Project Presentation 3

18102082 Su Kyoung Oh
18102092 Won Ryeol Jeong
16102284 Sung Ho Lee

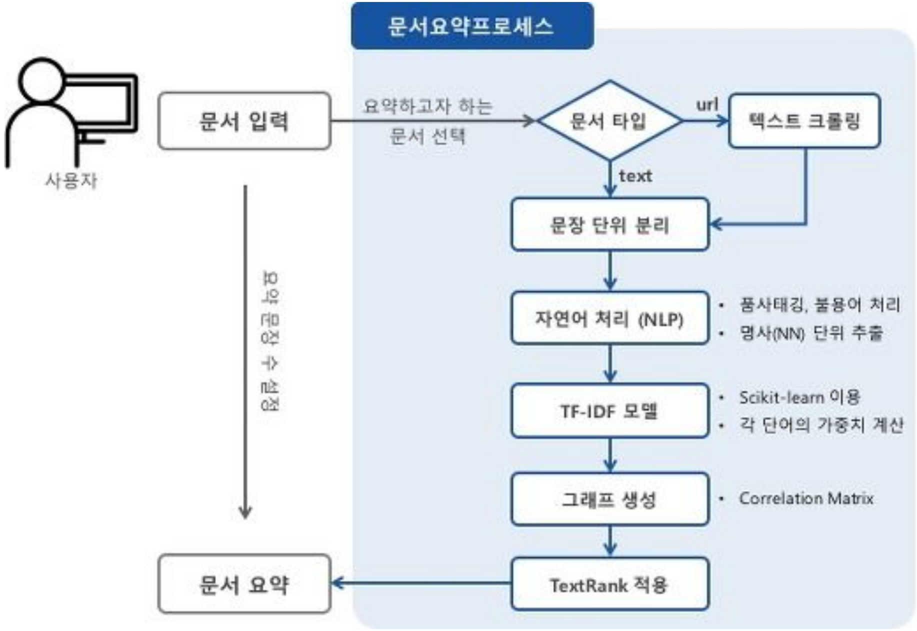
Contents

1. TextRank Algorithm
2. Variable Visualization
3. Train Model(Naive Bayes)
4. Evaluate Model(Naive Bayes)

TextRank Algorithm

Text-rank algorithm was introduced for preprocessing of sentimental analysis.

We tried to tokenize the reviews and classify the parts, and then applied the method of weighting the words such as nouns, adjectives, verbs, etc., which are expected to have a significant impact on the ratings according to their relative importance.

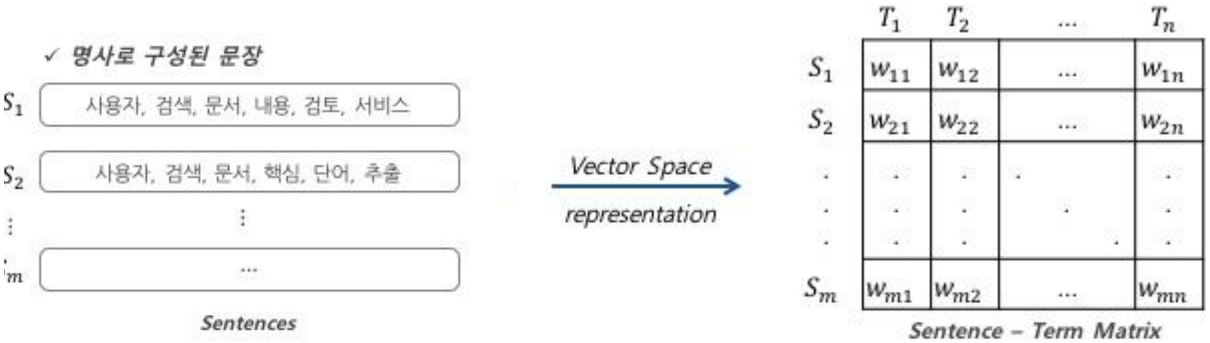


TextRank Algorithm

TF-IDF

: A statistical weight that indicates how important a word is within a particular document when there is a group of documents consisting of multiple documents.

=> Using this approach, when a particular word that is rarely found in other reviews appears in this specific review, it can be said that this particular word represents the review.



S_1: 사용자, 검색, 문서, 내용, 검토, 서비스

S_2: 사용자, 검색, 문서, 핵심, 단어, 추출

⋮

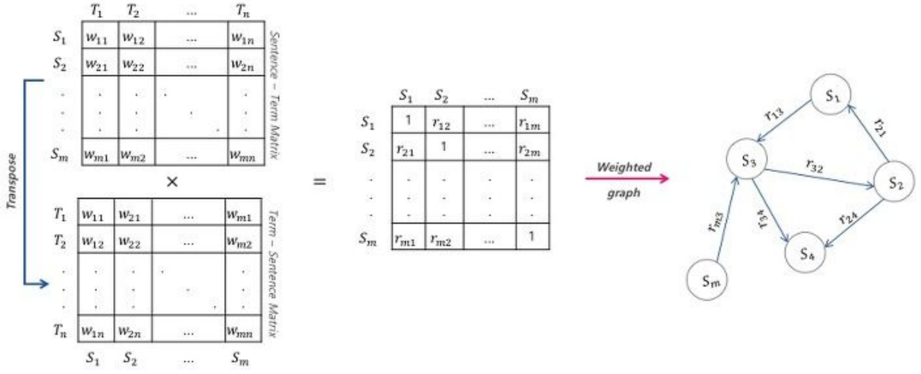
S_m: …

$$\begin{array}{c|cccc} & T_1 & T_2 & \cdots & T_n \\ \hline S_1 & w_{11} & w_{12} & \cdots & w_{1n} \\ S_2 & w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_m & w_{m1} & w_{m2} & \cdots & w_{mn} \end{array}$$

TextRank Algorithm

Weighted Graph

: The Tf-Idf matrix obtained previously and its transposition matrix will be multiplied to obtain the correlation matrix. Through the correlation matrix, the weighted graph between reviews or words can be expressed. This matrix can also be thought of as 'Adjancey Matrix', so a graph made of nodes and edges can be created as shown below. (T is word, S is review)

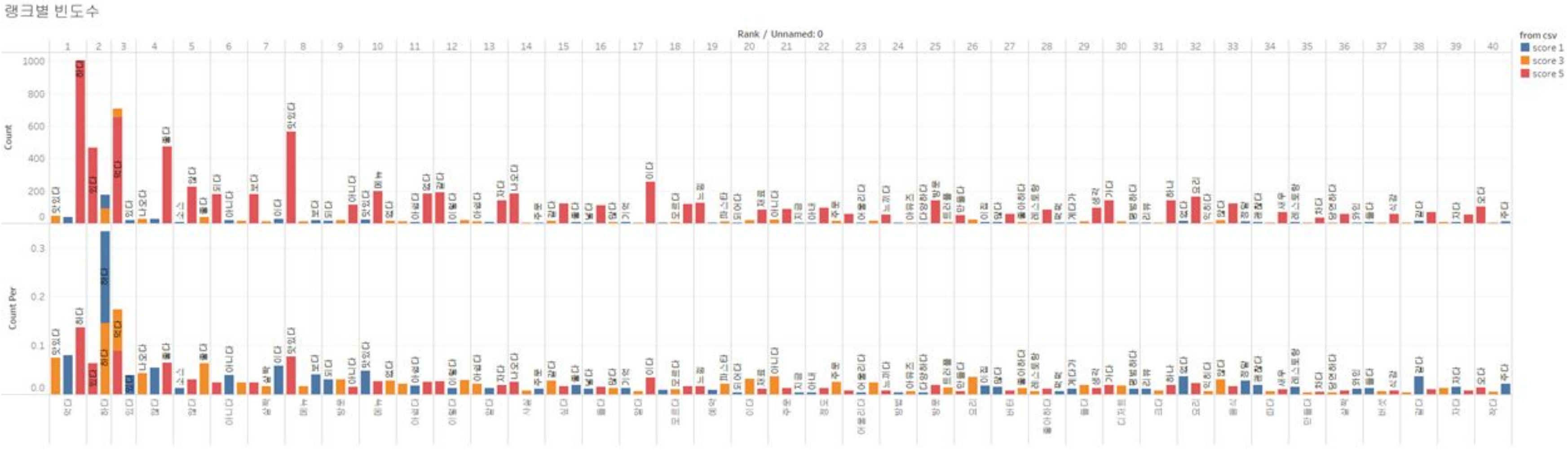


$$R = A^TA = \begin{bmatrix} A_1^T \\ A_2^T \\ \dots \\ A_n^T \end{bmatrix} \begin{bmatrix} A_1 & A_2 & \dots & A_n \end{bmatrix} = \begin{bmatrix} A_1^TA_1 & A_1^TA_2 & \dots & A_1^TA_n \\ A_2^TA_1 & A_2^TA_2 & \dots & A_2^TA_n \\ \dots & \dots & \dots & \dots \\ A_n^TA_1 & A_n^TA_2 & \dots & A_n^TA_n \end{bmatrix}$$

TextRank Algorithm

： We can apply the TextRank algorithm using the weighted graph of the words generated previously． Through the TextRank algorithm， we will sort the words in order of highest Ranking value and create a set of words that represent the reviews of that class．

Variable Visualization



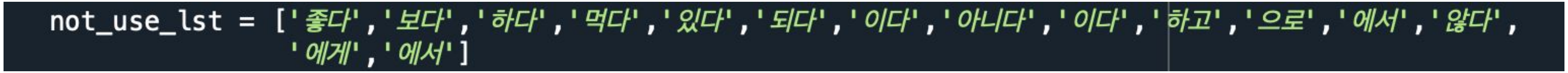
Rank별로 구분된 각 Unnamed: 0에 대해 Count의 합계 및 Count Per의 합계입니다. 색상이 from csv의 세부 정보를 표시합니다. 마크에 레이블이 Unnamed: 0에 의해 지정되었습니다. score 1, score 3 및 score 5을(를) 유지하는 from csv에 대한 뷰가 필터링되었습니다.

Frequency of words between each score class by ‘rank’

Variable Visualization

As a result of applying the text ranking algorithm without any preprocessing, words with significant insights tend to be buried because of meaningless data such as '하다', 좋다' and '이다'.

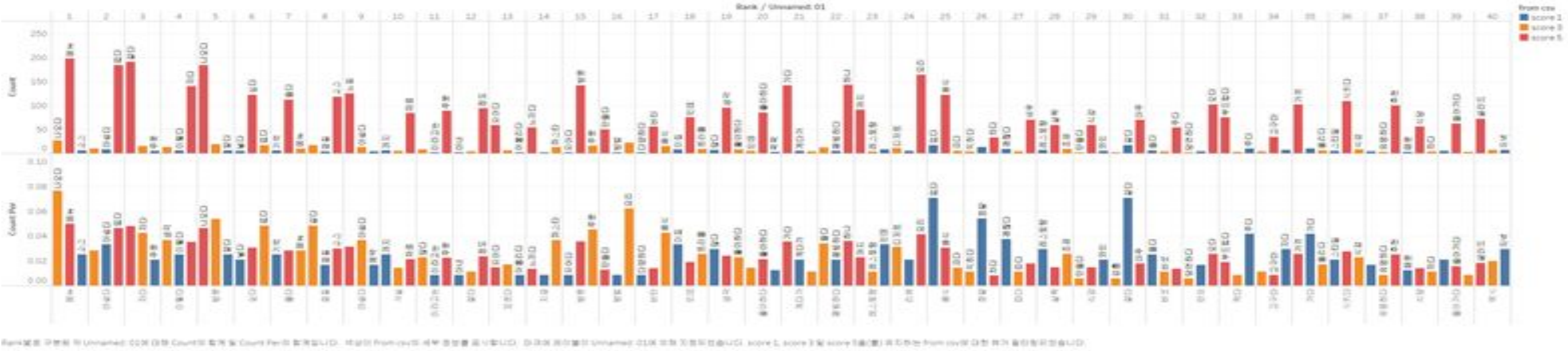
So we'll take these words out and reapply the text ranking algorithm again!



```
not_use_lst = ['좋다','보다','하다','먹다','있다','되다','이다','아니다','이다','하고','으로','에서','않다',
               '에게','에서']
```

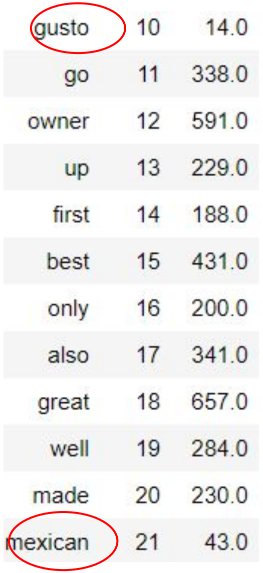
<Examples of Unnecessary words>

Variable Visualization



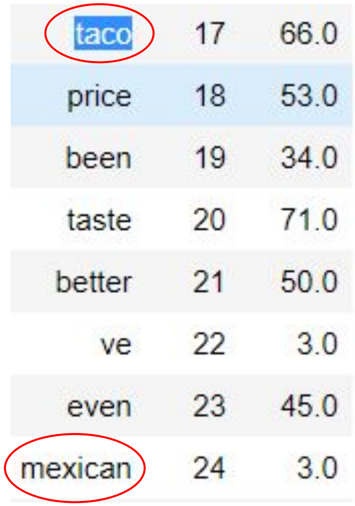
Frequency of words between each score class by ‘rank’

Variable Visualization



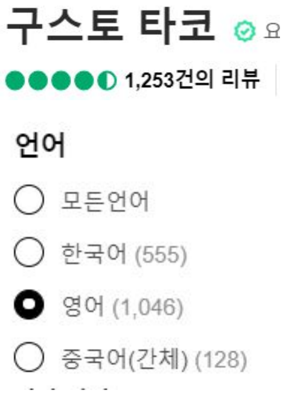
gusto	10	14.0
go	11	338.0
owner	12	591.0
up	13	229.0
first	14	188.0
best	15	431.0
only	16	200.0
also	17	341.0
great	18	657.0
well	19	284.0
made	20	230.0
mexican	21	43.0

Rank of Evaluate_Good



taco	17	66.0
price	18	53.0
been	19	34.0
taste	20	71.0
better	21	50.0
ve	22	3.0
even	23	45.0
mexican	24	3.0

Rank of Evaluate_Bad



구스토 타코

●●●●◐ 1,253건의 리뷰

언어

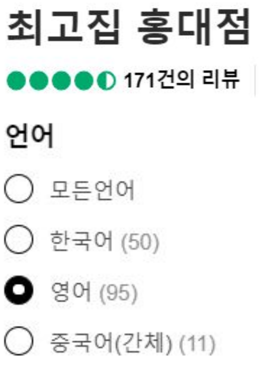
○ 모든언어

○ 한국어 (555)

● 영어 (1,046)

○ 중국어(간체) (128)

Rank 1 of HongDae Restaurant



최고집 홍대점

●●●●◐ 171건의 리뷰

언어

○ 모든언어

○ 한국어 (50)

● 영어 (95)

○ 중국어(간체) (11)

Rank 2 of HongDae Restaurant

It seems necessary to limit the number of reviews for a each store

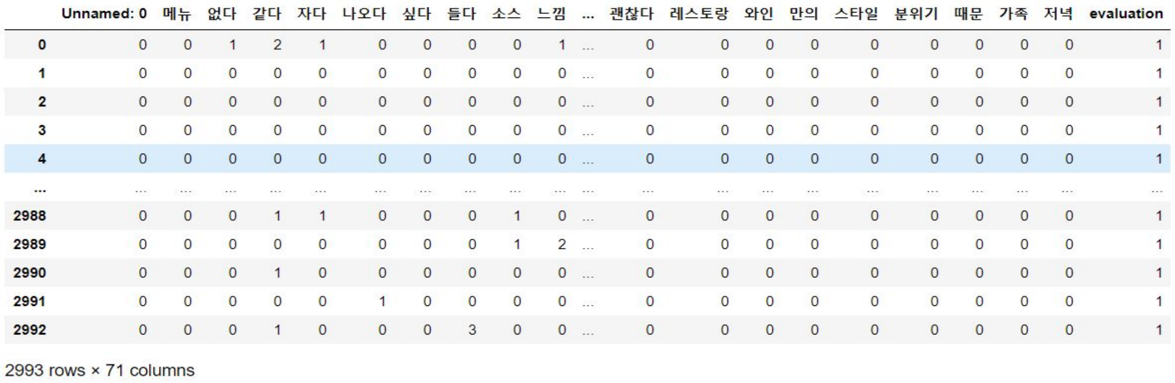
Train Model(Naive Bayes)

Naive Bayes

: A machine learning technique generally used for text classification and is based on the theorem of bays.

$$\begin{aligned}\log p(C_k|\mathbf{x}) &\propto \log\left(p(C_k)\prod_{i=1}^{n}{p_{ki}}^{x_i}\right)\\ &= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki}\\ &= b + \mathbf{w}_k^\top \mathbf{x}\end{aligned}$$

Train Model(Naive Bayes)

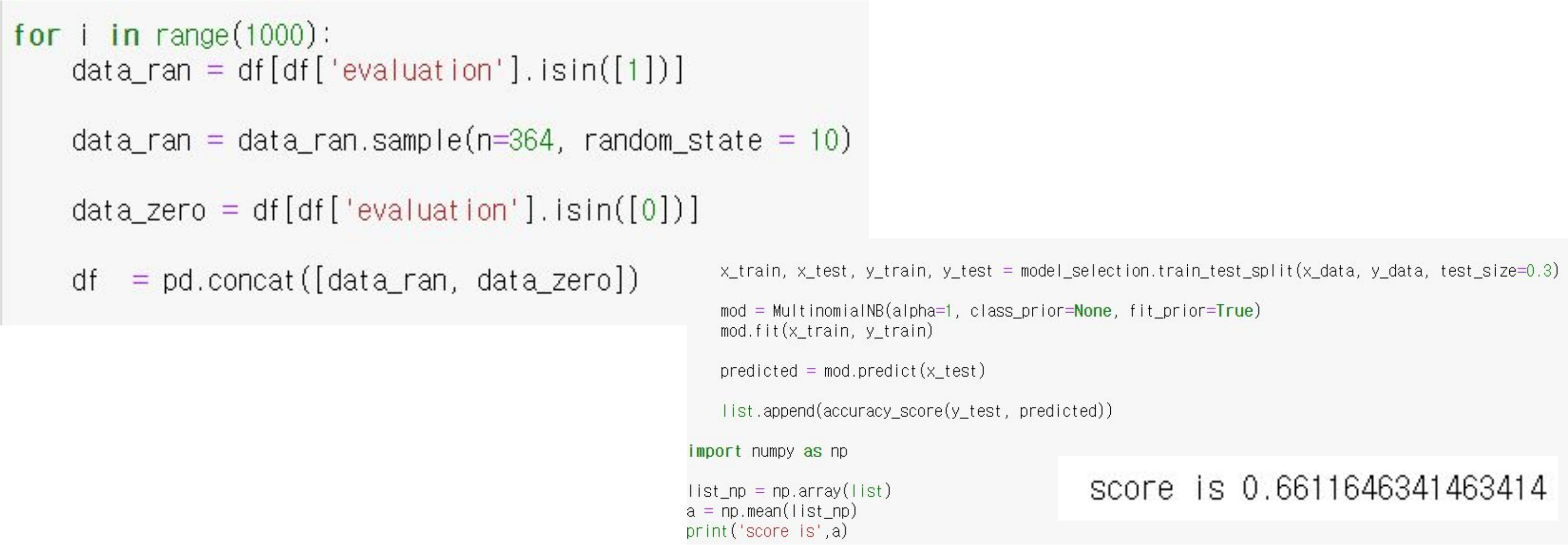


	Unnamed: 0	메뉴	없다	같다	자다	나오다	싶다	들다	소스	느낌	...	괜찮다	레스토랑	와인	만의	스타일	분위기	때문	가족	저녁	evaluation
0	0	0	1	2	1	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2988	0	0	0	1	1	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	1
2989	0	0	0	0	0	0	0	0	1	2	...	0	0	0	0	0	0	0	0	0	1
2990	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
2991	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
2992	0	0	0	1	0	0	0	3	0	0	...	0	0	0	0	0	0	0	0	0	1

2993 rows × 71 columns

TF matrix was created based on selected words with text rank algorithm. The label value (evaluation) is 1 for a positive review and 0 for a negative review.

Evaluate Model(Naive Bayes)



```
for i in range(1000):
    data_ran = df[df['evaluation'].isin([1])]

    data_ran = data_ran.sample(n=364, random_state = 10)

    data_zero = df[df['evaluation'].isin([0])]

    df  = pd.concat([data_ran, data_zero])
```

```
    x_train, x_test, y_train, y_test = model_selection.train_test_split(x_data, y_data, test_size=0.3)

    mod = MultinomialNB(alpha=1, class_prior=None, fit_prior=True)
    mod.fit(x_train, y_train)

    predicted = mod.predict(x_test)

    list.append(accuracy_score(y_test, predicted))

import numpy as np

list_np = np.array(list)
a = np.mean(list_np)
print('score is',a)
```

```
score is 0.6611646341463414
```

Evaluate Model(Naive Bayes)

- The number of rows in the data is insufficient.
- The target ratio of the data is tilted to one side.
- Just words can tell if it's positive or negative.

- Carry out further crawling to solve the problem of data imbalance and data shortages.
- Sentimental scores of each sentence classified by review attribute can be additionally calculated to infer emotional scores for sentences in each review by assessment

THANKS
FOR
LISTENING