딥러닝 프로젝트

DL_PJT 2조

#김정빈 #박건우

Table of Contents

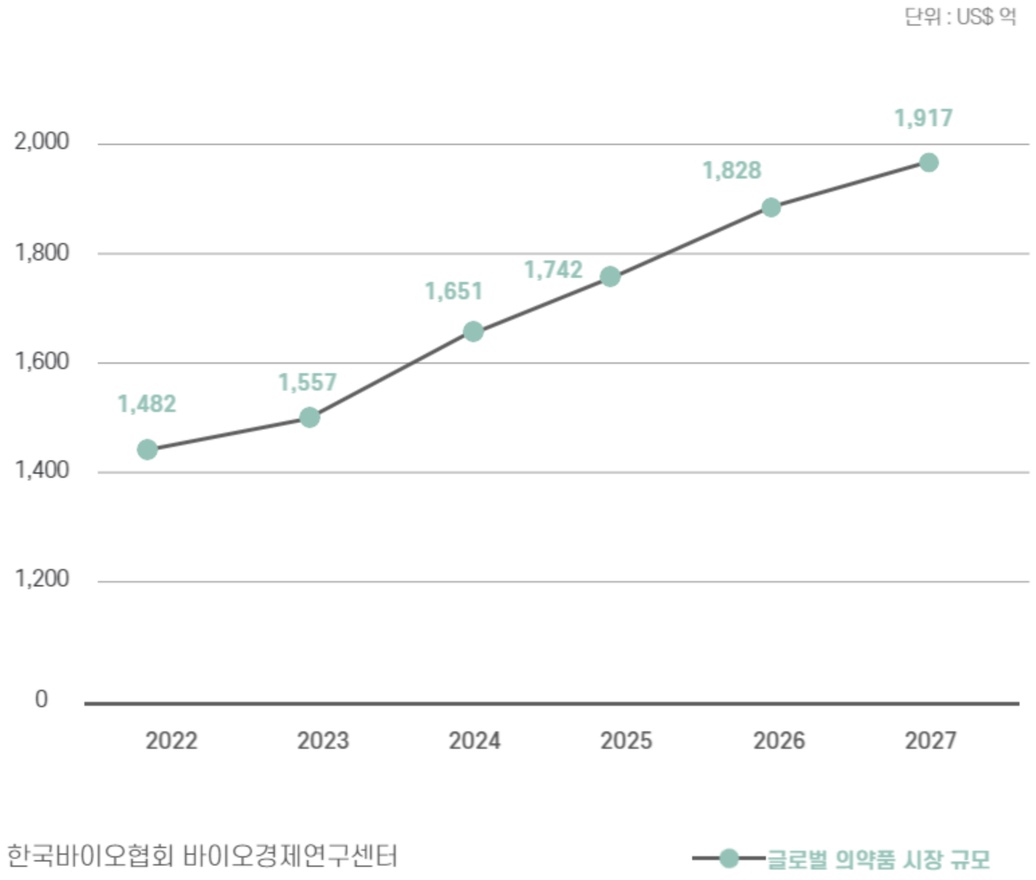
리뷰에 따른 의약품 추천시스템

dataset에 포함된 컬럼들을 이용해
리뷰에 따른 total_pred 생성

>>>

추천시스템을 통해
의약품 추천

INFOGRAPHIC



한국바이오협회 바이오경제연구센터

시장 연평균 **성장률**

3 ~ 6%

지난 몇년간 지속적인 성장세

2027년 시장 규모

1조 9,170억

향후 5년간 250개 이상의 신약 출시

AI 성능 향상 속도

10배

2010년 ~ 2020년 처리능력은 매년
약 10배씩 성장

#1 drugsComTrain_raw.csv

- 161297개의 row와 7개의 컬럼으로 구성된 train data
  - drugName : 의약품 명칭
  - condition : 증상
  - review : 복용자들의 리뷰
  - rating : 복용자 점수
  - date : 리뷰가 등록된 날짜
  - usefulCount : 리뷰가 유용하다고 생각하는 사용자 수

#2 drugsComTest_raw.csv

- 53766개의 row와 7개의 컬럼으로 구성된 test data
  - drugsComTrain_raw.csv와 컬럼 동일

#3 inquirerbasic.csv

- 11788개의 row와 4개의 컬럼으로 구성된 하버드 감성사전 data
  - entry : 단어
  - Positiv : 긍정 반응
  - Negativ : 부정 반응

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37
...	...	...	...	...	...	...	...
161292	191035	Campral	Alcohol Dependence	"I wrote my first report in Mid-October of 201...	10	31-May-15	125
161293	127085	Metoclopramide	Nausea/Vomiting	"I was given this in IV before surgey. I immed...	1	1-Nov-11	34
161294	187382	Orencia	Rheumatoid Arthritis	"Limited improvement after 4 months, developed...	2	15-Mar-14	35
161295	47128	Thyroid desiccated	Underactive Thyroid	"I've been on thyroid medication 49 years...	10	19-Sep-15	79
161296	215220	Lubiprostone	Constipation, Chronic	"I've had chronic constipation all my adu...	9	13-Dec-14	116

51297 rows × 7 columns

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	163740	Mirtazapine	Depression	"I've tried a few antidepressants over th...	10	28-Feb-12	22
1	206473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has done ...	8	17-May-09	17
2	159672	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9	29-Sep-17	3
3	39293	Contrave	Weight Loss	"Contrave combines drugs that were used for al...	9	5-Mar-17	35
4	97768	Cyclafem 1 / 35	Birth Control	"I have been on this birth control for one cyc...	9	22-Oct-15	4
...	...	...	...	...	...	...	...
53761	159999	Tamoxifen	Breast Cancer, Prevention	"I have taken Tamoxifen for 5 years. Side effe...	10	13-Sep-14	43
53762	140714	Escitalopram	Anxiety	"I've been taking Lexapro (escitaloprgra...	9	8-Oct-16	11
53763	130945	Levonorgestrel	Birth Control	"I'm married, 34 years old and I have no ...	8	15-Nov-10	7
53764	47656	Tapentadol	Pain	"I was prescribed Nucynta for severe neck/shou...	1	28-Nov-11	20
53765	113712	Arthrotec	Sciatica	"It works!!!"	9	13-Sep-09	46

53766 rows × 7 columns

	Entry	Source	Positiv	Negativ
0	A	H4Lvd	NaN	NaN
1	ABANDON	H4Lvd	NaN	Negativ
2	ABANDONMENT	H4	NaN	Negativ
3	ABATE	H4Lvd	NaN	Negativ
4	ABATEMENT	Lvd	NaN	NaN
...	...	...	...	...
11783	ZENITH	H4	Positiv	NaN
11784	ZERO	H4Lvd	NaN	NaN
11785	ZEST	H4	Positiv	NaN
11786	ZINC	H4Lvd	NaN	NaN
11787	ZONE	H4Lvd	NaN	NaN

11788 rows × 4 columns

EDA

1. feature간 상관관계 분석
2. review, drugName, condition 데이터 분석
3. date를 활용해 분석
4. review 말뭉치별 rating 분석

data preprocessing

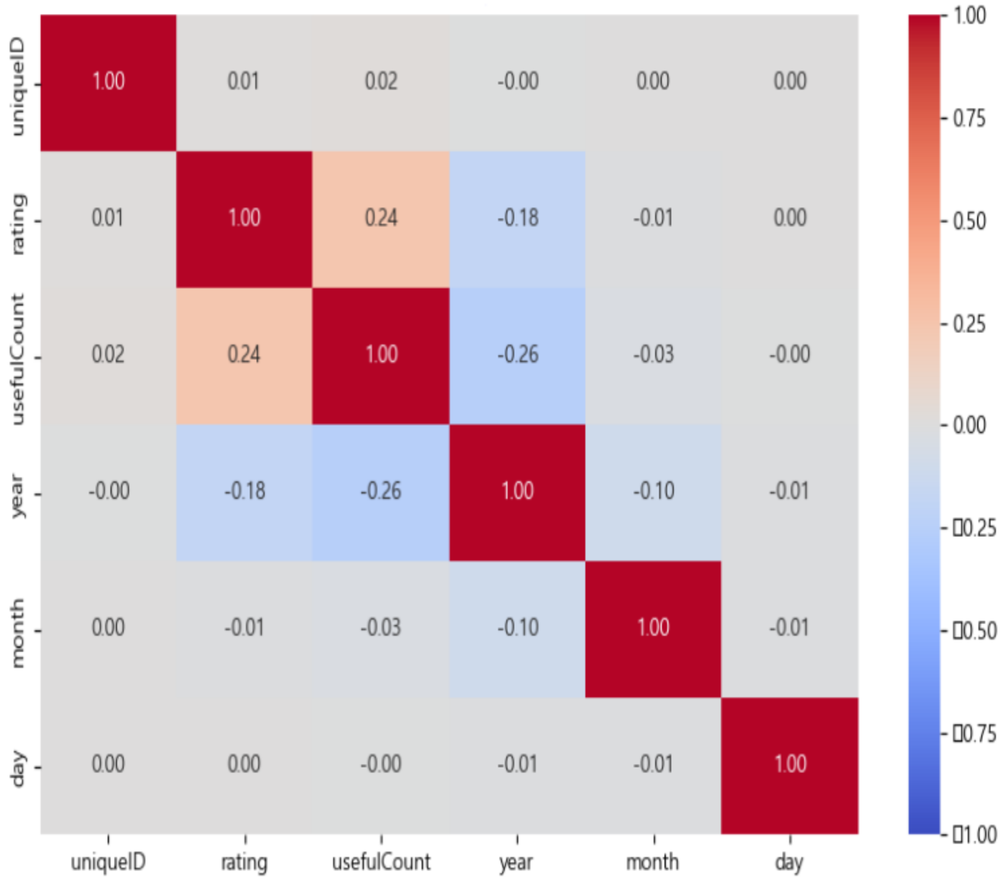
1. 결측치 제거
2. condition 전처리
3. review 전처리

modeling

1. 딥러닝 모델(N-gram)
2. Dictionary Sentiment Analysis

evaluate

1. 평가지표를 활용한 모델 평가

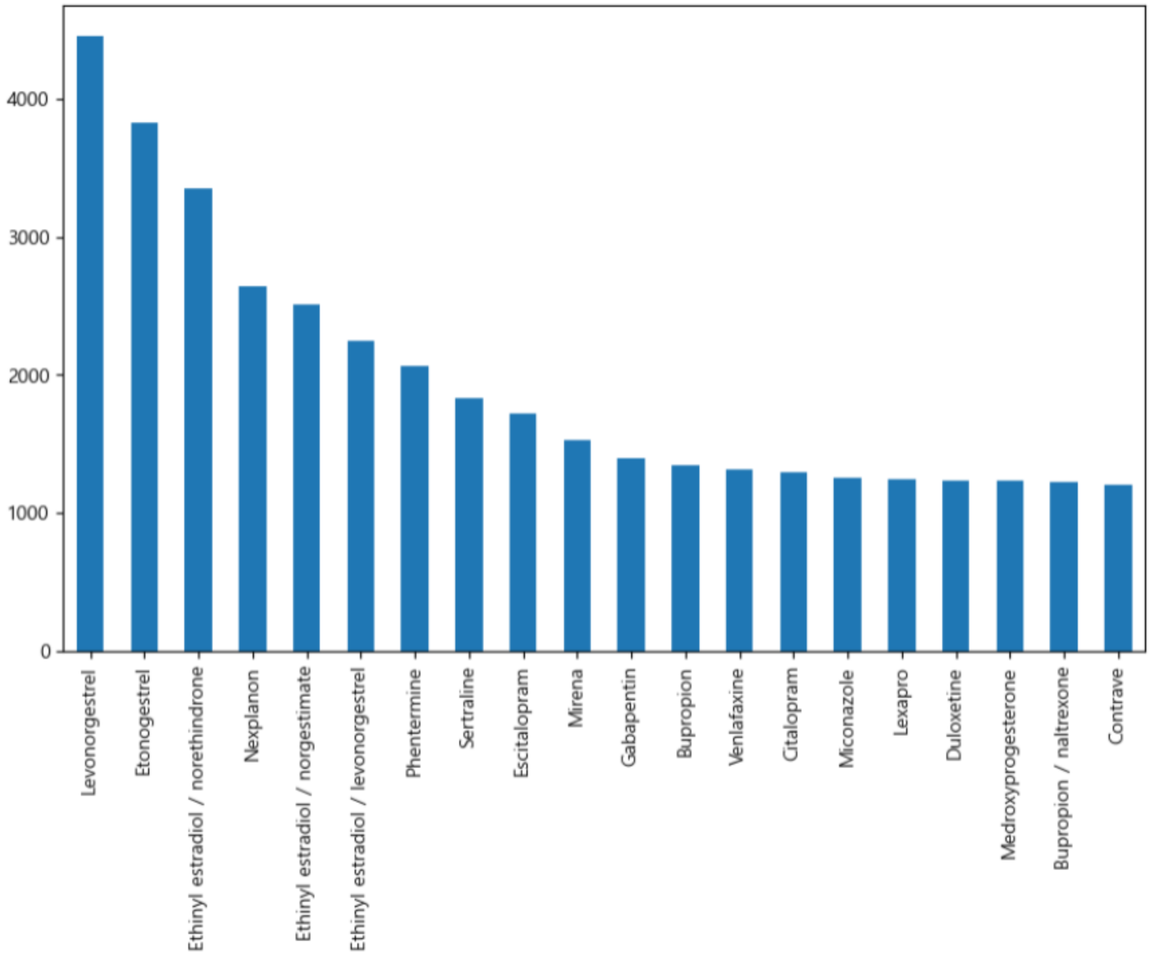


FEATURE

상관관계

가장 높은 상관관계는 usefulCount와 year(-0.26)

=> dataset내의 feature들간의 상관관계가 높지 않다.



리뷰수가 많은 의약품 상위 20개

가장 리뷰수가 많은 의약품은 levonorgestrel(응급 피임약) 이다.

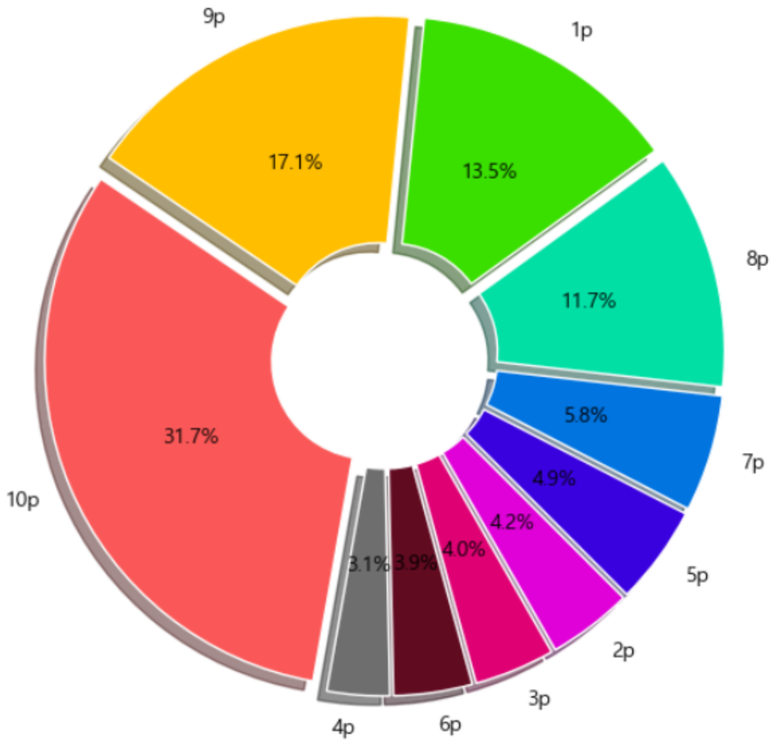
그 다음으로 ETONOGESTREL(피임약), ethinyl estradiol(여성 호르몬제)가 리뷰수가 많았다.

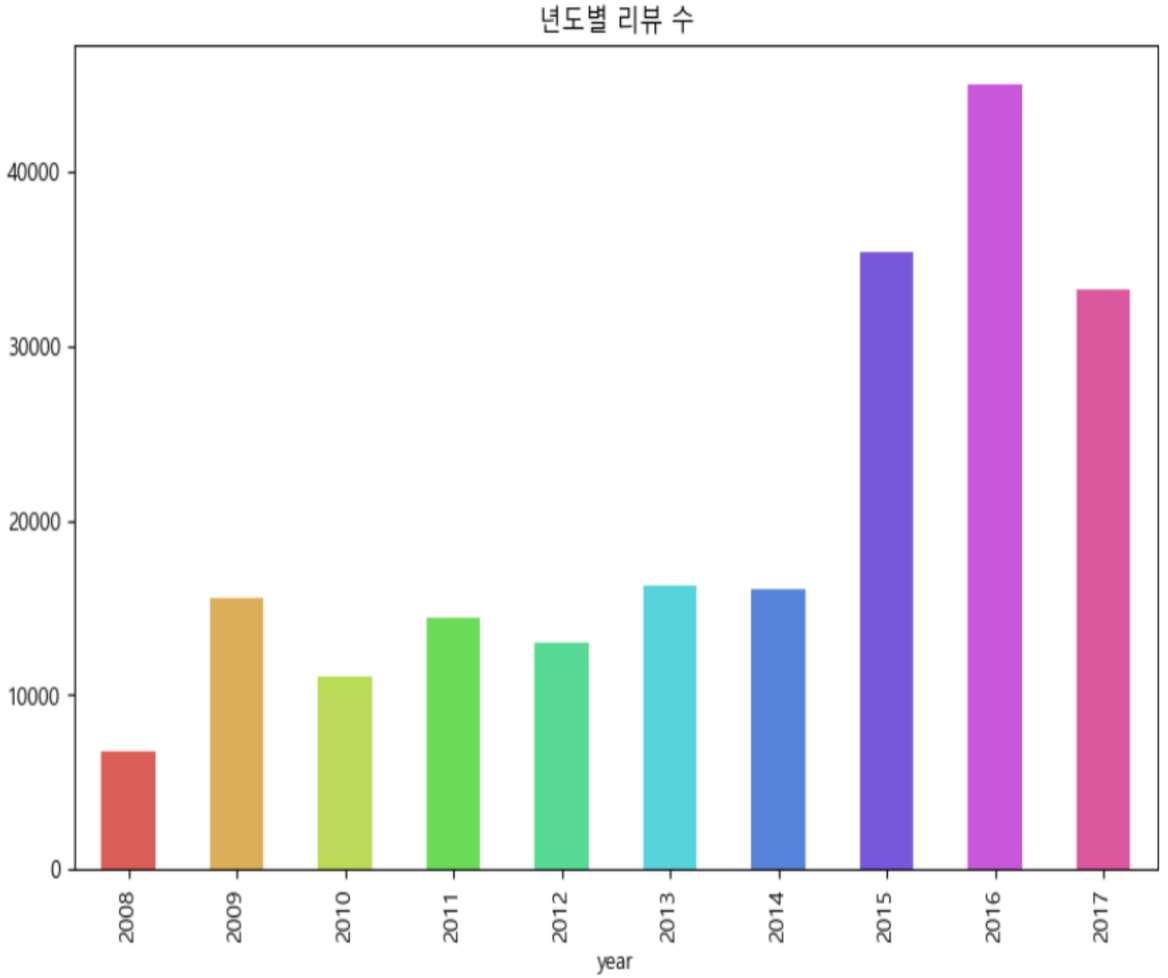
가장 리뷰수가 적은 의약품은 contrave(다이어트약) 이다.

=> 증상별 의약품 수는 피임이 3번째로 많았고, 리뷰수는 가장 많다.

'RATING'별 리뷰수 비율 파이차트

의약품을 복용한 환자들은 대부분의 긍정적인 점수를 줬다.
반면에, 3번째로 높은 비율을 가진 rating은 1점이다.
=> 환자들은 자신이 복용한 의약품이 매우 긍정적이거나
매우 부정적일때 리뷰를 작성한다.



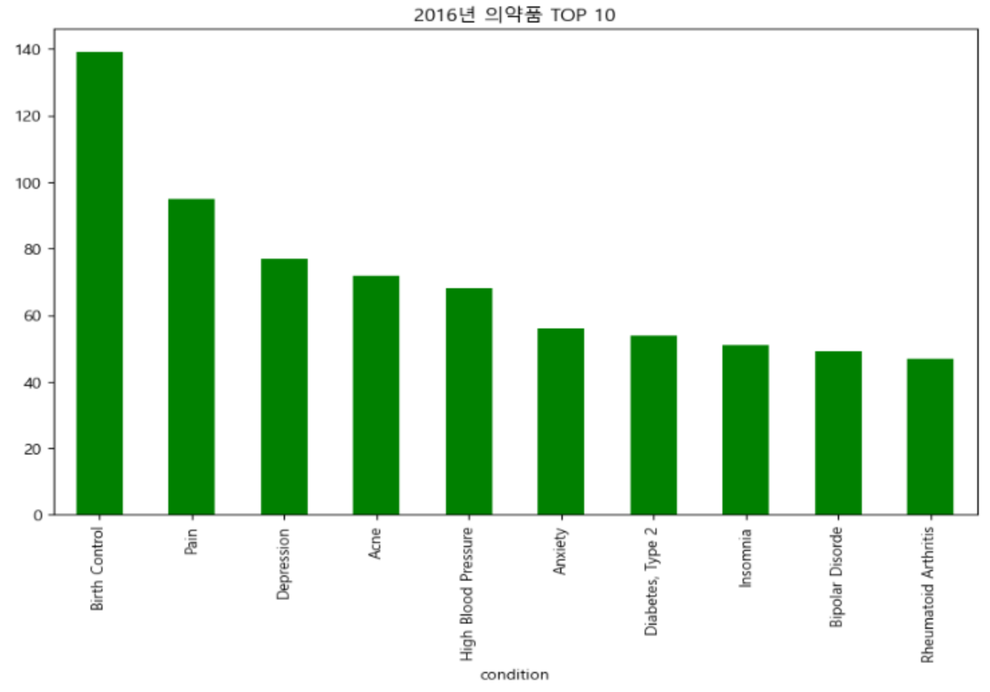


년도별리뷰

2014년 이후 리뷰의 양이 급증했다.

리뷰가 가장 많았던 년도는 2016년도이다.

=> 2016년도에 Birth Control(피임)이 가장 많이 복용되었다.



리뷰에 따른 WORDCLOUD

first : 처음 약을 복용한 환자들은 리뷰를 많이 쓰는 경향이 있다.

combination : 여러가지 약을 함께 복용하는 환자가 많다.

side + effect : 부작용에 대한 리뷰

N-GRAM

rating 점수에 따라 1~3점은 부정, 4~7점은 보통, 8~10점은 긍정으로 분류

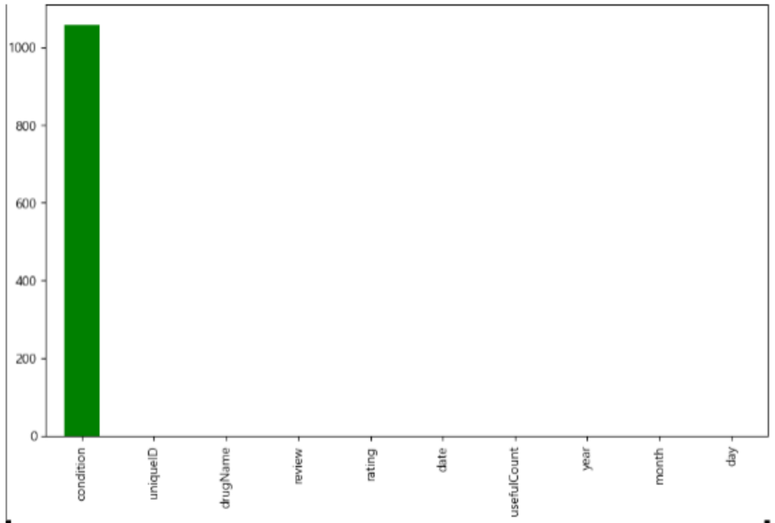
1 ~ 4 grams을 통해 어떤 말뭉치가 가장 감성을 잘 분류하는지 판단한다.

=> 4-gram에서 가장 차이가 뚜렷하다.

ex) 1~3점 : will never take again

4~7점 : started taking ~

8~10점 : side effects went away



#MISSING VALUE

각 컬럼별 결측치 확인

dropna를 통해 condition에 존재했던

결측치 제거

```python
# <span> 태그 삭제
all_list = set(df_all.index)
span_list = []
for i,j in enumerate(df_all['condition']):
    if '</span>' in j:
        span_list.append(i)
new_idx = all_list.difference(set(span_list))
df_all = df_all.iloc[list(new_idx)].reset_index()
del df_all['index']

#약물 추천이 목표이기 때문에, 증상에 대해 처방할 약품이 하나인 경우는 목표에 맞지 않기 때문에 제거한다.
df_condition = df_all.groupby(['condition'])['drugName'].nunique().sort_values(ascending=False)
df_condition = pd.DataFrame(df_condition).reset_index()

# 1개의 증상에 대한 처방 약물 종류가 1개인 경우의 약물이름 추출
df_condition_1 = df_condition[df_condition['drugName']==1].reset_index()

#해당 증상 제거
all_list = set(df_all.index)
condition_list = []
for i,j in enumerate(df_all['condition']):
    for c in list(df_condition_1['condition']):
        if j == c:
            condition_list.append(i)

new_idx = all_list.difference(set(condition_list))
df_all = df_all.iloc[list(new_idx)].reset_index()
del df_all['index']
```

#CONDITION

태그를 삭제한 후 증상에 대해

처방한 의약품이 1개인 경우 제거

```python
# 객체를 만들어 영어 어근 추출
stemmer = SnowballStemmer('english')

# 함수를 정의
def review_to_words(raw_review):

    # 1. HTML 태그 제거
    review_text = BeautifulSoup(raw_review, 'html.parser').get_text()

    # 2. 문자 이외 공백 치환, 문자열 소문자 변환
    letters_only = re.sub('[^a-zA-Z]', ' ', review_text)
    words = letters_only.lower().split()

    #stopwords 제거
    stops = set(stopwords.words('english'))
    meaningful_words = [w for w in words if not w in stops]

    #어근추출
    stemming_words = [stemmer.stem(w) for w in meaningful_words]

    return (' '.join(stemming_words))
```

#REVIEW

HTML 태그 제거

문자 이외 공백 치환, 문자열 소문자 변환

stopwords 제거

어근추출

딥러닝모델(N-GRAM)

모든 predict_sentiment를 2라고 값을 주었을 경우 accuracy가 0.611

=> 성능강화 필요

=> Havard 감성사전 데이터 이용

DL_PJT 2조

#CountVectorizer

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.pipeline import Pipeline

vectorizer = CountVectorizer(analyzer = 'word',
                             tokenizer = None,
                             preprocessor = None,
                             stop_words = None,
                             min_df = 2,
                             ngram_range=(4, 4),
                             max_features = 20000
                            )
vectorizer
```

```
CountVectorizer
CountVectorizer(max_features=20000, min_df=2, ngram_range=(4, 4))
```

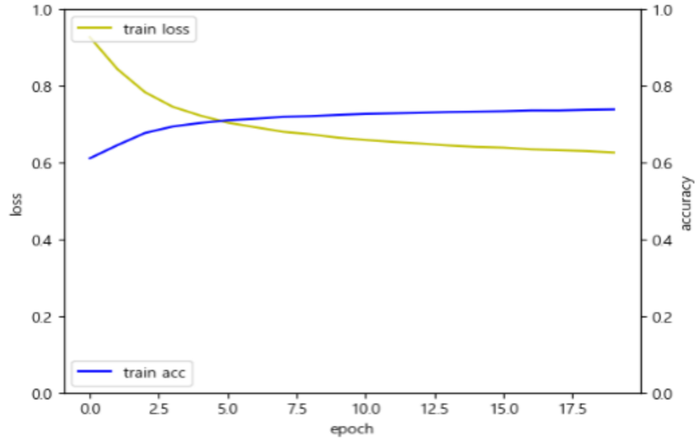
#Model Structure

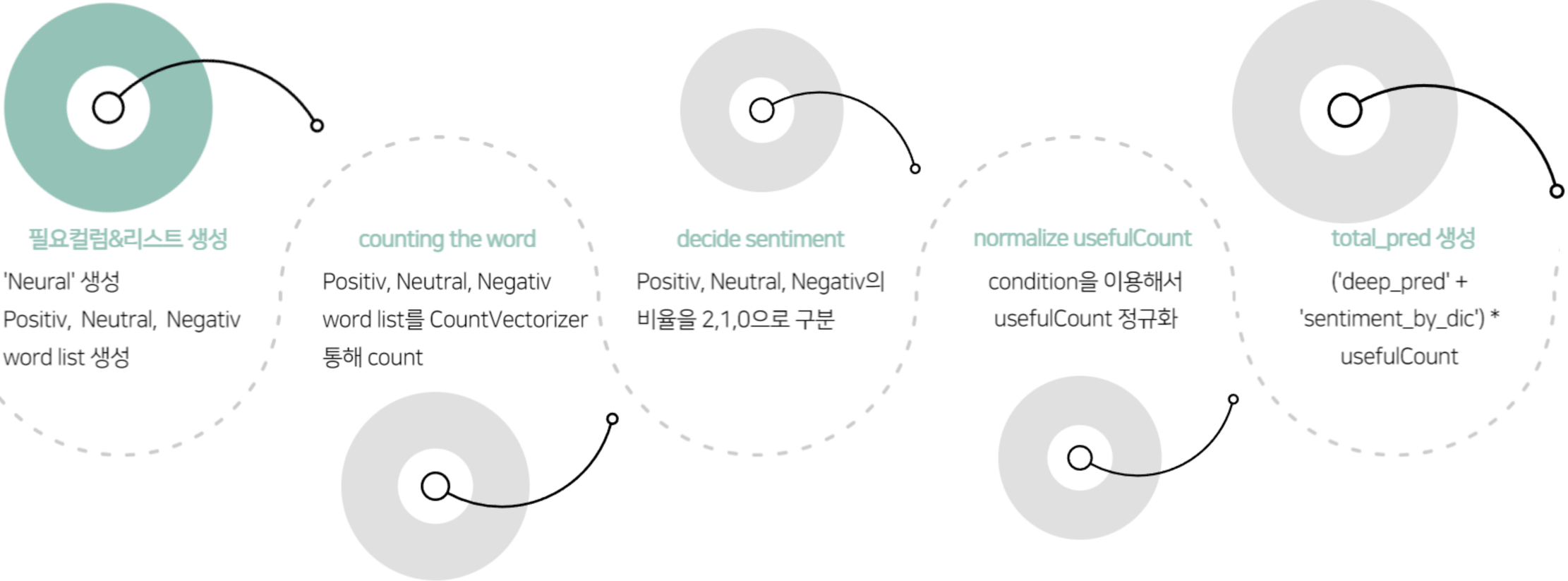
```
model = keras.models.Sequential()

model.add(keras.layers.Dense(200, input_shape=(20000,)))
model.add(keras.layers.BatchNormalization())
model.add(keras.layers.Activation('relu'))
model.add(keras.layers.Dropout(0.5))

model.add(keras.layers.Dense(300))
model.add(keras.layers.BatchNormalization())
model.add(keras.layers.Activation('relu'))
model.add(keras.layers.Dropout(0.5))

model.add(keras.layers.Dense(100, activation='relu'))
model.add(keras.layers.Dense(3, activation='softmax'))
```

#Model compile&Training

```
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

hist = model.fit(train_data_features, y_train, epochs=20, batch_size=16)

%matplotlib inline
import matplotlib.pyplot as plt

fig, loss_ax = plt.subplots()

acc_ax = loss_ax.twinx()

loss_ax.set_ylim([0.0, 1.0])
acc_ax.set_ylim([0.0, 1.0])

loss_ax.plot(hist.history['loss'], 'y', label='train loss')
acc_ax.plot(hist.history['accuracy'], 'b', label='train acc')

loss_ax.set_xlabel('epoch')
loss_ax.set_ylabel('loss')
acc_ax.set_ylabel('accuracy')

loss_ax.legend(loc='upper left')
acc_ax.legend(loc='lower left')

plt.show()
```

#Evaluation




필요컬럼&리스트 생성
'Neural' 생성
Positiv, Neutral, Negativ
word list 생성
counting the word
Positiv, Neutral, Negativ
word list를 CountVectorizer
통해 count
decide sentiment
Positiv, Neutral, Negativ의
비율을 2,1,0으로 구분
normalize usefulCount
condition을 이용해서
usefulCount 정규화
total_pred 생성
('deep_pred' +
'sentiment_by_dic') *
usefulCount

가치 제공

: 해당 프로젝트를 통해 제공하는 가치

RECOMMENDATION SYSTEM

condition과 review를 통한 의약품 추천 시스템

소비자들에게 의약품 정보 제공

관심이 있는 증상에 관하여 '의약품', '리뷰', 'total_pred' 확인


```python
def recommend_system(condition, review):
    print('Top 10 drug for {} (based on your condition & review)'.format(condition))
    condition_name = df_test[df_test['condition'] == condition]
    condition_value = condition_name[['condition', 'drugName', 'review_clean', 'total_pred']]
    condition_dataframe = condition_value.sort_values('total_pred', ascending=False)

    review_dataframe = condition_dataframe[condition_dataframe['review_clean'].str.contains(review)]
    review_top10 = review_dataframe.iloc[0:10]

    return review_top10
```

```python
recommend_system('Insomnia', 'fall asleep')
```

```
Top 10 drug for Insomnia (based on your condition & review)
```

	condition	drugName	review_clean	total_pred
45784	Insomnia	Trazodone	take month anxieti depress fall asleep mind co...	0.374526
28171	Insomnia	Trazodone	would feel sleepi go bed find still tri fall a...	0.347269
11079	Insomnia	Mirtazapine	use take good hour fall asleep night take arou...	0.335238
19782	Insomnia	Clonazepam	struggl moder insomnia year fall asleep wake h...	0.310408
27838	Insomnia	Ambien	sleep problem sinc colleg year old took normal...	0.286271
28453	Insomnia	Trazodone	suffer insomnia year tri ambien mg keep asleep...	0.266267
42914	Insomnia	Quetiapine	diagnos depress sever time gradual got older t...	0.233054
45142	Insomnia	Mirtazapine	start take mg bedtim month ago realli help fal...	0.199092
12945	Insomnia	Trazodone	trazodon work well u year ago start sleep poor...	0.184779
39898	Insomnia	Eszopiclone	taken lunesta mg around year take minut make s...	0.155204

```python
recommend_system('High Blood Pressure', 'blood pressur')
```

```
Top 10 drug for High Blood Pressure (based on your condition & review)
```

	condition	drugName	review_clean	total_pred
33472	High Blood Pressure	Ramipril	start altac blood pressur brought bp period yr...	0.865296
26876	High Blood Pressure	Amlodipine	take amlodipin norvasc month norvasc amlodipin...	0.739427
37354	High Blood Pressure	Losartan	high blood pressur year first put enalapril hy...	0.619501
35354	High Blood Pressure	Metoprolol	medicin lower blood pressur side effect horrib...	0.569323
19908	High Blood Pressure	Losartan	switch losartan lisinipril caus horrifi cough ...	0.541204
33714	High Blood Pressure	Clonidine	high blood pressur sinc start uncontrol blood ...	0.490702
43148	High Blood Pressure	Metoprolol Tartrate	medic keep hypertens control also seem effect ...	0.458378
34626	High Blood Pressure	Cozaar	year maintain blood pressur	0.447327
37124	High Blood Pressure	Norvasc	norvasc caus extrem leg ankl edema improv hype...	0.432661
32416	High Blood Pressure	Valsartan	work great blood pressur howev stop medicin se...	0.432140

THANK YOU

DL_PJT 2조

#김정빈 #박건우