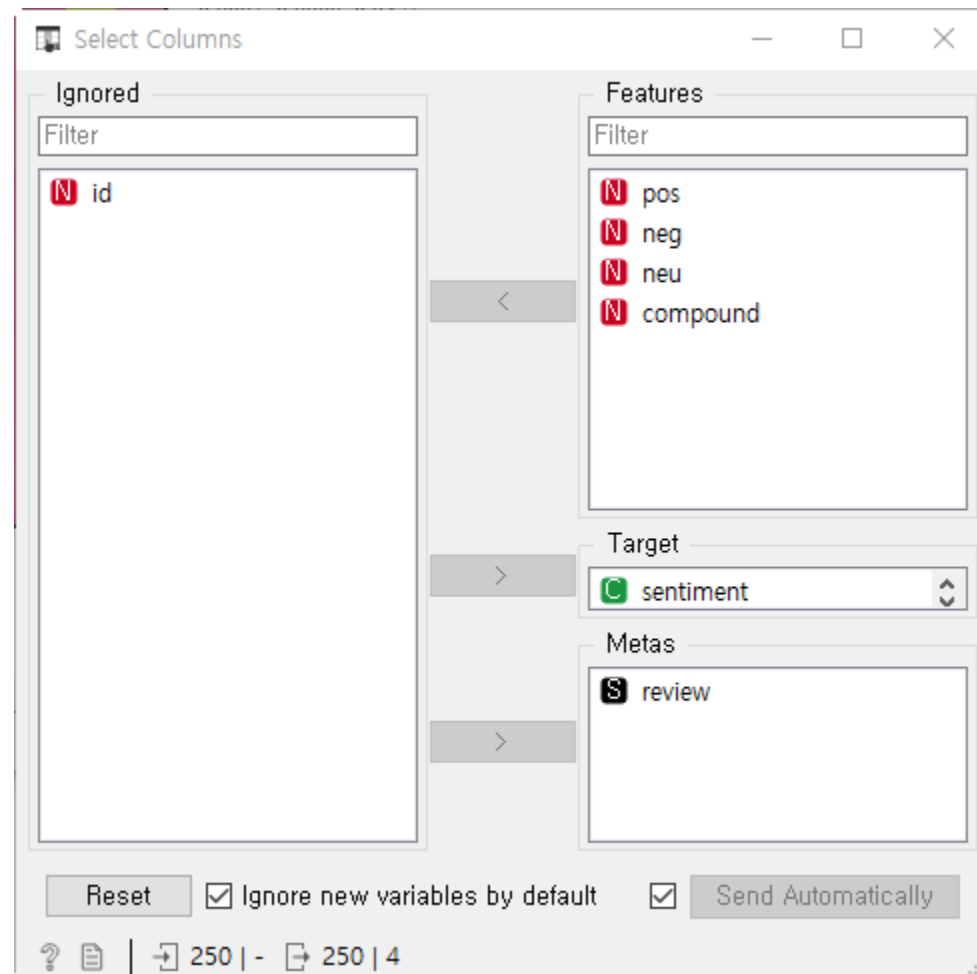
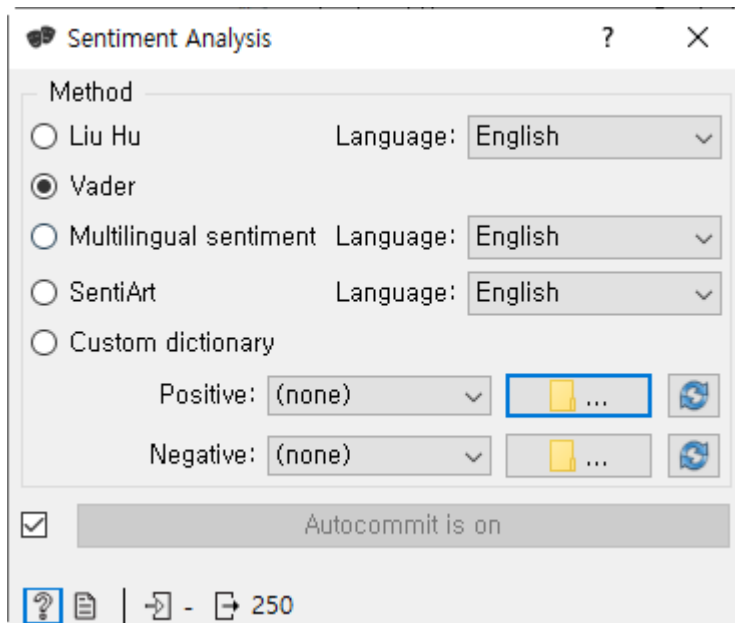


07. 텍스트 분석



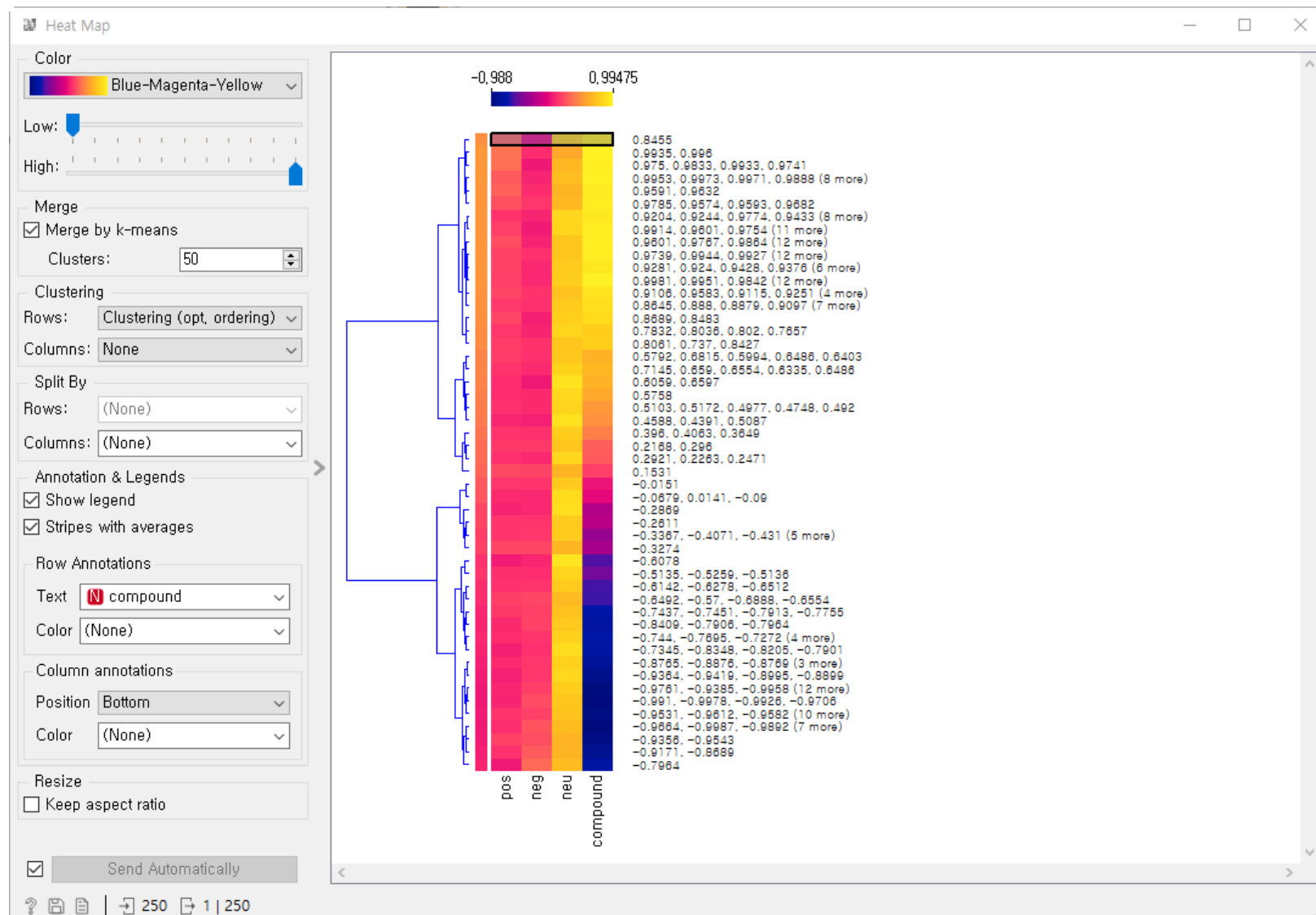


07. 텍스트 분석





07. 텍스트 분석



Any Questions?

