#3. 5장 연합뉴스 타이틀 주제 분류

https://bit.ly/2023어쩌다텍스트분석

어쩌다 한국어 텍스트 분석 with 파이썬

경복고 김선경

이미지 출처 : https://www.sedaily.com/NewsView/29S01GJ0YH

[5장]연합뉴스 타이틀 주제 분류 : 머신러닝을 통한 뉴스 텍스트 분류

https://dacon.io/competitions/official/235747/data

For Example,

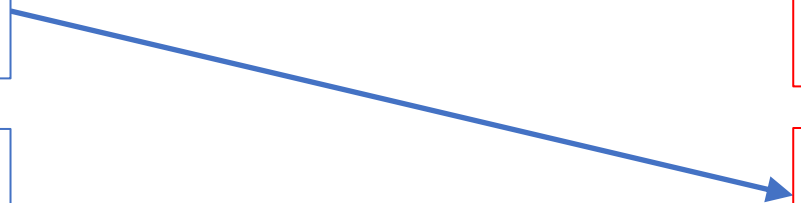
유튜브 내달 2일까지 크리에이터 지원 공간 운영

어버이날 맑다가 흐려져…남부지방 옅은 황사

내년부터 국가RD 평가 때 논문건수는 반영 않는다

김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것

회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간



IT과학

경제

사회

생활문화

세계

스포츠

정치

For Example,

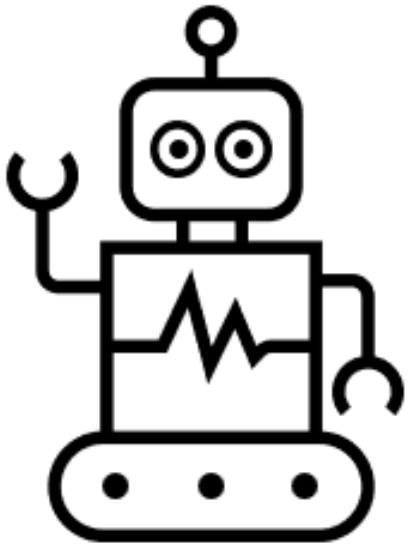
유튜브 내달 2일까지 크리에이터 지원 공간 운영

어버이날 맑다가 흐려져…남부지방 옅은 황사

내년부터 국가RD 평가 때 논문건수는 반영 않는다

김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것

회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간



IT과학

경제

사회

생활문화

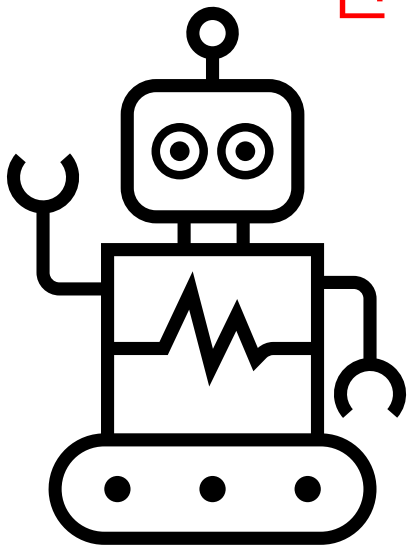
세계

스포츠

정치

그러면, 이 기계는?

단어를 공부해야 한다

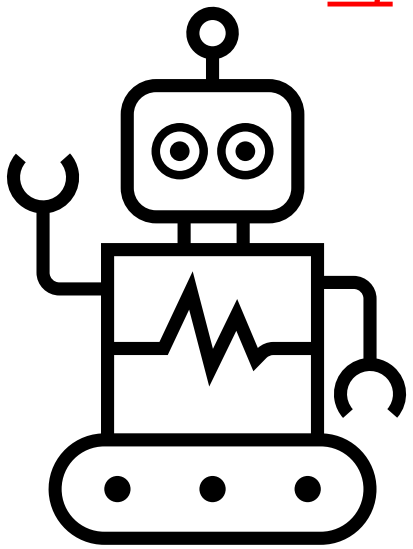


유튜브 내달 2일까지 크리에이터 지원 공간 운영

유튜브	내달	2일까지	크리에이터	지원	공간	운영
유튜브	내달	2일	크리에이터	지원	공간	운영

그러면, 이 기계는?

그런데..난 숫자밖에 몰라.



유튜브 내달 2일까지 크리에이터 지원 공간 운영

유튜브	내달	2일까지	크리에이터	지원	공간	운영
유튜브	내달	2일	크리에이터	지원	공간	운영

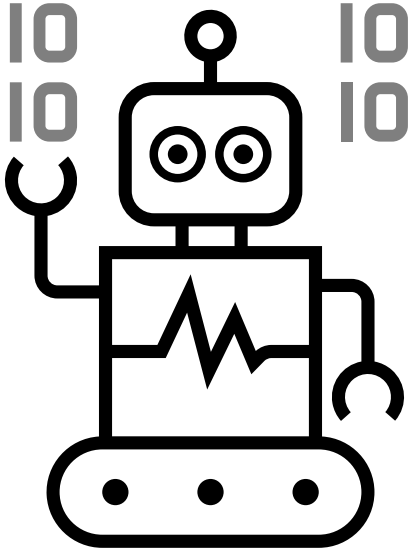
["유튜브", "내달",
"2일", "크리에이터",
"지원", "공간", "운영"]

(1) [1, 0, 1, 0, 0, 1, 1, 0]

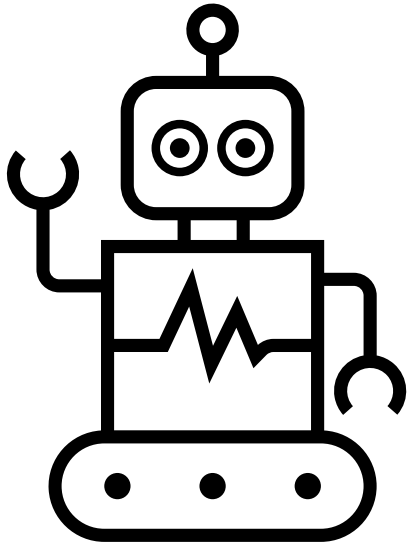
(2) [0, 1, 0, 1, 0, 0, 0, 1]

(3) [0, 1, 0, 0, 1, 0, 0, 1]

그러면, 이 기계는? 공부해라아아아아



얘는 어떻게 공부해?



얘=기계=Machine

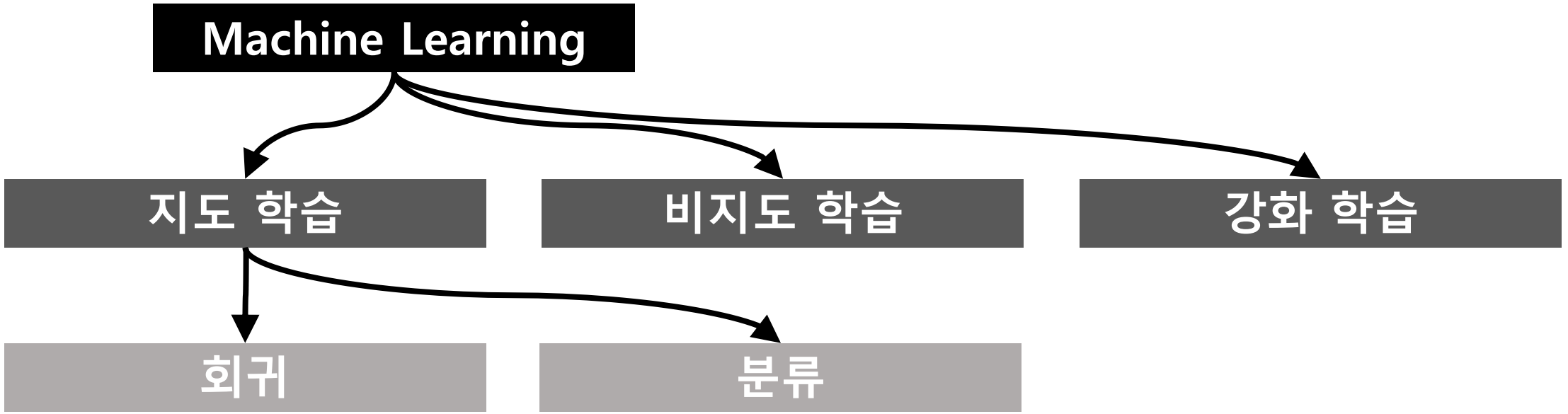
공부=학습=Learning

Machine Learning

어떻게 공부해?

- 지도 학습
- 비지도 학습
- 강화 학습

머신 러닝 지도 학습 모델



Machine Learning Model

선형 회귀

로지스틱 회귀

랜덤 포레스트

lightGBM

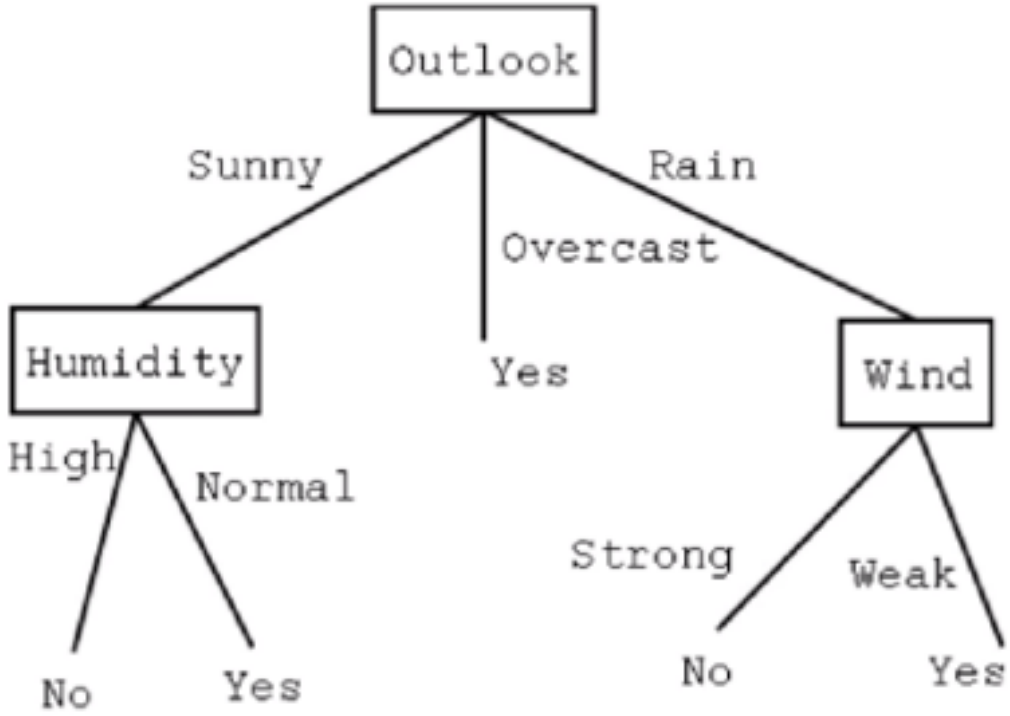
인공 신경망

k-NN

오늘의 모델 : 랜덤포레스트

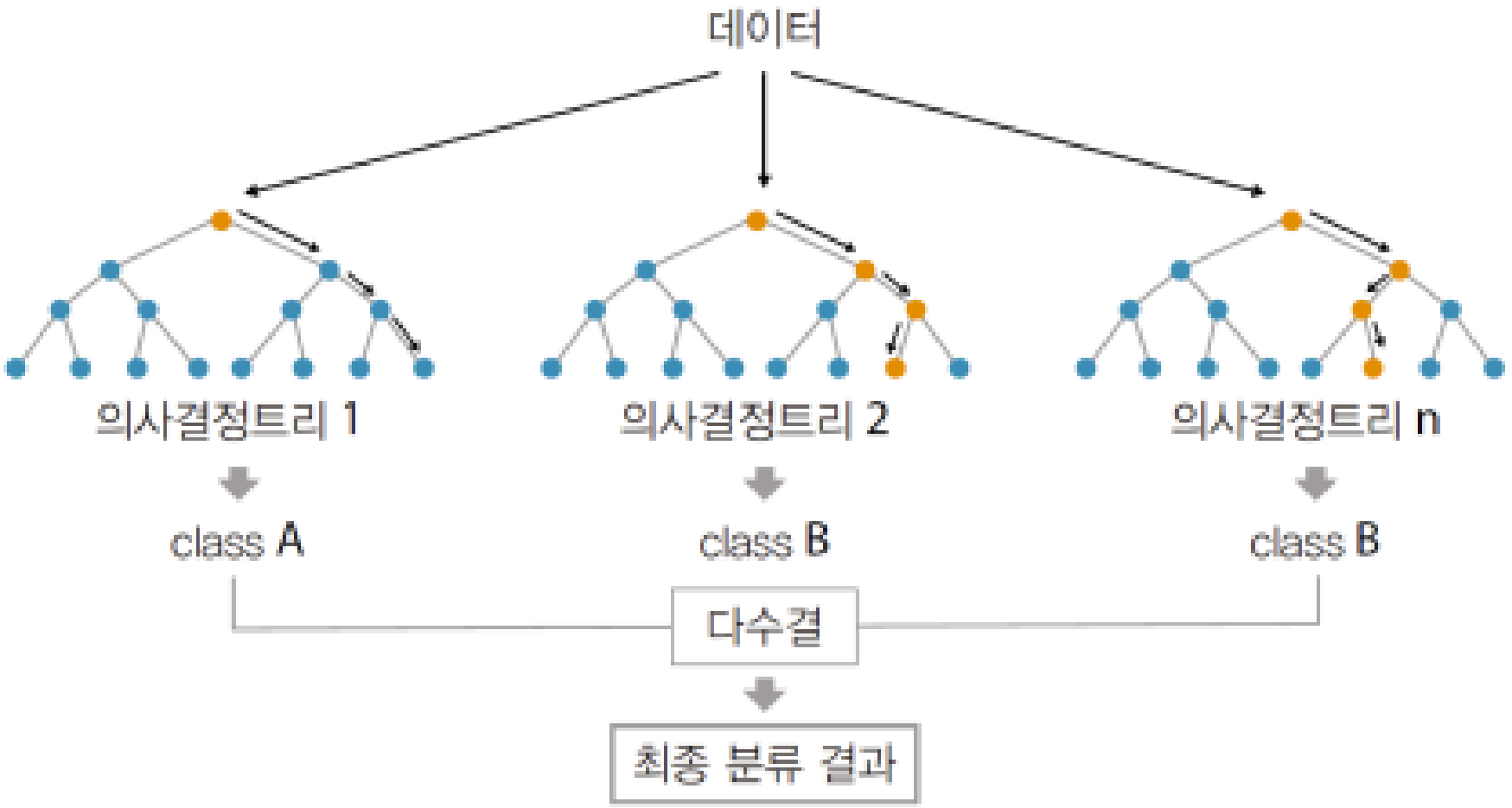
숲이 되려면 나무가 먼저 필요하다 : Decision Tree(의사 결정 트리)

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



오늘의 모델 : 랜덤포레스트

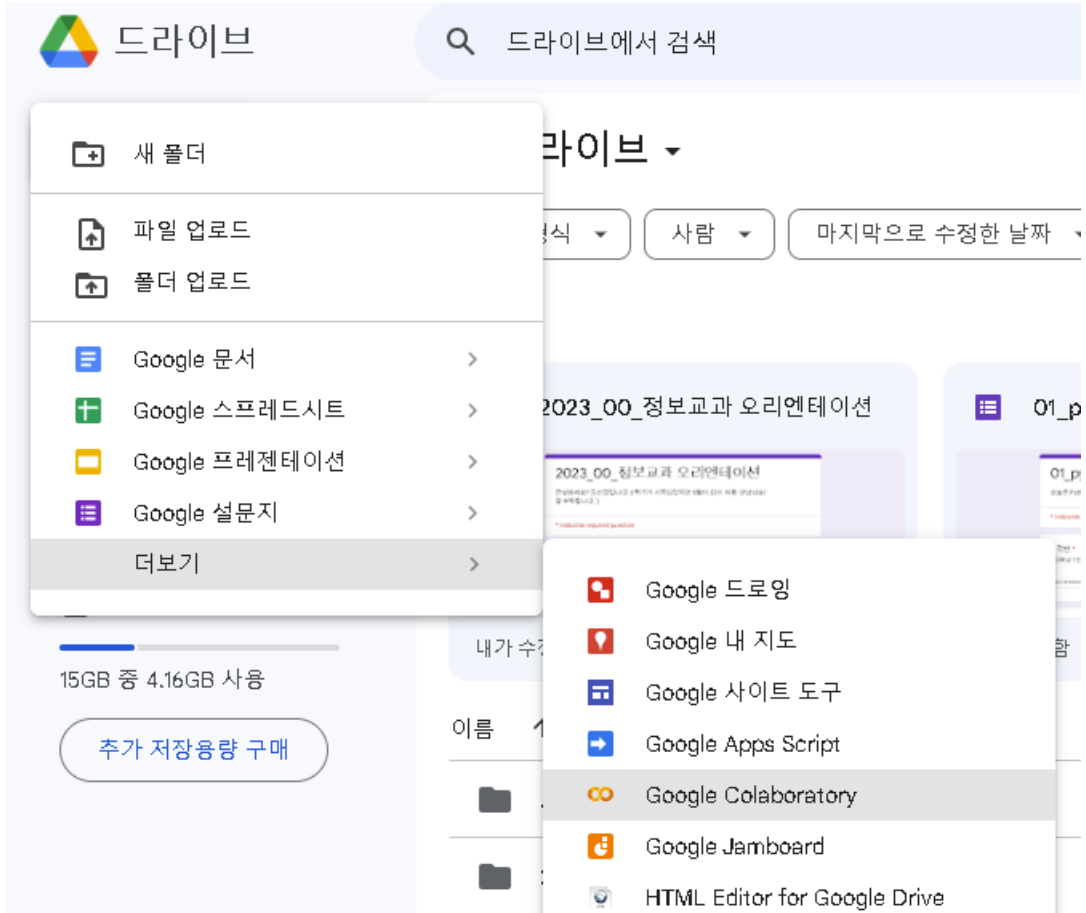
나무를 여러 개 심으면 숲 : Random Forest(랜덤 포레스트)



구글 코랩 실행하기

1. 구글 드라이브 드라이브

2. 새로 만들기 – Google Colaboratory



GitHub의 [참고code] 링크 클릭하여 내 구글드라이브로 가져가세요!

Home

swKyungbock edited this page now · 16 revisions

Welcome to the <2023 ML with Korea Text Data>!

(2023.08.13/08.15)

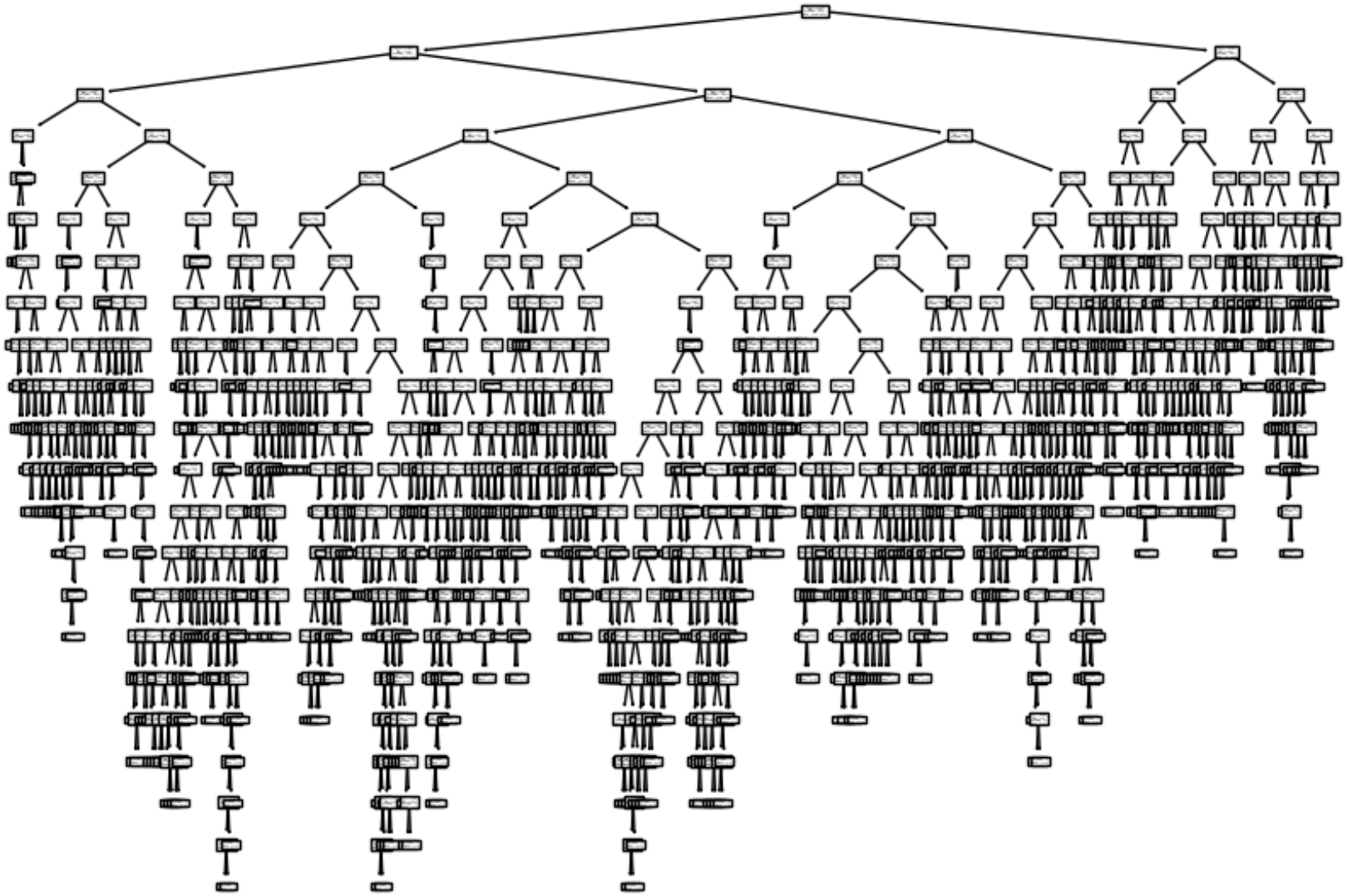
수업목차

8월 13일(1일차)

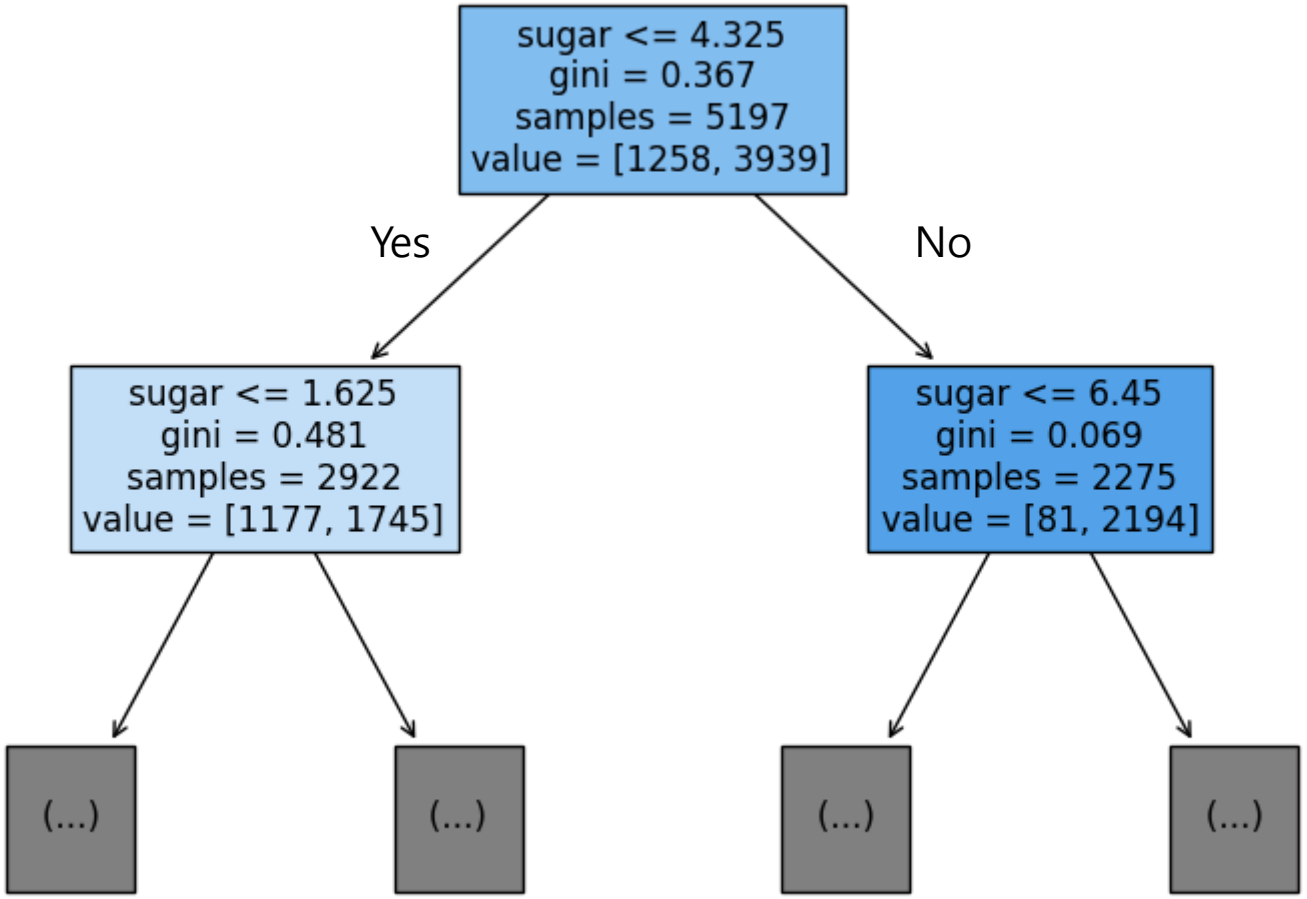
1. 오리엔테이션-어쩌다 한국어 텍스트 분석 [ppt]
2. 오늘,우리에게 필요한 머신러닝 살펴보기1 [ppt]
3. 5장 연합뉴스 타이틀 주제 분류 [ppt],[참고code][5장code]
4. 6장 국민청원 데이터 시각화와 분류 [ppt],[code]

[참고] 레드와인과 화이트 와인을 결정트리를 이용하여 분류하기

```
dt=DecisionTreeClassifier(random_state=42)
```



[참고] 레드와인과 화이트 와인을 결정트리를 이용하여 분류하기

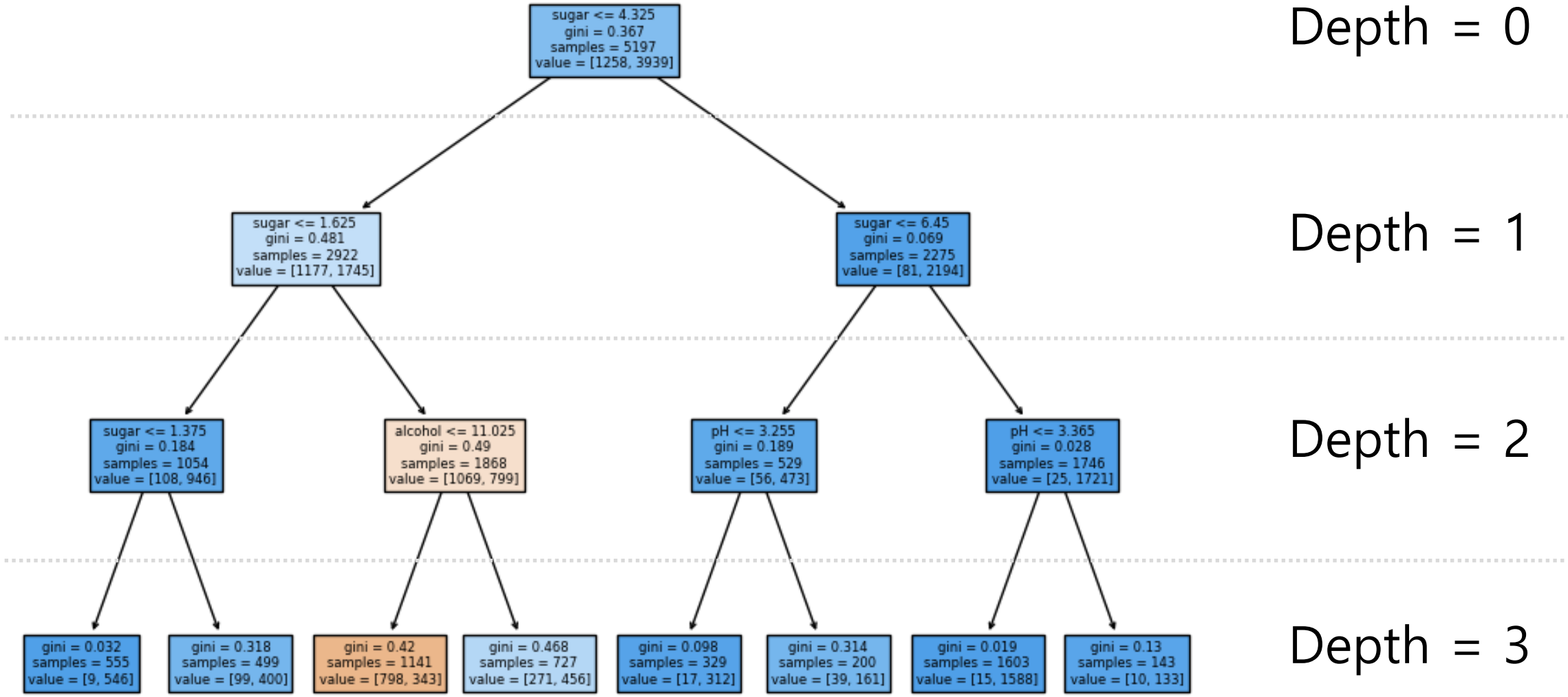


Depth = 0

Depth = 1

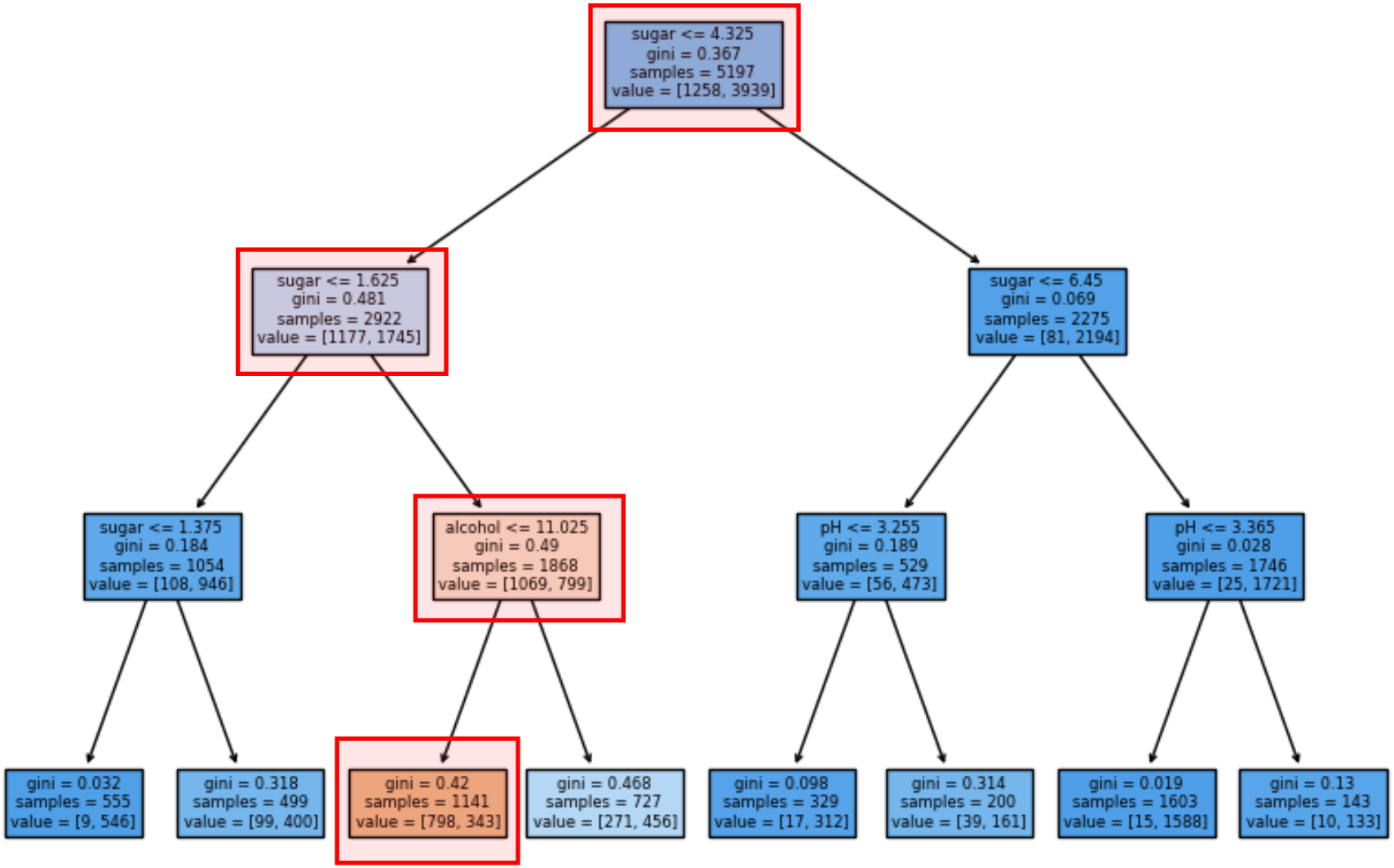
[참고] 레드와인과 화이트 와인을 결정트리를 이용하여 분류하기

```
dt2=DecisionTreeClassifier(max_depth=3, random_state=42)
```



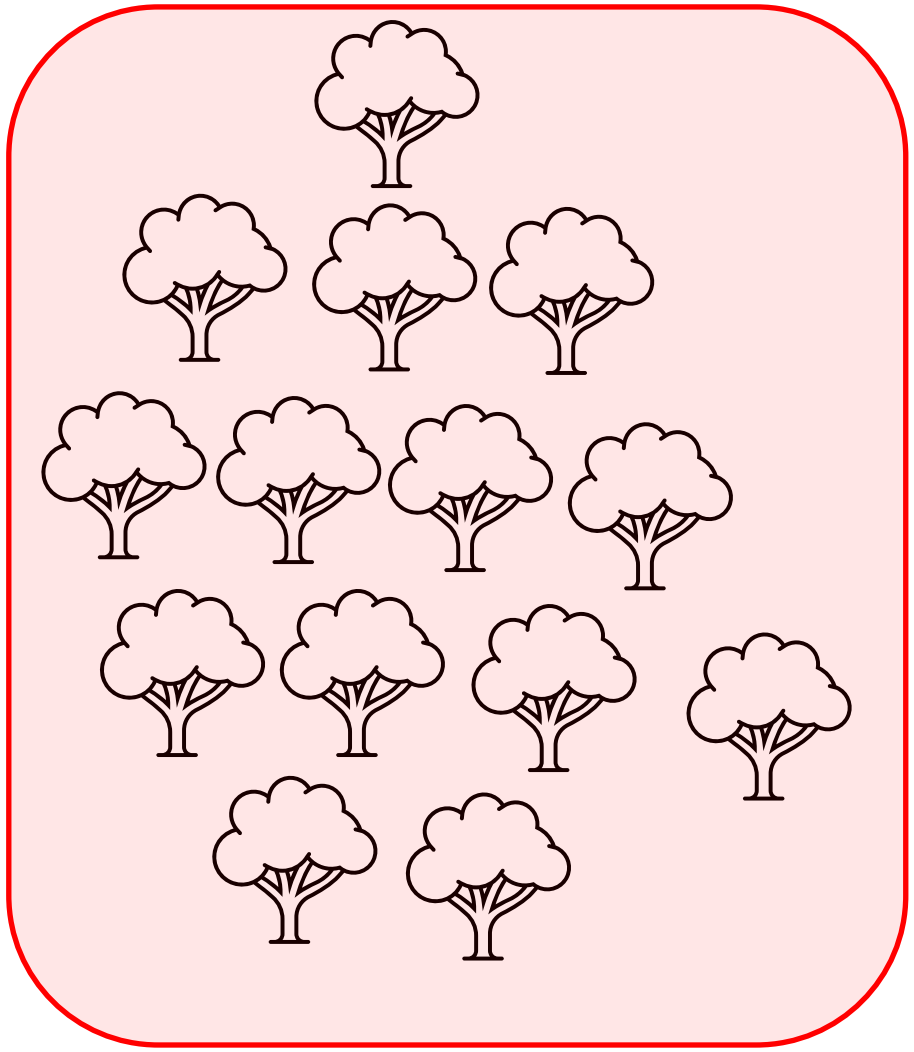
[참고] 레드와인과 화이트 와인을 결정트리를 이용하여 분류하기

레드 와인의 조건 : 1.625< sugar <=4.325 이고, alcohol <= 11.025



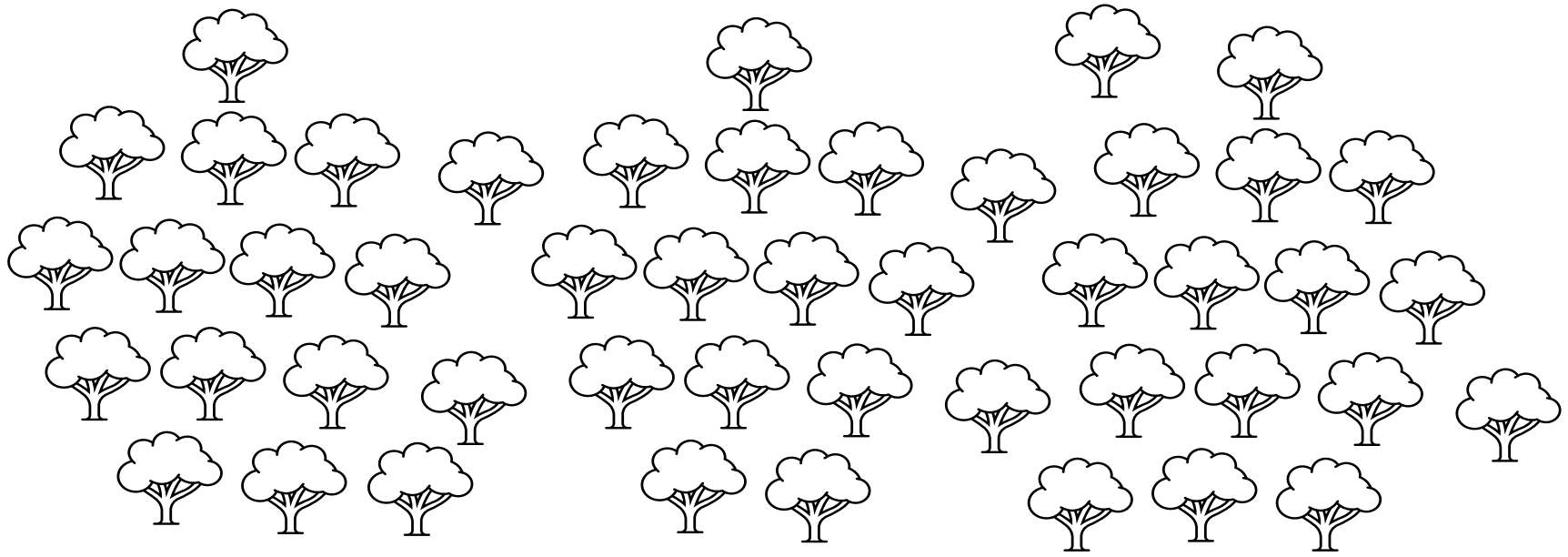
[참고] 레드와인과 화이트 와인을 랜덤포레스트를 이용하여 분류하기

결정트리



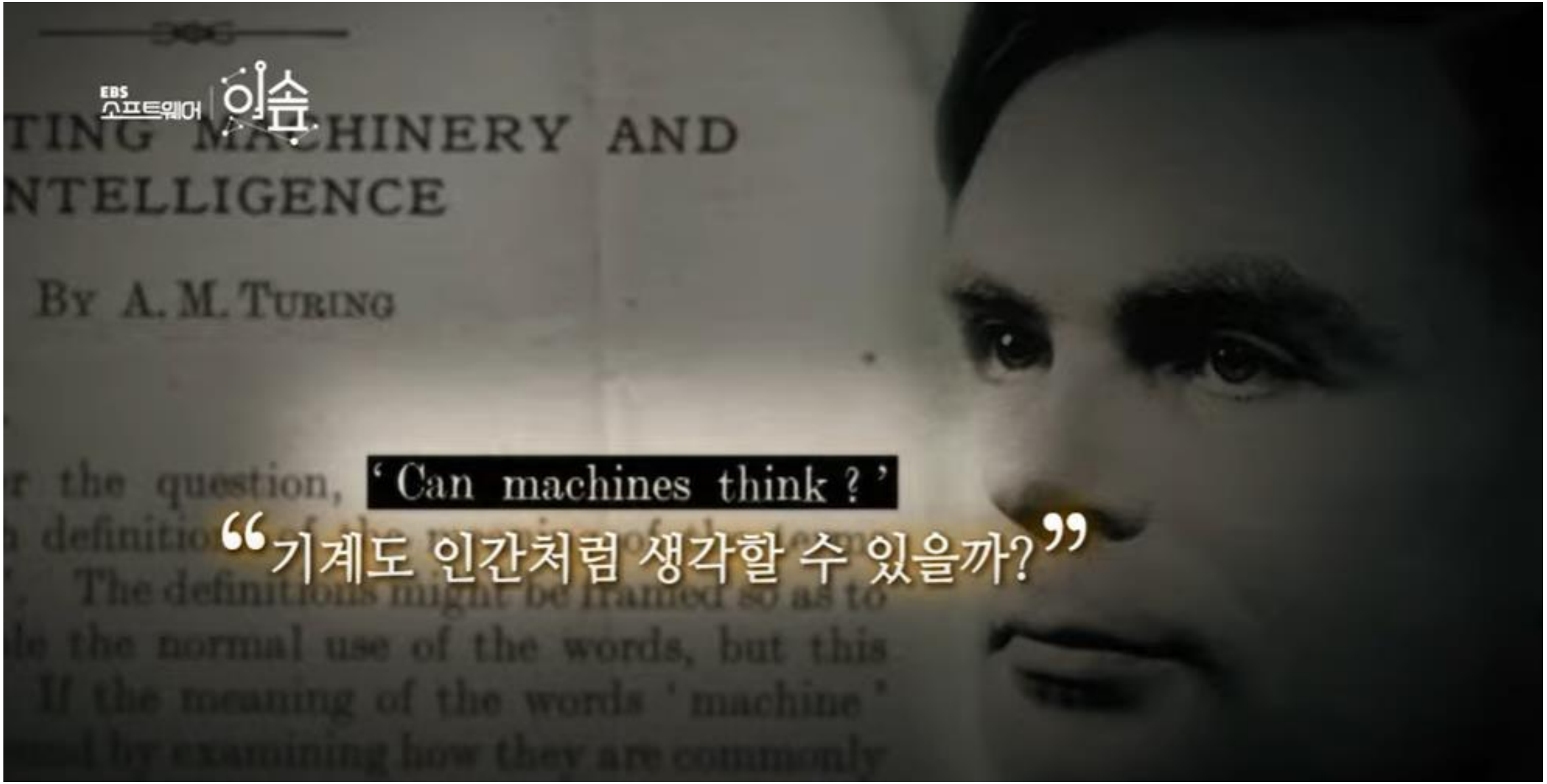
랜덤포레스트

[참고] 레드와인과 화이트 와인을 랜덤포레스트를 이용하여 분류하기



```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier(n_estimators=100, n_jobs=-1, random_state=42)
rf.fit(train_input, train_target)

print(rf.score(train_input, train_target))
print(rf.score(test_input, test_target))
print(rf.feature_importances_)
```

결정 트리도 알겠고, 랜덤 포레스트도 알겠는데 왜 이렇는 걸까요?

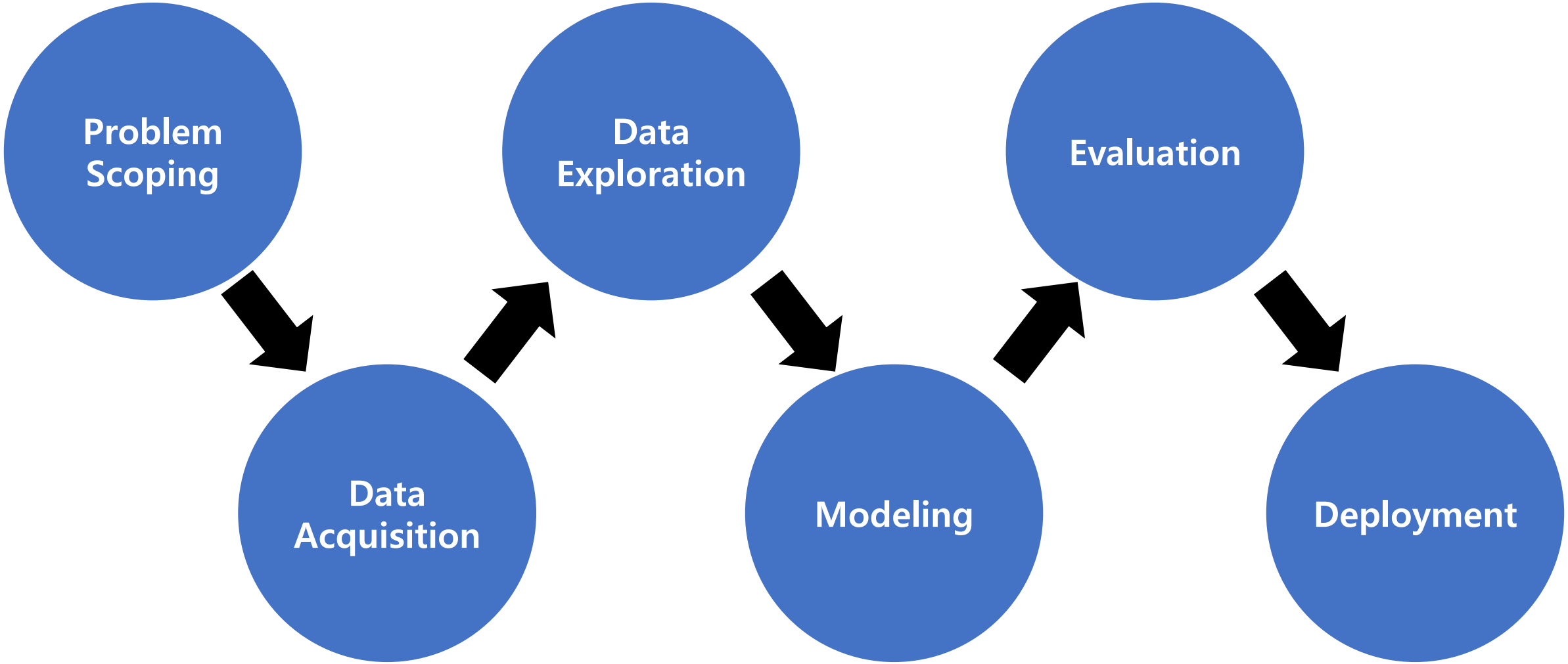

EBS 소프트웨어 이슈
TING MACHINERY AND
NTELLIGENCE
By A. M. Turing
'Can machines think?'
"기계도 인간처럼 생각할 수 있을까?"

연합뉴스 타이틀 주제 분류 : 머신러닝을 통한 뉴스 텍스트 분류

모델 : 랜덤 포레스트

https://dacon.io/competitions/official/235747/data

AI 프로젝트 작업 흐름도



AI 프로젝트 작업 흐름도 : Problem Scoping

대회안내 | 데이터 | 코드 공유 | 토크 | 리더보드 | 제출

- 개요
- 규칙
- 일정
- 상금
- 동의사항

1.배경

안녕하세요 여러분! 🙌 **뉴스 토픽 분류 AI 경진대회**에 오신 것을 환영합니다.

텍스트 주제를 추론하는 것은 언어 이해 시스템이 보유해야 하는 핵심 기능입니다.

YNAT(주제 분류를 위한 연합 뉴스 헤드라인) 데이터 세트를 활용해 주제 분류 알고리즘을 개발해 주세요.

국내 최초 오픈 데이터 세트인 KLUE(Korean Language Understanding Evaluation) 데이터 세트를 이용하여 다양한 언어 모델의 성능을 비교해 한국어 자연어처리 분야의 발전에 기여할 것으로 예상합니다.

2. 목적

AI 프로젝트 작업 흐름도 : Data Acquisition

대회안내 | 데이터 | 코드 공유 | 토크 | 리더보드 | 제출

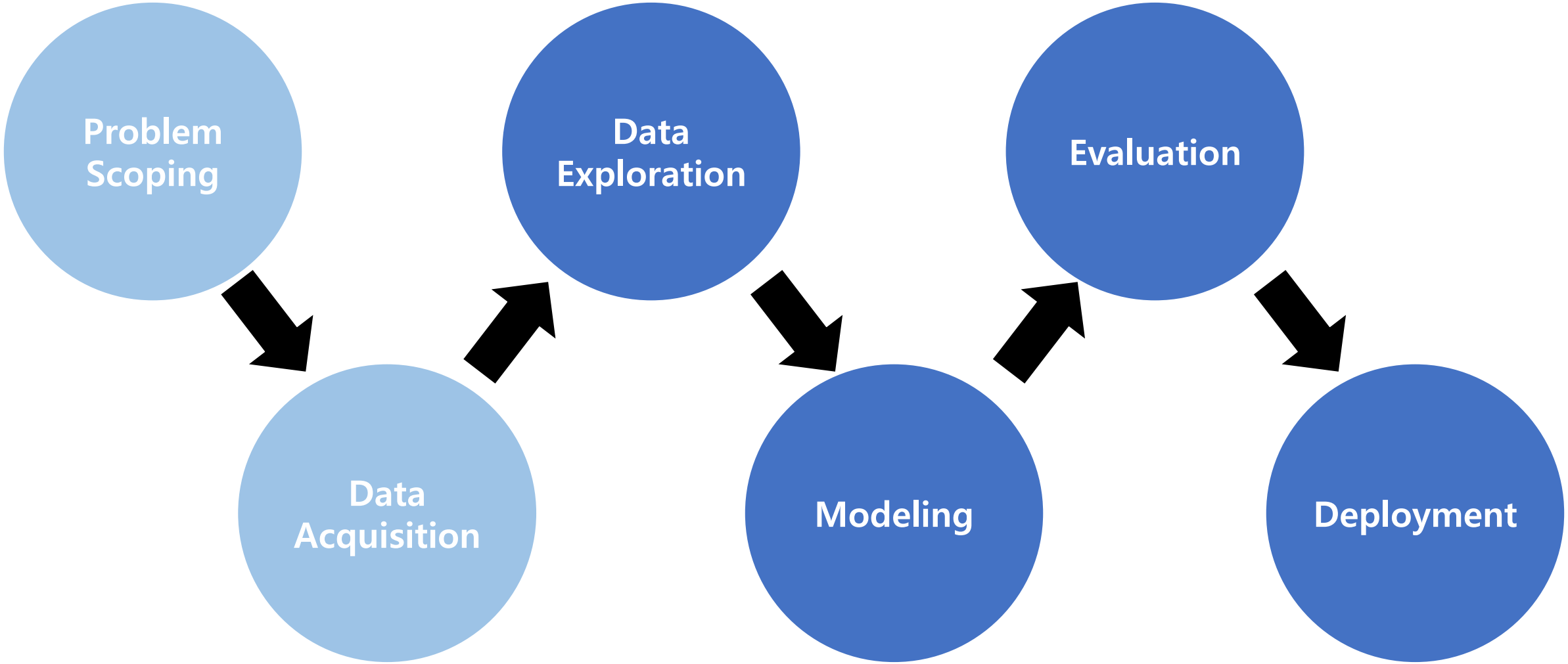
설명

train_data.csv

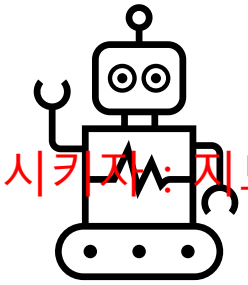
- index : 헤드라인 인덱스
- title : 뉴스 헤드라인
- topic_idx : 뉴스 주제 인덱스 값(label)

다운로드

AI 프로젝트 작업 흐름도



연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



얘를 공부 시키자: 지도 학습으로

랜덤포레스트

데이터 로드

데이터 전처리

데이터 나누기
Train-test-split

텍스트 데이터 벡터화
BOW, TF-IDF

학습 데이터 세트
(x_train, y_train)

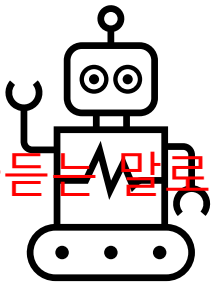
시험 데이터 세트
(x_test, y_test)

모델

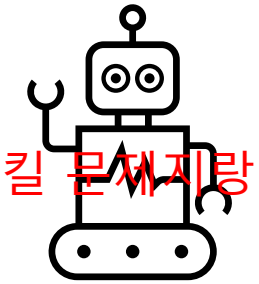
학습
fit(x_train, y_train)

예측
predict(x_test)

평가
evaluate(y_test, y-predict)



얘가 알아듣는 말로 바꿔준다



얘를 공부 시킬 문제지랑 시험지 준비

GitHub의 [5장code] 링크 클릭하여 내 구글드라이브로 가져가세요!

Home

swKyungbock edited this page now · 16 revisions

Welcome to the <2023 ML with Korea Text Data>!

(2023.08.13/08.15)

수업목차

8월 13일(1일차)

1. 오리엔테이션-어쩌다 한국어 텍스트 분석 [ppt]
2. 오늘,우리에게 필요한 머신러닝 살펴보기1 [ppt]
3. 5장 연합뉴스 타이틀 주제 분류 [ppt],[참고code][5장code]
4. 6장 국민청원 데이터 시각화와 분류 [ppt],[code]

기본 설정 : 필요한 라이브러리

1. 판다스(pandas)

데이터 다루기

2. 넘파이(numpy)

3. 맷플롯립(matplotlib)

데이터 시각화

4. 시본(seaborn)

기본 설정 : 한글 문제를 위한 설치

1. 시각화(맷플롯립 그래프)를 위한 한글 설치

   pip install koreanize-matplotlib

2. 한국어 정보 처리를 위한 파이썬 패키지 설치

   pip install konlpy --upgrade

연합뉴스 타이틀 주제 분류 : 데이터 로드

https://dacon.io/competitions/official/235747/data

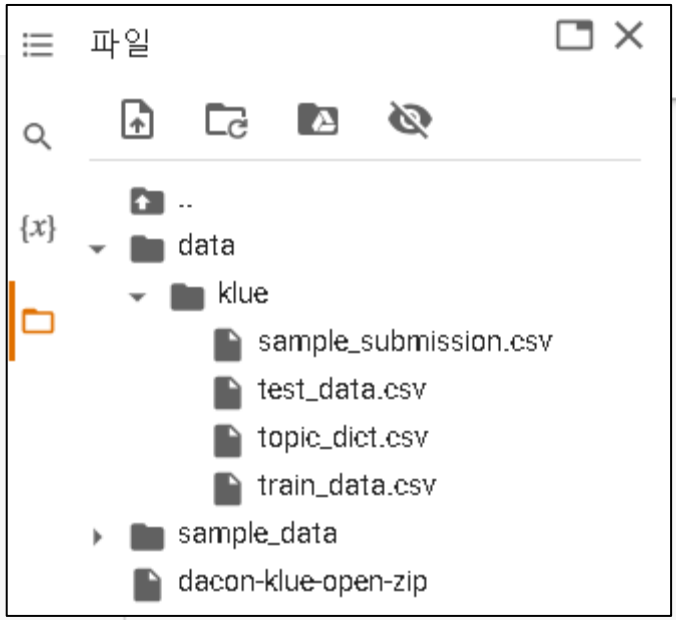
설명

train_data.csv

- index : 헤드라인 인덱스
- title : 뉴스 헤드라인
- topic_idx : 뉴스 주제 인덱스 값(label)

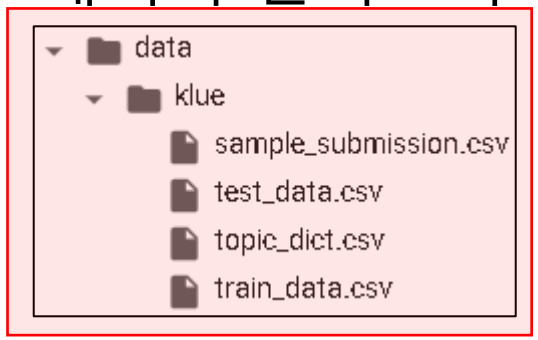
연합뉴스 타이틀 주제 분류 : 데이터 로드

```python
import os
import platform

base_path="data/klue/"
file_name="dacon-klue-open-zip"

#파일이 있는지 확인하는 함수 작성
def file_exist_check(base_path) :
  #파일 경로 만들기
  if os.path.exists(f'{base_path}train_data.csv') : #해당 경로에 파일이 있다면
    print(f'{os.getcwd()}/{base_path} 경로에 파일이 있음')
    return
  if not os.path.exists(f'{base_path}train_data.csv') : #해당 경로에 파일이 없다면
    os.makedirs(base_path)
  if platform.system()=="Linux" :
    print(f'파일을 다운로드 하고 {base_path} 경로에 압축을 해제함')
    !wget https://bit.ly/{file_name}
    !unzip {file_name} -d {base_path}
    return
  else :
    print(f"""https://dacon.io/competitions/official/235747/data 에서 다운로드 해 실습 경로 {os.getcwd()}/{base_path}에 옮겨주세요.""")
    return

file_exist_check(base_path)
```



연합뉴스 타이틀 주제 분류 : 데이터 로드

데이터 살펴보기



index	title
45654	유튜브 내달 2일까지 크리에이터 지원 공간 운영
45655	어버이날 맑다가 흐려져…남부지방 옅은 황사
45656	내년부터 국가RD 평가 때 논문건수는 반영 않는다
45657	김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것
45658	회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간
45659	야외서 생방송 하세요…액션캠 전용 요금제 잇따라
45660	월드컵 태극전사 16강 전초기지 레오강 입성종합
45661	미세먼지 속 출근길
45662	왓츠앱稅 230원에 성난 레바논 민심…총리사퇴로 이어져종합2보
45663	베트남 경제 고성장 지속…2분기 GDP 6.71% 성장
45664	그리스서 한국전 참전 기념식…참전용사 한반도 평화 기원

topic	topic_idx
IT과학	0
경제	1
사회	2
생활문화	3
세계	4
스포츠	5
정치	6

index	title	topic_idx
0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4
1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4
2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4
3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4
4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4
5	팔레스타인 가자지구서 16세 소년 이스라엘군 총격에 사망	4
6	인도 48년 만에 파키스탄 공습…테러 캠프 폭격종합2보	4
7	美대선 TV토론 음담패설 만회실패 트럼프…사과 대신 빌클린턴 공격해 역효과	4
8	푸틴 한반도 상황 진전 위한 방안 김정은 위원장과 논의	4
9	특검 면죄부 받은 트럼프 스캔들 보도 언론 맹공…국민의 적	4
10	日 오키나와서 열린 강제징용 노동자 추도식	4

연합뉴스 타이틀 주제 분류 : 데이터 로드

지도학습

Test Data

index	title
45654	유튜브 내달 2일까지 크리에이터 지원 공간 운영
45655	어버이날 맑다가 흐려져…남부지방 옅은 황사
45656	내년부터 국가R&D 평가 때 논문건수는 반영 않는다
45657	김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것
45658	회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간
45659	야외서 생방송 하세요…액션캠 전용 요금제 잇따라
45660	월드컵 태극전사 16강 전초기지 레오강 입성종합
45661	미세먼지 속 출근길
45662	왓츠앱稅 230원에 성난 레바논 민심…총리사퇴로 이어져종합2보
45663	베트남 경제 고성장 지속…2분기 GDP 6.71% 성장
45664	그리스서 한국전 참전 기념식…참전용사 한반도 평화 기원

topic	topic_idx
IT과학	0
경제	1
사회	2
생활문화	3
세계	4
스포츠	5
정치	6

Training Data

index	title	topic_idx
0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4
1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4
2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4
3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4
4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4
5	팔레스타인 가자지구서 16세 소년 이스라엘군 총격에 사망	4
6	인도 48년 만에 파키스탄 공습…테러 캠프 폭격종합2보	4
7	美대선 TV토론 음담패설 만회실패 트럼프…사과 대신 빌클린턴 공격해 역효과	4
8	푸틴 한반도 상황 진전 위한 방안 김정은 위원장과 논의	4
9	특검 면죄부 받은 트럼프 스캔들 보도 언론 맹공…국민의 적	4
10	日 오키나와서 열린 강제징용 노동자 추도식	4

연합뉴스 타이틀 주제 분류 : 데이터 로드

```
train=pd.read_csv(os.path.join(base_path, "train_data.csv"))
test=pd.read_csv(os.path.join(base_path,"test_data.csv"))
```

train변수
(45654, 3)

```
1 train.head()
```

	index	title	topic_idx
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4

test변수
(9131, 2)

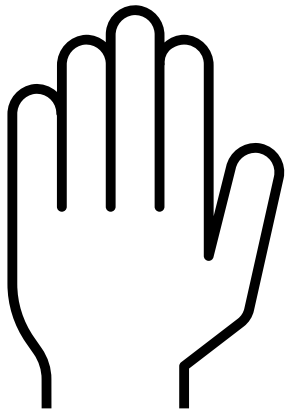
```
1 test.head()
```

	index	title
0	45654	유튜브 내달 2일까지 크리에이터 지원 공간 운영
1	45655	어버이날 맑다가 흐려져…남부지방 옅은 황사
2	45656	내년부터 국가RD 평가 때 논문건수는 반영 않는다
3	45657	김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것
4	45658	회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간

연합뉴스 타이틀 주제 분류 : 데이터 전처리

1. 전체적인 데이터 특성 파악 : 문자 길이, 단어의 등장 빈도 확인(시각화)

2. 불필요한 기호를 제거 후 중요 내용 추출



잠깐!

머신러닝에서 알고리즘이나 좋은 컴퓨터 환경만큼 중요한 것이
바로 “제대로 된 데이터를 준비하는 일” 입니다.

그래서, 판다스와 넘파이, 데이터 시각화를 통해
다각적으로 데이터를 관찰하고 효율적으로 다루는 연습을 하는 것이
중요합니다!!!

연합뉴스 타이틀 주제 분류 : 데이터 전처리(전체적인 데이터 파악 中)

- 트레이닝용 데이터와 테스트 데이터 모두 전처리를 해 줘야 하므로 병합해서 처리한 뒤, 나중에 다시 분류

```
raw=pd.concat([train,test])
print(raw.shape)
raw.head()
raw.tail()
```

(54785, 3)

	index	title	topic_idx
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4.0
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0

	index	title	topic_idx
9126	54780	인천 오후 3시35분 대설주의보…눈 3.1cm 쌓여	NaN
9127	54781	노래방에서 지인 성추행 외교부 사무관 불구속 입건종합	NaN
9128	54782	40년 전 부마항쟁 부산 시위 사진 2점 최초 공개	NaN
9129	54783	게시판 아리랑TV 아프리카개발은행 총회 개회식 생중계	NaN
9130	54784	유영민 과기장관 강소특구는 지역 혁신의 중심…지원책 강구	NaN

연합뉴스 타이틀 주제 분류 : 데이터 전처리(전체적인 데이터 파악 中)

- 토픽 이름 찾아주기

raw. shape : (54785, 3)

	index	title	topic_idx
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4.0
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0

```
df=raw.merge(topic, how="left")
```

df. shape : (54785, 4)

	index	title	topic_idx	topic
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0	세계
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4.0	세계
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0	세계
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0	세계
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0	세계

연합뉴스 타이틀 주제 분류 : 데이터 전처리(전체적인 데이터 파악 中)

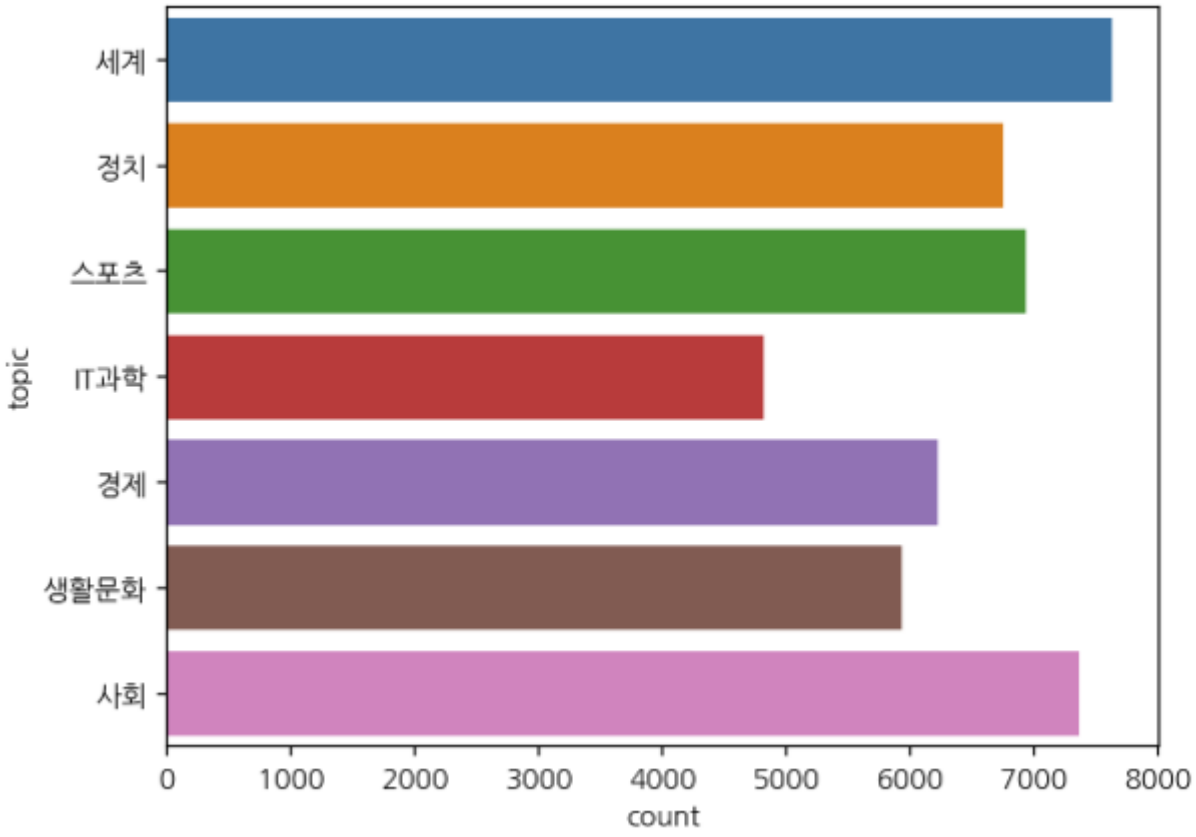
- 정답값에 대한 빈도수 확인

```
df["topic_idx"].value_counts()
```

4.0	7629
2.0	7362
5.0	6933
6.0	6751
1.0	6222
3.0	5933
0.0	4824

시각화

```
sns.countplot(data=df, y="topic")
```



연합뉴스 타이틀 주제 분류 : 데이터 전처리(전체적인 데이터 파악 中)

- 문자의 길이 확인 : 긴 텍스트(의미 파악 Good)

```
df["len"]=df["title"].apply(lambda x : len(x))
df["word_count"]=df["title"].apply(lambda x : len(x.split()))
df["unique_word_count"]=df["title"].apply(lambda x : len(set(x.split())))
```

	index	title	topic_idx	topic	len	word_count	unique_word_count
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0	세계	24	5	5
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4.0	세계	30	6	6
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0	세계	30	8	8
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0	세계	32	7	7
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0	세계	25	7	7

시각화

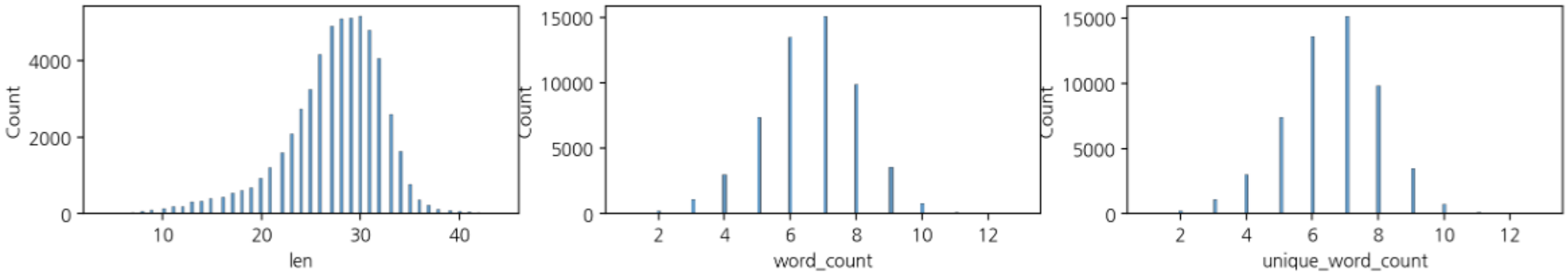
```
fig, axes=plt.subplots(1,3,figsize=(15,2))
sns.histplot(df["len"], ax=axes[0])
sns.histplot(df["word_count"],ax=axes[1])
sns.histplot(df["unique_word_count"],ax=axes[2])
```

연합뉴스 타이틀 주제 분류 : 데이터 전처리(전체적인 데이터 파악 中)

- 문자의 길이 확인

	index	title	topic_idx	topic	len	word_count	unique_word_count
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0	세계	24	5	5
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4.0	세계	30	6	6
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0	세계	30	8	8
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0	세계	32	7	7
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0	세계	25	7	7

시각화

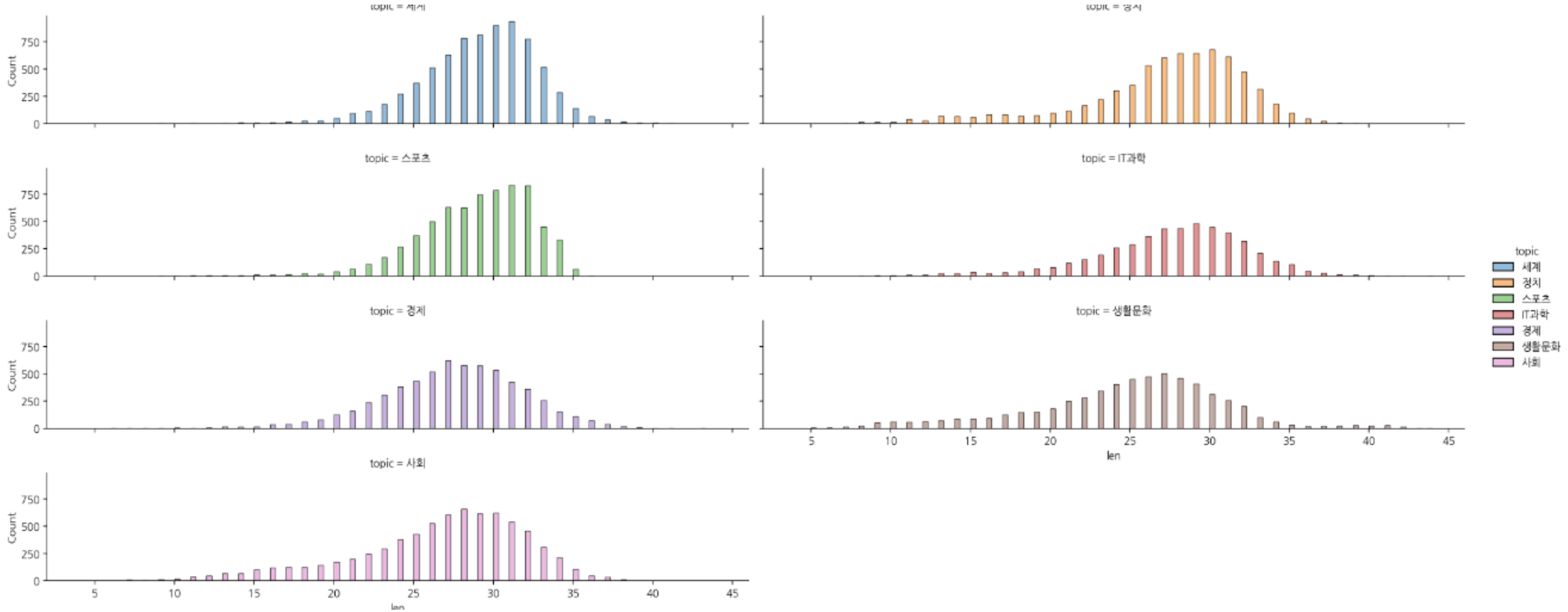


연합뉴스 타이틀 주제 분류 : 데이터 전처리(전체적인 데이터 파악 中)

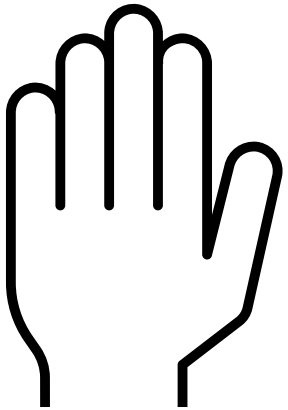
- 주제별 글자와 단어의 빈도 확인(시각화)

```
sns.displot(data=df, x="len", hue="topic", col="topic", col_wrap=2, aspect=5, height=2)
```

시각화



[참고] 데이터 시각화 : seaborn



잠깐! seaborn의 displot 등 대해 자세히 알고 싶다면
공식 레퍼런스를 참고하면 됩니다.

https://seaborn.pydata.org/generated/seaborn.distplot.html

matplotlib

https://matplotlib.org/stable/api/index

연합뉴스 타이틀 주제 분류 : 데이터 전처리

1. 전체적인 데이터 특성 파악 : 문자 길이, 단어의 등장 빈도 확인(시각화)

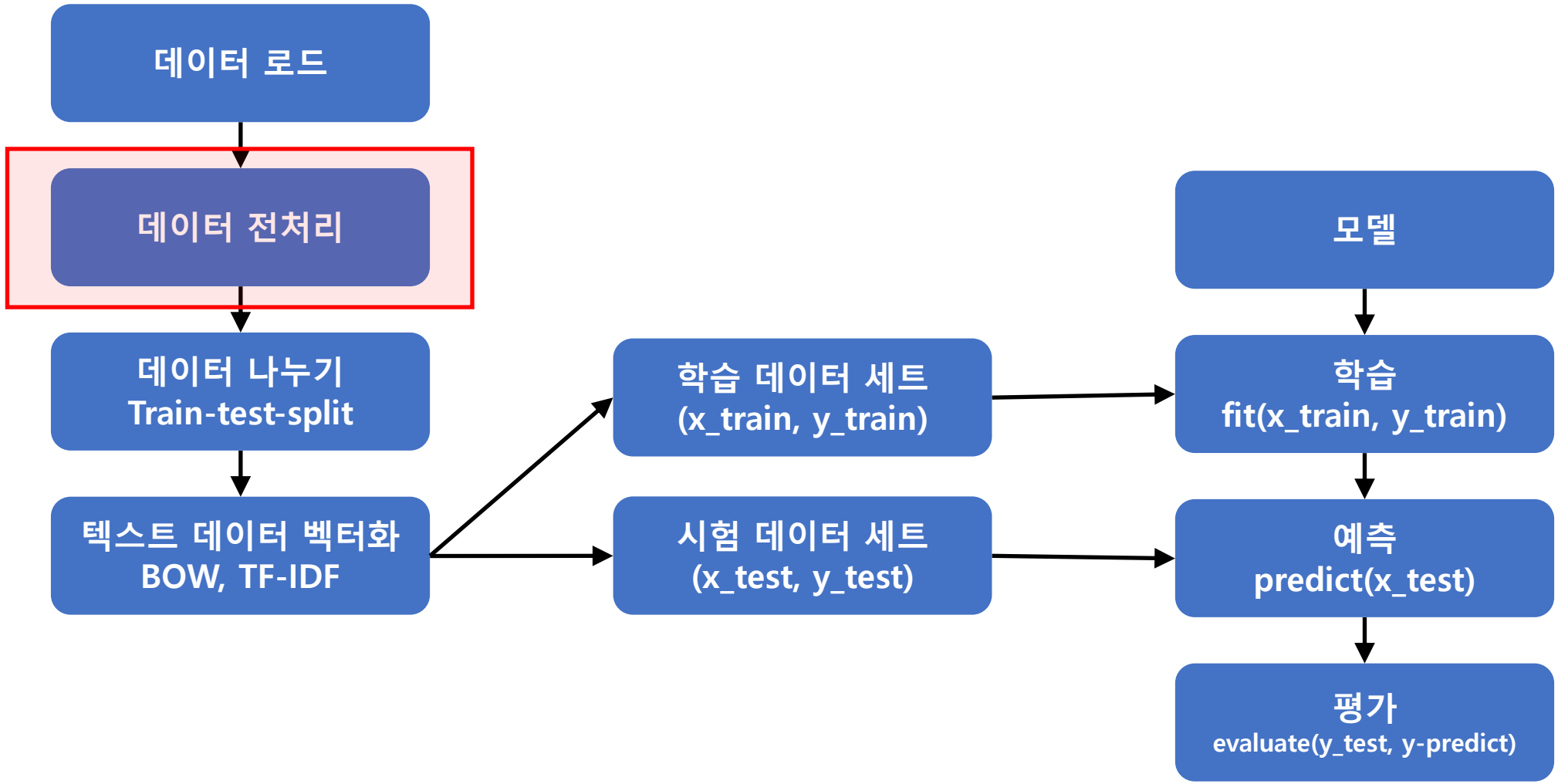
2. 불필요한 기호를 제거 후 중요 내용 추출

연합뉴스 타이틀 주제 분류 : 데이터 전처리(불필요한 기호 제거, 중요 내용 추출)

- 불필요한 기호 제거

  - 숫자에 의미가 없으므로 숫자 제거
  - 영어 대소문자에 따라 다른 단어로 분류할 필요 없으므로 소문자로 통일
  - 형태소 분석기(KoNLPy)로 조사, 어미, 구두점 제거
  - 의미없는 단어(불용어) 제거

연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



연합뉴스 타이틀 주제 분류 : 데이터 나누기

(54785, 4)

	index	title	topic_idx	topic
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0	세계
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4.0	세계
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0	세계
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0	세계
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0	세계

트레이닝용 데이터

테스트용 데이터

	index	title	topic_idx	topic
54780	54780	인천 오후 3시35분 대설주의보…눈 3.1cm 쌓여	NaN	NaN
54781	54781	노래방에서 지인 성추행 외교부 사무관 불구속 입건종합	NaN	NaN
54782	54782	40년 전 부마항쟁 부산 시위 사진 2점 최초 공개	NaN	NaN
54783	54783	게시판 아리랑TV 아프리카개발은행 총회 개회식 생중계	NaN	NaN
54784	54784	유영민 과기장관 강소특구는 지역 혁신의 중심…지원책 강구	NaN	NaN

연합뉴스 타이틀 주제 분류 : 데이터 나누기

Training Data

	index	title	topic_idx	topic	len	word_count	unique_word_count
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4.0	세계	24	5	5
1	1	실리콘밸리 넘어서겠다…구글 조원 들여 美전역 거점화	4.0	세계	30	6	6
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4.0	세계	30	8	8
3	3	nyt 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4.0	세계	32	7	7
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4.0	세계	25	7	7

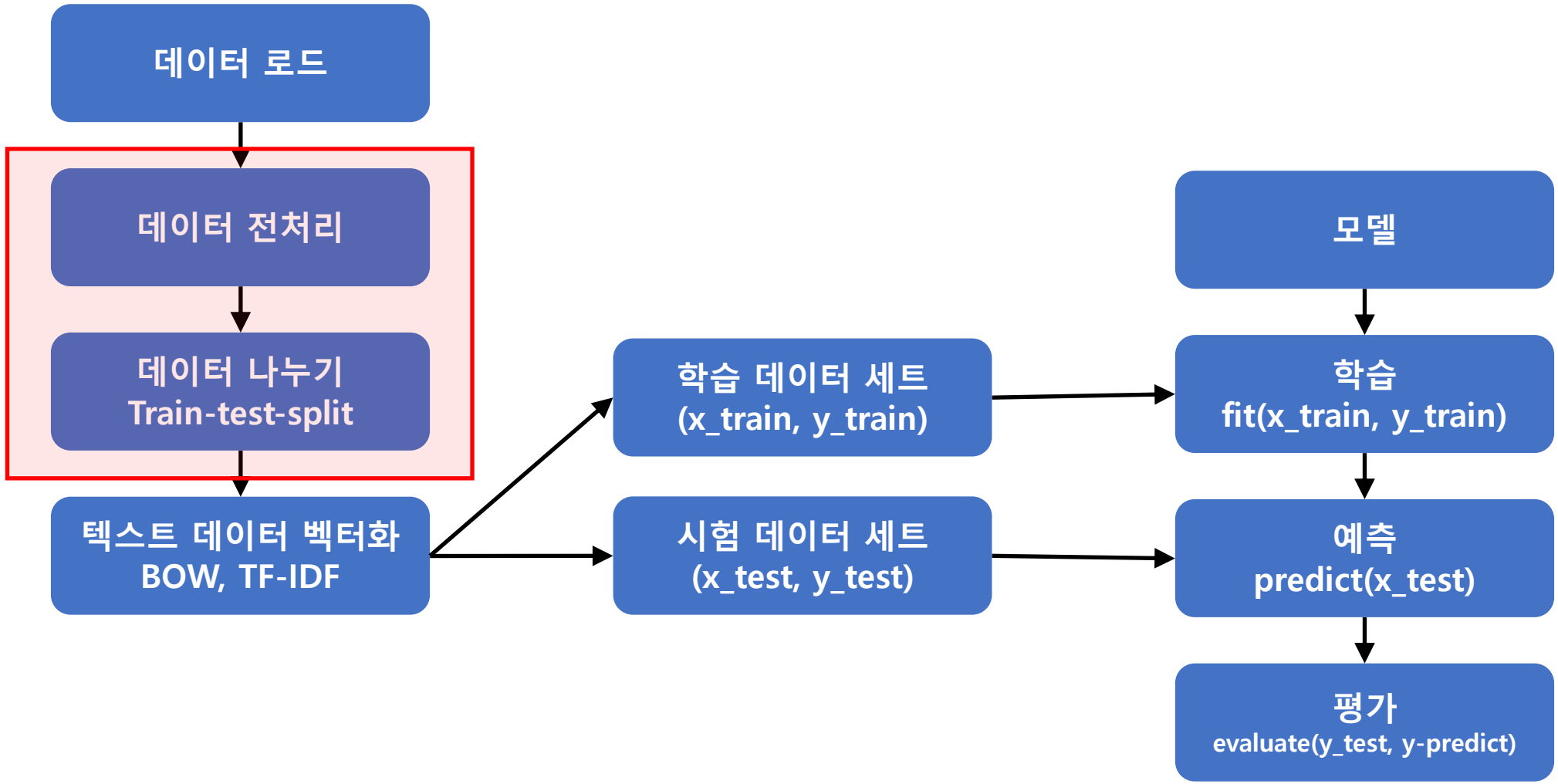
(45654, 7)

Test Data

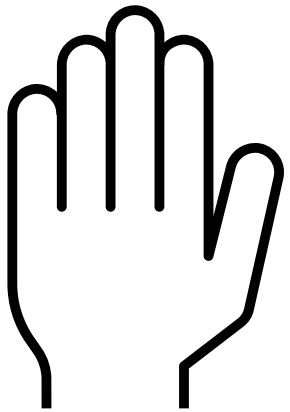
	index	title	topic_idx	topic	len	word_count	unique_word_count
45654	45654	유튜브 내달 일까지 크리에이터 지원 공간 운영	NaN	NaN	26	7	7
45655	45655	어버이날 맑다가 흐려져…남부지방 옅은 황사	NaN	NaN	23	5	5
45656	45656	내년부터 국가rd 평가 때 논문건수는 반영 않는다	NaN	NaN	27	7	7
45657	45657	김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것	NaN	NaN	31	10	10
45658	45658	회색인간 작가 김동식 양심고백 등 새 소설집 권 출간	NaN	NaN	30	9	9

(9131, 7)

연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



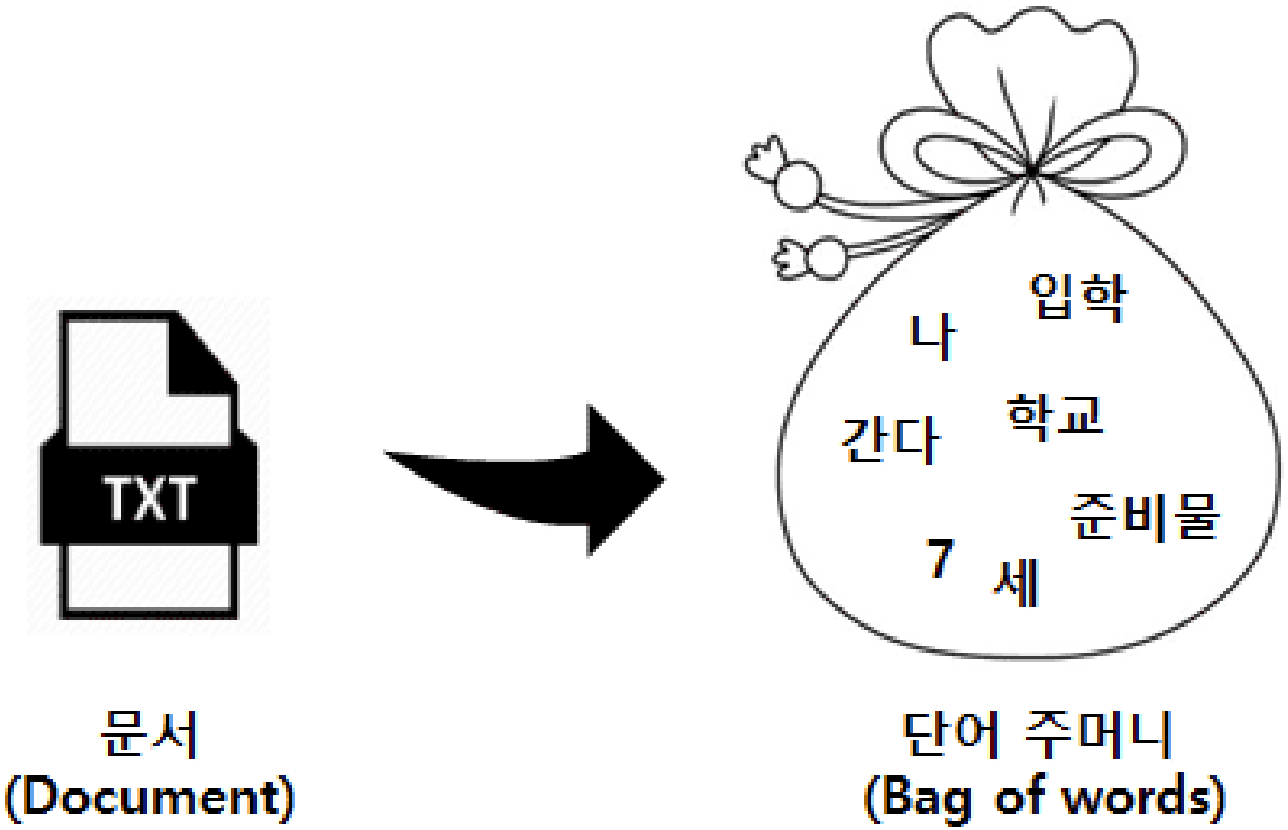
연합뉴스 타이틀 주제 분류 : 단어 벡터화 하기



잠깐!

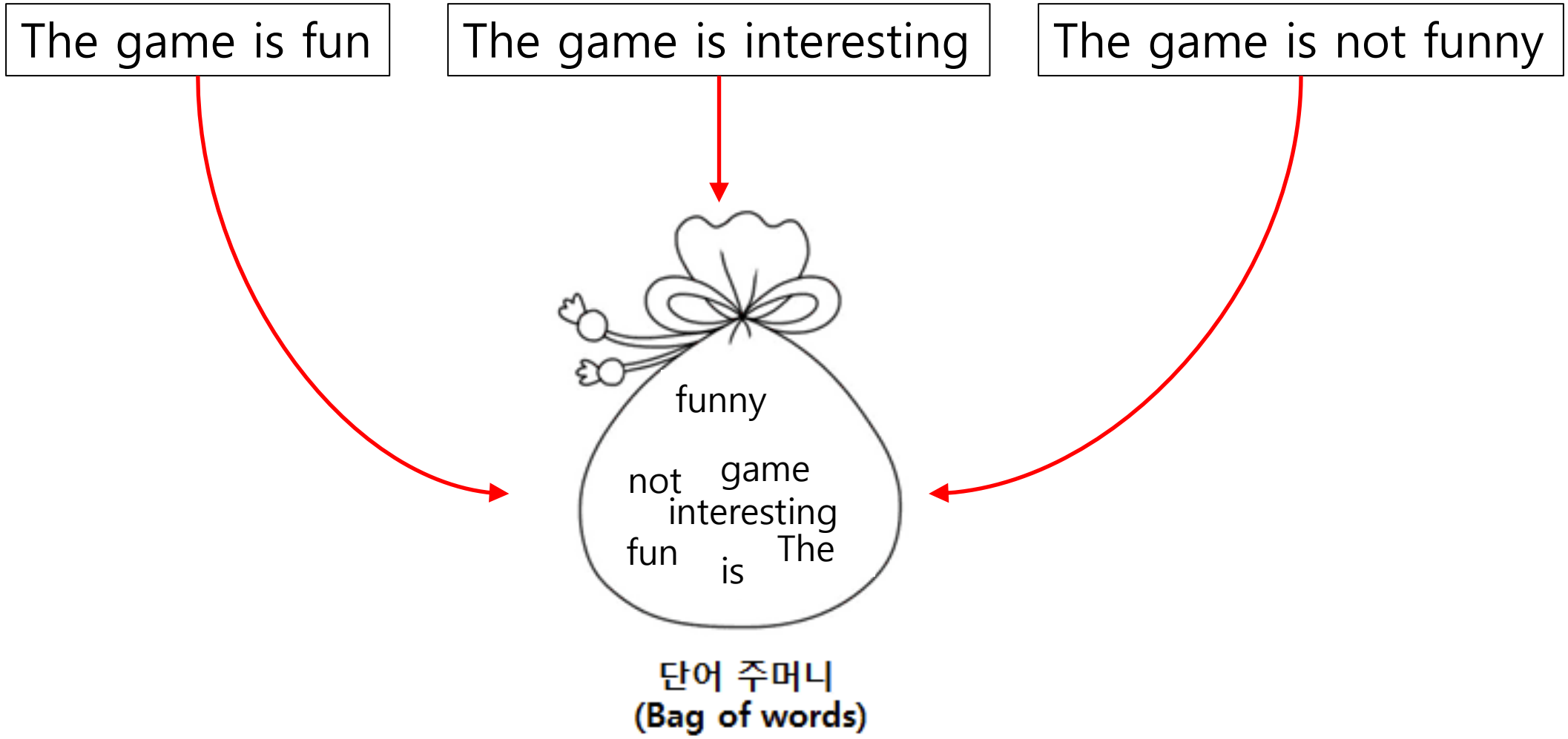
[4장 단어 가방 모형]의 내용을 잠깐 보고 갈게요!

단어 가방 모형(Bag Of Words) : 텍스트 분석 기법



이미지 및 내용 출처 : https://team-platform.tistory.com/34

단어 가방 모형(Bag Of Words) : 텍스트 분석 기법



이미지 및 내용 출처 : https://team-platform.tistory.com/34

단어 가방 모형(Bag Of Words) : 텍스트 분석 기법

단어 주머니
(Bag of words)

the	game	is	fun	interesting	not	funny
1	1	1	1	0	0	0

> The game is fun : [1, 1, 1, 1, 0, 0, 0]

the	game	is	fun	interesting	not	funny
1	1	1	0	1	0	0

> The game is interesting : [1, 1, 1, 0, 1, 0, 0]

the	game	is	fun	interesting	not	funny
1	1	1	0	0	1	1

> The game is not funny : [1, 1, 1, 0, 0, 1, 1]

Binary Vector로 표현

이미지 및 내용 출처 : https://team-platform.tistory.com/34

단어 가방 모형(Bag Of Words)의 단점 : 텍스트 분석 기법

- **Sparsity(희소행렬)**

  실제 사전에는 100만개가 넘는 단어들이 있을 수도 있다.

  the game is fun [1,1,1,1,0,0,0,0,0,0,0,,,,,,0,0,0,0,0]

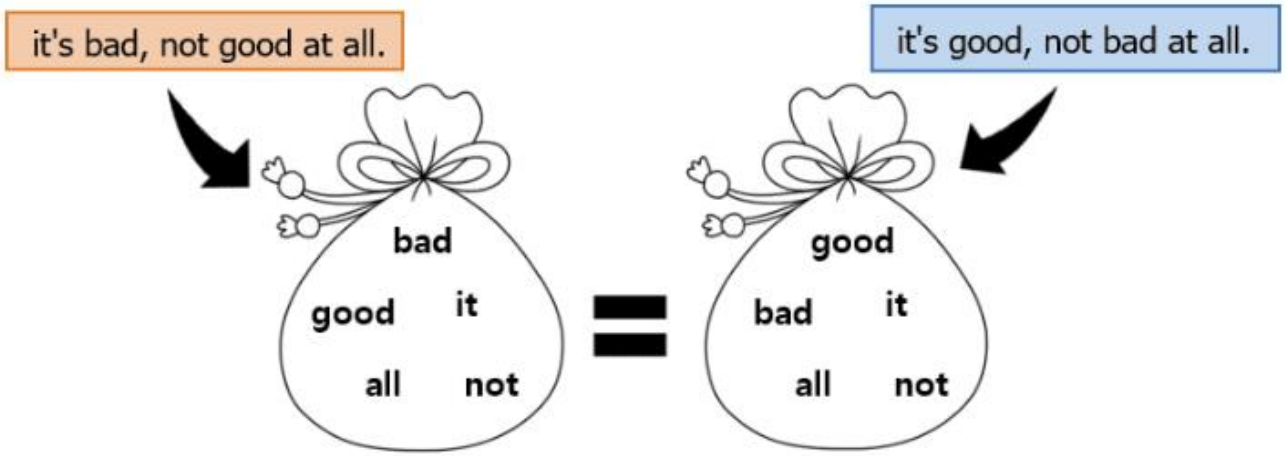
- **빈번한 단어는 더 많은 힘을 가진다.**

  만약, 의미 없는 단어들이 많이 사용되었다면?

- **Out of vocabulary**

  오타나 신조어?

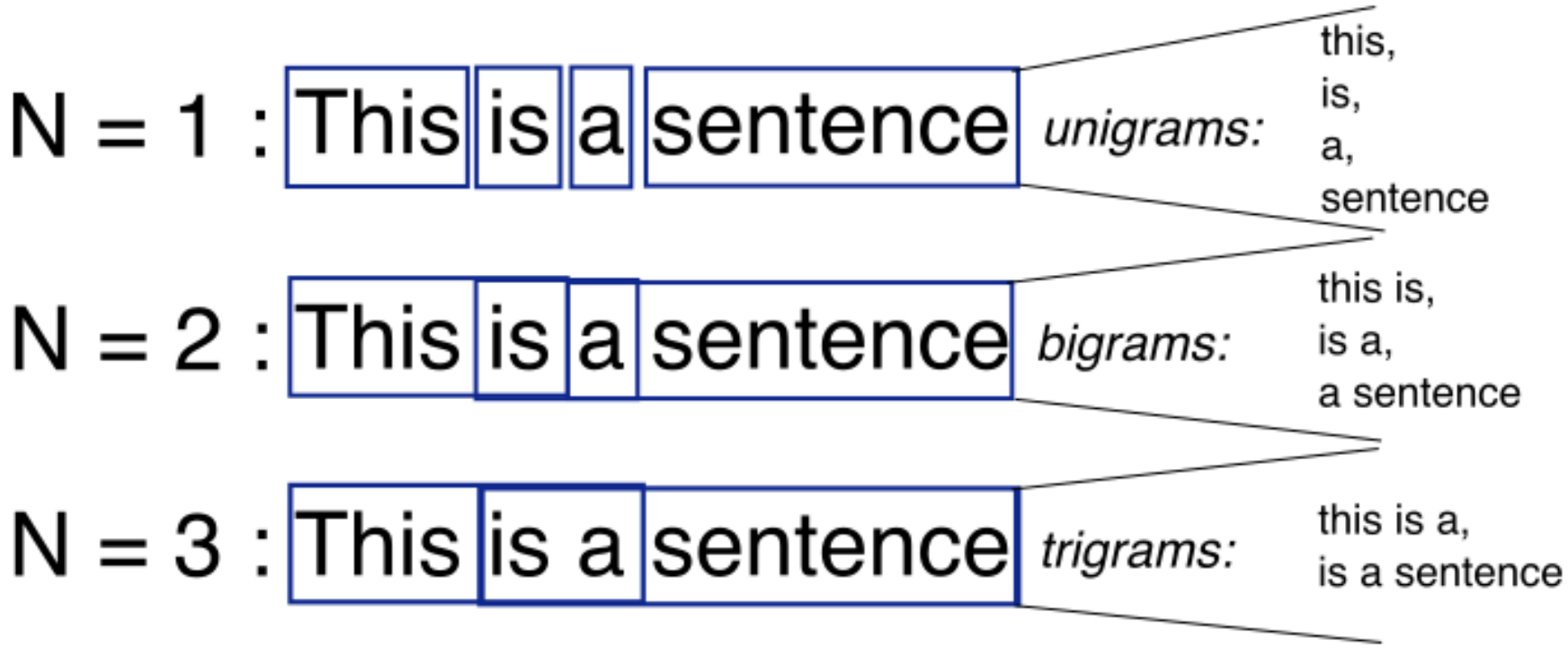
- **단어의 순서가 무시됨**



이미지 및 내용 출처 : https://team-platform.tistory.com/34

단어 가방 모형(Bag Of Words)의 단점 극복하기 : 텍스트 분석 기법

n-gram : 앞 뒤 토큰 몇 개(n)을 묶어서 보자!



이미지 및 내용 출처 https://deepai.org/machine-learning-glossary-and-terms/n-gram

텍스트의 전처리

정규화 normalization (입니닼ㅋㅋ -> 입니다 ㅋㅋ, 샤릉해 -> 사랑해)
한국어를 처리하는 예시입니닼ㅋㅋㅋㅋㅋ -> 한국어를 처리하는 예시입니다 ㅋㅋ

토큰화 tokenization
한국어를 처리하는 예시입니다 ㅋㅋ -> 한국어Noun, 를Josa, 처리Noun, 하는Verb, 예시Noun, 입Adjective, 니다Eomi ㅋㅋKoreanParticle

어근화 stemming (입니다 -> 이다)
한국어를 처리하는 예시입니다 ㅋㅋ -> 한국어Noun, 를Josa, 처리Noun, 하다Verb, 예시Noun, 이다Adjective, ㅋㅋKoreanParticle

어근 추출 phrase extraction
한국어를 처리하는 예시입니다 ㅋㅋ -> 한국어, 처리, 예시, 처리하는 예시

이미지 및 내용 출처 : https://team-platform.tistory.com/34

텍스트의 전처리

토큰화(Tokenization)

I loved you. machine learning

I | loved | you | machine | learning

나는 머신 러닝을 사랑합니다

나는 | 머신 | 러닝을 | 사랑합니다

나 | 머신 | 러닝 | 사랑합니다

한국어 형태소 분석기 필요
우리는,
KoNLPy(코엔엘파이) 사용

텍스트의 전처리

불용어(stopword)

I loved you. machine learning

I | loved | you | machine | learning

나는 머신 러닝을 사랑합니다

나는 | 머신 | 러닝을 | 사랑합니다

나 | 머신 | 러닝 | 사랑합니다

한국어 형태소 분석기 필요
우리는, KoNLPy의
Okt(Opensource Korean Text Processor) 사용

[참고] KoNLpy(코엔엘파이)의 형태소 분석기

1. KKma

Kkma는 서울대학교 IDS 연구실에서 자연어 처리를 위해 개발한 한국어 형태소 분석기입니다.

'꼬꼬마'로 발음합니다.

KKma(꼬꼬마) 홈페이지 : http://kkma.snu.ac.kr/

위의 홈페이지에서 자세한 정보를 확인할 수 있습니다.

2. Komoran

Komoran은 Shineware에서 개발한 자바(Java) 기반의 한국어 형태소 분석기입니다.

'코모란'으로 발음합니다.

코모란은 공백이 포함된 형태소 단위로 분석이 가능합니다.

Komoran(코모란) 홈페이지 : https://www.shineware.co.kr/products/komoran/

3. Okt (Open-source Korean Text Processor)

Okt는 트위터에서 개발한 Twitter 한국어 처리기에서 파생된 오픈소스 한국어 처리기입니다.

Okt GitHub : https://github.com/open-korean-text/open-korean-text

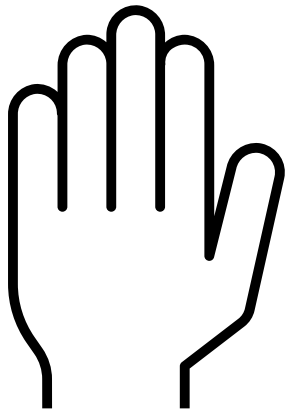
이미지 및 내용 출처 : https://seokii.tistory.com/28

연합뉴스 타이틀 주제 분류 : 단어 벡터화 하기

- 사이킷런(sklearn)을 이용하여 *단어 가방 벡터*를 만든다.

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vect=TfidfVectorizer(tokenizer=None,
                           ngram_range=(1,2),
                           min_df=3,
                           max_df=0.95)

tfidf_vect.fit(X_train)
```

연합뉴스 타이틀 주제 분류 : 단어 벡터화 하기



잠깐!

BOW(Bag Of Words)의 TF-IDF내용을 잠깐 보고 갈게요!

[참고] BOW(Bag Of Word)

```
corpus=["코로나 거리두기와 코로나 상생지원금 문의입니다.",
        "지하철 운행시간과 지하철 요금 문의입니다.",
        "지하철 승강장 문의입니다.",
        "택시 승강장 문의입니다."]
```

```
from sklearn.feature_extraction.text import CountVectorizer
cvect=CountVectorizer()
cvect.fit(corpus)
dtm=cvect.transform(corpus)
dtm
```

```
<4x9 sparse matrix of type '<class 'numpy.int64'>'
        with 14 stored elements in Compressed Sparse Row format>
```

→ 4X9의 정수형 숫자가 들어가 있는 희소행렬이며 14개의 요소를 가지고 있다

내용 출처 : 딥 러닝을 이용한 자연어 처리 입문(유원준, 안상준)

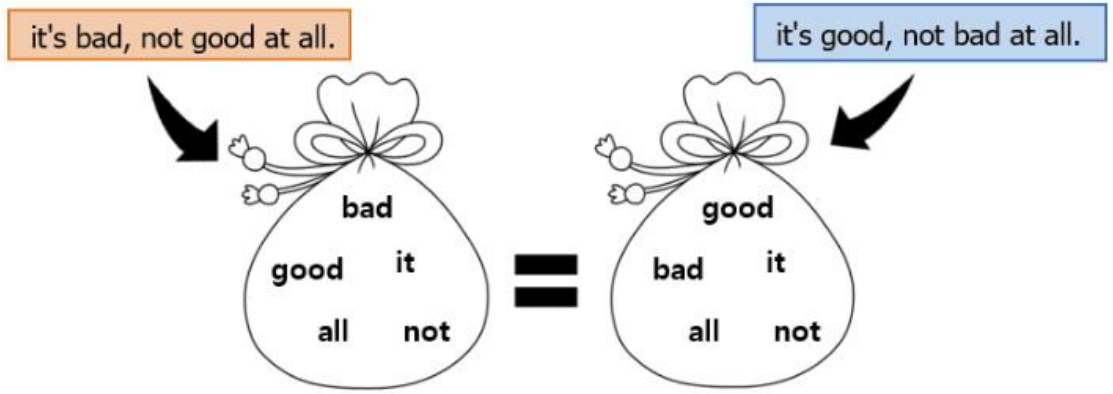
[참고] BOW(Bag Of Word) : 문서 단어 행렬(Document-Term Matrix, DTM)

	거리두기와	문의입니다	상생지원금	승강장	요금	운행시간과	지하철	코로나	택시
0	1	1	1	0	0	0	0	2	0
1	0	1	0	0	1	1	2	0	0
2	0	1	0	1	0	0	1	0	0
3	0	1	0	1	0	0	0	0	1

	거리두기와	문의입니다	상생지원금	승강장	요금	운행시간과	지하철	코로나	택시
0	1	4	1	2	1	1	3	2	1

단어의 등장 빈도만으로 행렬을 구성하면

앞뒤 맥락을 잃어버릴 수 있다.



[참고] BOW(Bag Of Word) : n-gram 앞뒤 단어 묶어서 사용

```
1 cvect=CountVectorizer(ngram_range=(1,2))
2 dtm=cvect.fit_transform(corpus)
3 dtm
```

```
<4x20 sparse matrix of type '<class 'numpy.int64'>'
        with 26 stored elements in Compressed Sparse Row format>
```

	거리두기와	거리두기와 코로나	문의입니다	상생지원금	상생지원금 문의입니다	승강장	승강장 문의입니다	요금	요금 문의입니다
0	1	1	1	1	1	0	0	0	0
1	0	0	1	0	0	0	0	1	1
2	0	0	1	0	0	1	1	0	0
3	0	0	1	0	0	1	1	0	0

	거리두기와	거리두기와 코로나	문의입니다	상생지원금	상생지원금 문의입니다	승강장	승강장 문의입니다	요금	요금 문의입니다
0	1	1	4	1	1	2	2	1	1

[참고] BOW(Bag Of Word) TF-IDF

빈도만 고려해 단어 가방 모형을 만들면 빈도수가 높은 단어일수록 중요한 단어라고 생각한다.

각 문서의 특성을 구분할 수 있는 단어는 높은 가중치를 주고, 그렇지 않은 단어는 낮은 가중치를 주자.

구분	의미	내용
TF	단어 빈도 Term Frequency	• 특정한 단어가 문서 안에서 얼마나 자주 등장하는지 나타내는 값 • 이 값이 높을수록 문서에서 중요하다고 생각할 수 있음
DF	문서 빈도 Document Frequency	• 특정 단어가 등장한 문서의 수 • 단어 자체가 문서군 안에서 자주 사용되고, 흔하게 등장한다는 의미
IDF	역문서 빈도 Inverse Document Frequency	• DF의 역수로 DF에 반비례하는 수
TF-IDF	TF와 IDF를 곱한 값	• 대부분의 문서에 자주 등장하는 단어는 낮은 중요도로 계산 • 특정 문서에만 자주 등장하는 단어는 높은 중요도로 계산

[참고] BOW(Bag Of Word) TF-IDF

```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2 tfidfvect=TfidfVectorizer()
3 tfidfvect.fit(corpus)
4 dtm=tfidfvect.transform(corpus)
5 dtm
```

```
<4x9 sparse matrix of type '<class 'numpy.float64'>'
        with 14 stored elements in Compressed Sparse Row format>
```

	거리두기와	문의입니다	상생지원금	승강장	요금	운행시간과	지하철	코로나	택시
0	0.399288	0.208365	0.399288	0.000000	0.000000	0.000000	0.000000	0.798575	0.000000
1	0.000000	0.239219	0.000000	0.000000	0.458412	0.458412	0.722835	0.000000	0.000000
2	0.000000	0.423897	0.000000	0.640434	0.000000	0.000000	0.640434	0.000000	0.000000
3	0.000000	0.379192	0.000000	0.572892	0.000000	0.000000	0.000000	0.000000	0.726641

['코로나 거리두기와 코로나 상생지원금 문의입니다.',
'지하철 운행시간과 지하철 요금 문의입니다.',
'지하철 승강장 문의입니다.',
'택시 승강장 문의입니다.']

[5장]연합뉴스 타이틀 주제 분류 : 머신러닝을 통한 뉴스 텍스트 분류

https://dacon.io/competitions/official/235747/data

이 문제로 다시 돌아갑시다!

연합뉴스 타이틀 주제 분류 : 단어 벡터화 하기

- 사이킷런(sklearn)을 이용하여 *단어 가방 벡터*를 만든다.

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vect=TfidfVectorizer(tokenizer=None,
                           ngram_range=(1,2),
                           min_df=3,
                           max_df=0.95)

tfidf_vect.fit(X_train)
```

- 벡터를 행렬(DTM)로 변환하기

```
1 train_feature_tfidf=tfidf_vect.transform(X_train)
2 test_feature_tfidf=tfidf_vect.transform(X_test)
3
4 train_feature_tfidf.shape, test_feature_tfidf.shape
```

```
((45654, 22385), (9131, 22385))
```

연합뉴스 타이틀 주제 분류 : 단어 벡터화 하기

- 생성된 단어 사전 확인

```
1 vocab=tfidf_vect.get_feature_names_out()
2 print(len(vocab))
3 vocab[:10]
```

```
22385
array(['aa로', 'abs', 'acl', 'afc', 'afc 챔스리그', 'afc 챔피언십', 'afc 회장', 'ag',
       'ag 우승', 'ai'], dtype=object)
```

- 전체 단어 사전에서 가중치 값의 합계 살펴보기

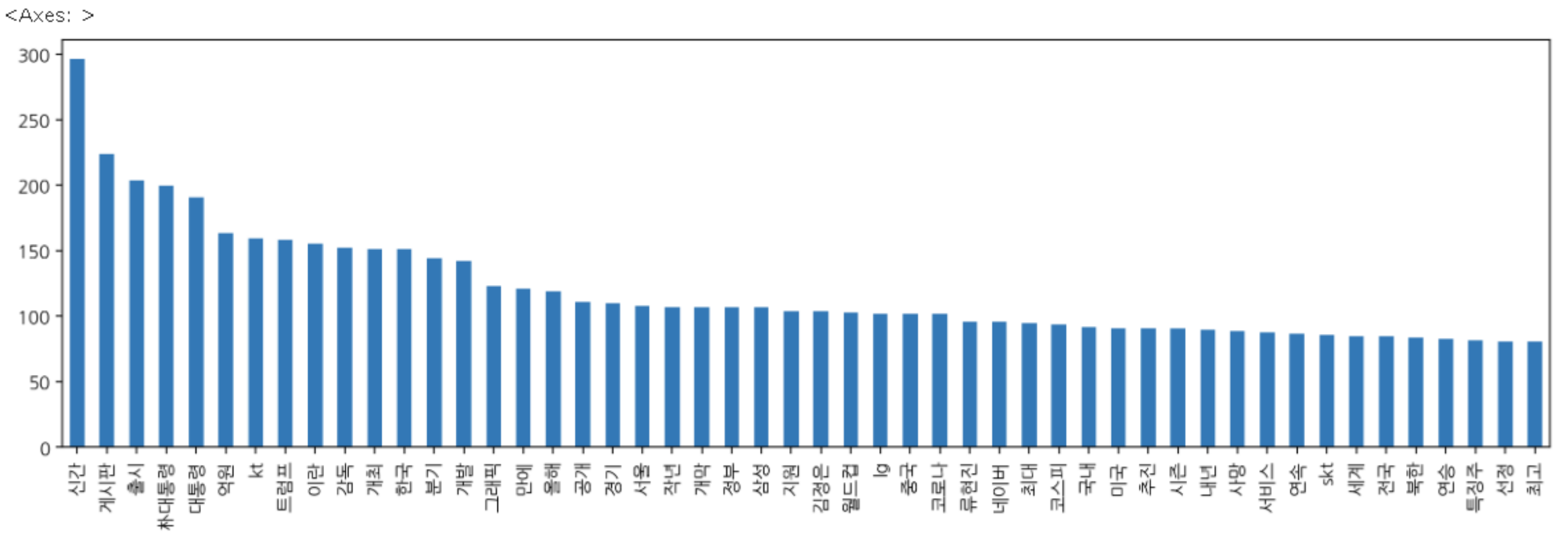
```
1 dist=np.sum(train_feature_tfidf, axis=0)
2
3 vocab_count=pd.DataFrame(dist, columns=vocab)
4 vocab_count
```

	aa로	abs	acl	afc	afc 챔스리그	afc 챔피언십	afc 회장	ag	ag 우승	ai	...	힘으로
0	1.374165	1.493937	4.560771	10.036045	3.516982	2.254818	1.220953	14.847285	1.557569	74.285975	...	4.326381

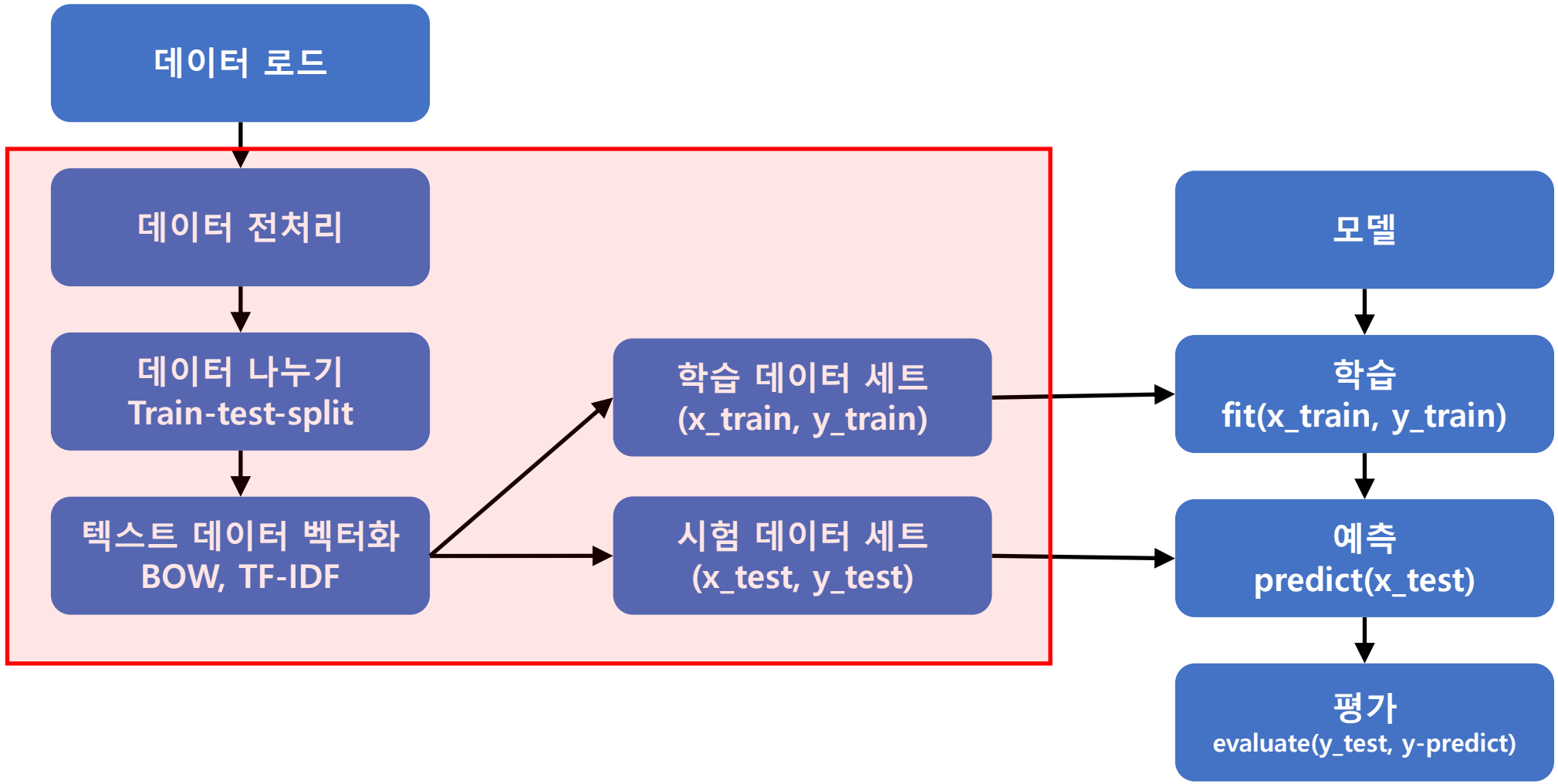
연합뉴스 타이틀 주제 분류 : 단어 벡터화 하기

- 시각화

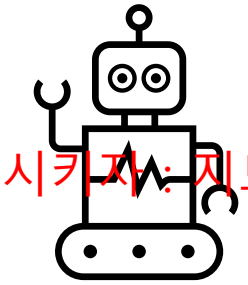
```
1 vocab_count.T[0].sort_values(ascending=False).head(50).plot.bar(figsize=(15,4))
```



연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



얘를 공부 시키자: 지도 학습으로

랜덤포레스트

데이터 로드

데이터 전처리

데이터 나누기
Train-test-split

텍스트 데이터 벡터화
BOW, TF-IDF

학습 데이터 세트
(x_train, y_train)

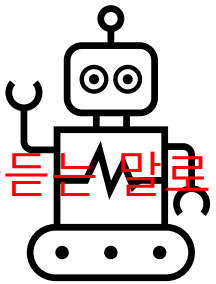
시험 데이터 세트
(x_test, y_test)

모델

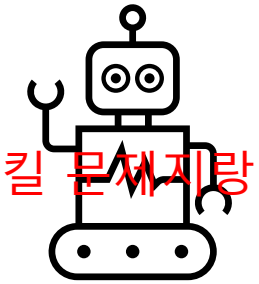
학습
fit(x_train, y_train)

예측
predict(x_test)

평가
evaluate(y_test, y-predict)



얘가 알아듣는 말로 바꿔준다



얘를 공부 시킬 문제지랑 시험지 준비

연합뉴스 타이틀 주제 분류 : 학습과 예측하기_랜덤포레스트분류기

사이킷런(scikit-Learn)의 랜덤포레스트분류기(RandomForestCalssifier)사용

```
1 #RandomForestClassifier를 불러 온다
2 from sklearn.ensemble import RandomForestClassifier
3
4 #랜덤 포레스트 분류기를 사용한다.
5 model=RandomForestClassifier(n_estimators=100, n_jobs=-1, random_state=42)
6 model
```

결정 트리의 개수 : 100개

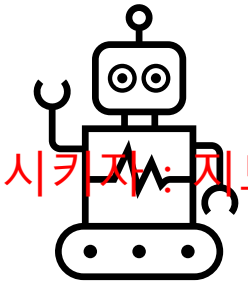
가능한 CPU 코어를 모두 사용하기

랜덤 시드를 지정해서 수행할 때마다 같은 결과를 얻도록 한다

RandomForestClassifier

RandomForestClassifier(n_jobs=-1, random_state=42)

연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



얘를 공부 시키자: 지도 학습으로

랜덤포레스트

데이터 로드

데이터 전처리

데이터 나누기
Train-test-split

텍스트 데이터 벡터화
BOW, TF-IDF

학습 데이터 세트
(x_train, y_train)

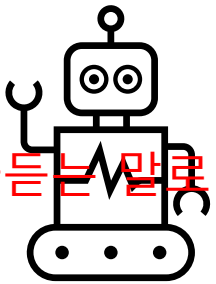
시험 데이터 세트
(x_test, y_test)

모델

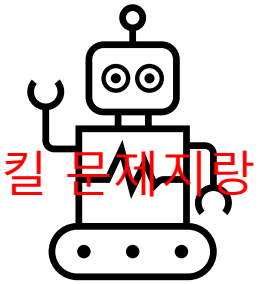
학습
fit(x_train, y_train)

예측
predict(x_test)

평가
evaluate(y_test, y-predict)



얘가 알아듣는 말로 바꿔준다



얘를 공부 시킬 문제지랑 시험지 준비

연합뉴스 타이틀 주제 분류 : 학습과 예측하기_학습

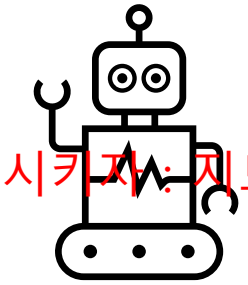
학습하기

```
%time model.fit(train_feature_tfidf, y_train)
```

단어 벡터 가방(TF-IDF)로 만든 테이블

학습 데이터 셋

연합뉴스 타이틀 주제 분류 작업 흐름도(p.114)



얘를 공부 시키자: 지도 학습으로

랜덤포레스트

데이터 로드

데이터 전처리

데이터 나누기
Train-test-split

텍스트 데이터 벡터화
BOW, TF-IDF

학습 데이터 세트
(x_train, y_train)

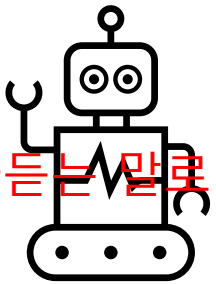
시험 데이터 세트
(x_test, y_test)

모델

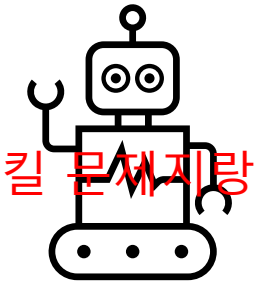
학습
fit(x_train, y_train)

예측
predict(x_test)

평가
evaluate(y_test, y-predict)



얘가 알아듣는 말로 바꿔준다



얘를 공부 시킬 문제지랑 시험지 준비

연합뉴스 타이틀 주제 분류 : 학습과 예측하기_예측

예측하기

```
1 y_predict=model.predict(test_feature_tfidf)
2 y_predict[:10]
```

```
array([2., 3., 2., 2., 3., 2., 5., 3., 4., 4.])
```

	index	title
0	45654	유튜브 내달 2일까지 크리에이터 지원 공간 운영
1	45655	어버이날 맑다가 흐려져…남부지방 옅은 황사
2	45656	내년부터 국가RD 평가 때 논문건수는 반영 않는다
3	45657	김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것
4	45658	회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간

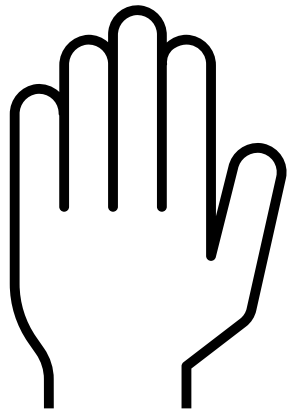
2
3
2
2
3

예측 값

topic	topic_idx
IT과학	0
경제	1
사회	2
생활문화	3
세계	4
스포츠	5
정치	6

사회
생활문화
사회
사회
생활문화

연합뉴스 타이틀 주제 분류



잠깐!

[과적합(overfitting)]을 아십니까?!

[참고] 과적합

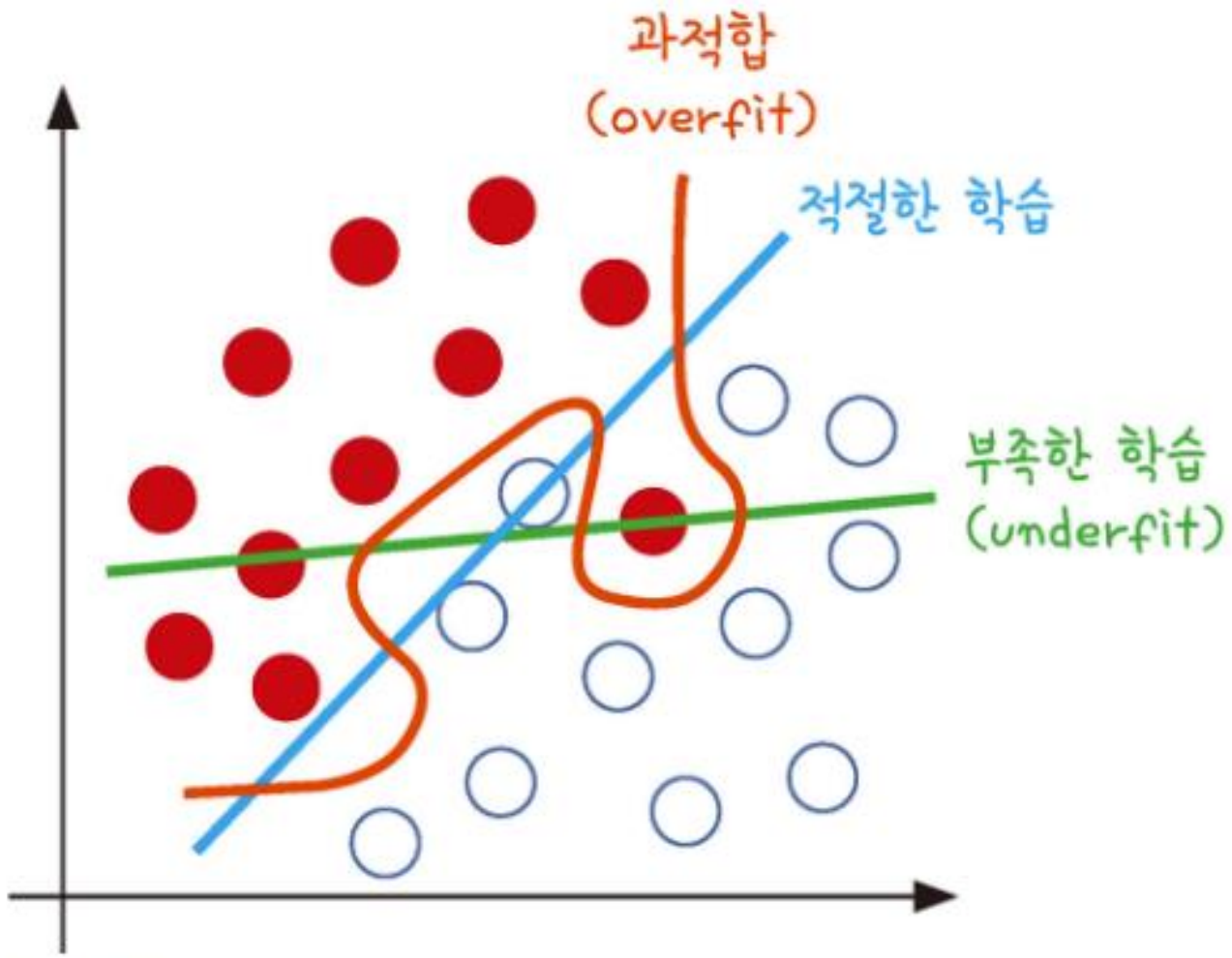


그림 13-1
과적합이 일어난 경우(빨간색)와 학습이 제대로 이루어지지 않은 경우(초록색)

모두의 딥러닝(조태호 저)

[참고] 과적합

- 과적합 피하기 : k-교차검증(cross validation)

그림 13-5 5겹 교차 검증의 도식

모두의 딥러닝(조태호 저)

 # 결론, 랜덤포레스트를 이용한 우리의 머신러닝을 학습한 기계는 이렇게 예측했다.
(정확도는 약 74%이다.)

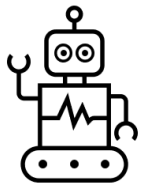
유튜브 내달 2일까지 크리에이터 지원 공간 운영

어버이날 맑다가 흐려져…남부지방 옅은 황사

내년부터 국가RD 평가 때 논문건수는 반영 않는다

김명자 신임 과총 회장 원로와 젊은 과학자 지혜 모을 것

회색인간 작가 김동식 양심고백 등 새 소설집 2권 출간



IT과학

경제

사회

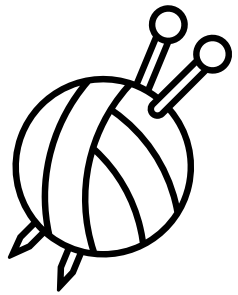
생활문화

세계

스포츠

정치

저자의 말로 돌아가보면...(p.10)



3장에서 4장으로 넘어갈 때 난이도가 **점프**한다는 느낌을 받을 것이다.

하지만 6장을 넘어서면 드디어 2년차 직장인처럼 **"아! 비슷비슷하구나!"** 하고 느낄 수 있기를 소망한다

쌤 생각에는...오늘 다 했어요!

[우리가 5장에서 경험해 본 것]

1. 데이터 다루기 : 판다스(Pandas), 넘파이(Numpy)

2. 시각화 : 맷플롯립(Matplotlib), 시본(Seaborn)

3. 머신러닝 : 사이킷런(Scikit-Learn)-랜덤포레스트

4. 텍스트 분석 : 단어 가방 모형, KoNLPy

다음 시간에는!

파이팅 해야지!