데이터분석캡스톤디자인

7주차 수행보고

Khupid 조

산업경영공학과 김동혁
관광학과 류연주
산업경영공학과 유정수

데이터 통합

1. 갈 곳(Go), 볼 것(Watch), 먹을 것(Eat) 각자 크롤링 한 data의 필드를 지난번에 정의한 필드로 통합
2. 결측치 제거 및 수정
3. 리뷰 통합 data를 text파일로 별도 저장 -> 태그 추출할 때 이용

```
eat_info.columns
```

```
Index(['name', 'address', 'category', 'main_mn', 'price', 'opng_tm', 'rating',
       'rvw_cnt', 'tags'],
      dtype='object')
```

```
new.columns
```

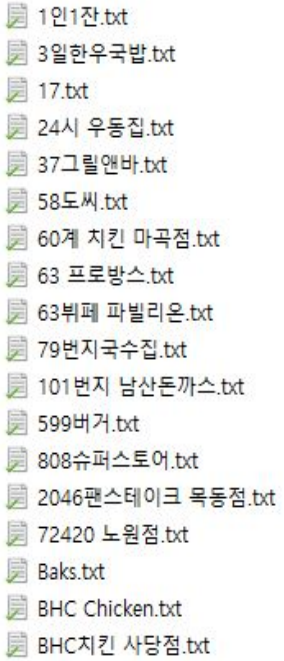
```
Index(['p_id', 'p_name', 'address', 'district', 'monday', 'tuesday',
       'wednesday', 'thursday', 'friday', 'saturday', 'sumday', 'category',
       'price', 'p_rate'],
      dtype='object')
```

결측치 제거 및 수정

```
# 결측치 계산
go_info.isnull().sum()
```

```
name          28
address       28
category    2360
main_mn      132
price        213
opng_tm      201
rating        30
rvw_cnt       30
tags          37
dtype: int64
```

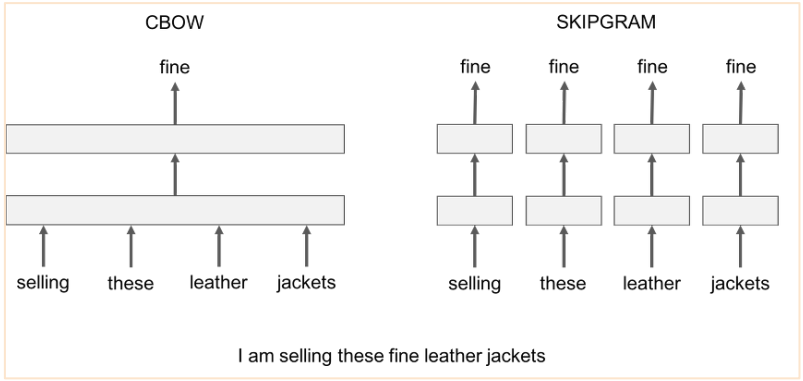
*장소별 리뷰 통합 **data** 구축*



FastText model 스터디

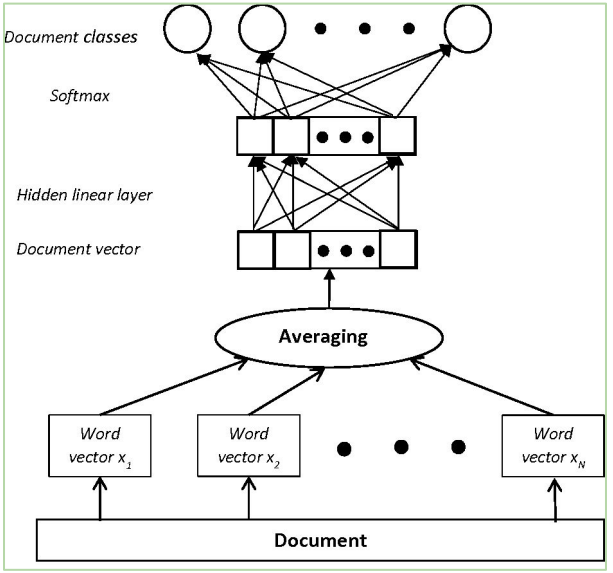
pre-trained 모델 이해 (한국어, 영어 모델 모두)

- CBOW
- 300 dimension
- window size: 5 and 10 negatives



Output: Feature Vector 이해

- word vector의 평균을 내서 input으로 넣어줌
- 은닉층의 노드 값을 가져와서 벡터로 나타냄



FastText model과 형용사 클러스터링

	comfortable	only	favorite	little	weak	first	want	female	overall	other	delicious
0	-0.000428	0.001228	-0.000180	0.000972	-0.000185	0.000038	0.000038	0.000261	0.000047	-0.000214	-0.000271
1	0.000183	0.000173	0.000307	0.001191	0.000233	0.000596	0.000923	0.000445	-0.000409	0.000150	-0.000466
2	0.000478	-0.000791	0.000173	-0.000499	0.001446	0.001369	0.000924	-0.000424	-0.000552	-0.000090	-0.000167
3	-0.000368	-0.001272	0.000377	0.000133	0.001009	0.000273	-0.000876	0.000082	-0.000340	0.000455	-0.000392
4	-0.000772	-0.000109	0.000543	-0.000014	0.000653	-0.000085	-0.000363	0.000610	-0.000167	-0.001110	-0.000231
...	...	...	...	...	...	...	...	...	...	...	...
295	0.000125	0.000061	-0.000115	0.000619	-0.001195	-0.001046	0.000390	0.000690	0.000301	0.000077	0.000119
296	-0.000071	0.000744	-0.000208	-0.000682	0.000181	0.000426	-0.000330	0.000058	-0.000014	0.000821	-0.000109
297	-0.000125	-0.000324	-0.000019	-0.000676	0.000222	0.000235	-0.000293	0.000132	-0.000472	-0.000355	0.000125
298	-0.000319	0.000006	-0.000451	0.000197	-0.000604	-0.000112	-0.000614	0.000107	-0.000610	0.000303	-0.000287
299	0.000050	-0.000063	-0.000577	0.000161	-0.000171	0.000441	0.000525	0.000271	0.000296	0.000548	0.000539

```
Y[Y['kmeans_id'] == 0].index
```

```
Index(['good', 'goodAnd', 'good^^', 'goodTo', 'good?', 'goodIt', 'good:-)'], dtype='object')
```

```
Y[Y['kmeans_id'] == 1].index
```

```
Index(['Warm', 'deep', 'wait', 'sleep', 'studyDeep', '"I', 'mic', 'intent'], dtype='object')
```

```
Y[Y['kmeans_id'] == 2].index
```

```
Index(['only', 'delicious', 'great', 'song', 'new', 'quiet', 'neat', 'dirty',
       'soggy', 'cold', 'hard', 'old', 'club', 'strange', 'shade', 'hot',
       'real', 'sound', '감사합니다', 'it'd', 'Yeongdeungpo', 'own', 'dusty', 'rid',
       'bike', 'eat', 'delicious*^^*', 'scary', 'senior', 'cool', 'dry',
       'remix-board', 'solid', 'central', 'weird', 'angry', 'three-hour',
       'red', 'minor', 'Myeong-dong', 'dirty;', 'malicious', 'Myeonmok-dong',
       'shitlol', 'mini', 'serious', '*^^*', 'deliciousFresh', 'lol', 'vague',
       'in?', 'dogs', 'Yeontral', 'mean', 'extra', 'cat', 'Clean', 'kind^^',
       'girl', 'deliciousNo', 'sexual', 'dirty-looking', 'sour', 'it?',
       'misogynistic', 'dinga'],
      dtype='object')
```

	긴밀하다	커다랗다	알맞다	삼삼하다	한결같다	눅눅하다	궁상맞다	복되다	우세하다	고약하다	...	어둑하다	엄밀하다
0	-0.000616	-0.000736	-0.001491	-0.001006	-0.000099	-0.000069	-0.001113	-0.000863	-0.000611	-0.000721	...	-0.000091	-0.00030
1	0.000533	0.000075	0.000391	-0.000457	0.000262	0.000014	0.001086	0.000544	-0.000148	-0.000420	...	-0.001072	-0.00113
2	-0.000339	0.000275	-0.000591	-0.000952	0.000549	-0.000961	0.001196	-0.000041	-0.000064	-0.000467	...	0.000119	-0.00068
3	0.000512	-0.000356	-0.000500	-0.000438	0.000390	-0.001120	0.000144	0.000694	0.000818	0.000098	...	0.000456	-0.00100
4	-0.001173	0.000177	-0.001402	-0.000714	0.000130	-0.000494	-0.000464	-0.001017	-0.000485	-0.001191	...	-0.000304	-0.00025
...	...	...	...	...	...	...	...	...	...	...	...	...	...
295	-0.001237	0.000938	-0.000158	0.000697	-0.000589	-0.000379	0.000483	0.001353	0.000515	0.000145	...	-0.000359	0.00018
296	-0.000204	0.000475	-0.000283	-0.001503	0.000210	-0.000091	0.000500	0.000400	-0.000230	-0.000401			
297	-0.000299	0.000860	0.000859	0.000676	0.000251	-0.001028	-0.000296	-0.000269	-0.000325	-0.000254			
298	-0.000950	0.000068	0.000398	-0.000279	0.000634	-0.000181	0.000949	0.001193	-0.001301	0.000395			
299	-0.000594	-0.000642	0.000012	0.000234	0.000639	0.000161	0.000294	-0.000961	-0.000350	0.000420			

```
array([ 6,  3, 11,  4,  5, 14, 11,  7, 11, 14,  2,  2, 11,  6,  1, 12,  7,
        5,  9,  2,  0,  7,  2, 11,  5, 11, 11, 11, 11,  3,  1,  1,  2, 12,  6,
       14, 14,  5,  5,  2,  1,  1,  1,  2,  5,  2,  8, 12, 14,  2,  6,  7,
        9,  3,  2,  0,  2,  0,  6, 11, 13, 13,  7, 13, 13,  7,  9,  2,  4,
        7,  5, 14,  0,  7, 11, 12, 14,  1,  6,  2,  9,  6, 11, 11,  2,  0,
        0, 14,  0,  9, 12, 11,  2,  0,  7, 10,  7,  7, 11, 14,  4,  7,  3,
        2,  2,  4,  7,  7, 13,  7, 11,  1,  2, 10,  7,  4,  2, 11, 12, 11,
        2, 11, 13,  5, 14,  3,  5,  3,  2, 11,  1,  2, 11,  2,  0,  6,  2,
       13,  2,  1,  5,  2,  4,  9,  2,  7,  3, 11, 14,  9,  7,  2, 14,  5,
       13,  7, 13,  6, 10,  3, 11,  1,  2,  6,  6, 10,  1,  2, 11, 14, 11,
        6,  2, 14, 11,  2,  2, 14, 11, 13,  2,  2,  2, 12, 14, 14,  8,  2,
        2, 12,  3,  0,  6, 11,  5,  7,  6, 14,  7, 13, 14,  6,  6,  0, 11,
        2,  6,  7,  7,  7,  6, 11,  2,  6, 12, 14, 14,  0,  0, 13,  7, 11,
       14, 11,  1,  6,  2, 12,  5, 14,  7,  1, 14,  2,  0,  4, 12,  5,  2,
        5,  2,  1,  2,  2, 14,  4, 14, 14,  2,  4, 11,  7,  7,  4,  4,  5,
        2, 11,  7,  6, 11,  4,  7,  7,  7, 11,  2,  2,  5, 11,  2, 14,  1,
        4,  3,  2, 14,  6,  0,  3,  6,  5,  1,  4,  1,  2,  5, 11, 11, 11,
        2, 11, 14, 13,  8,  0, 11,  6, 11, 10,  1])
```

```
{6: 24,
 3: 11,
 11: 42,
 4: 14,
 5: 19,
 14: 30,
 7: 32,
 2: 54,
 1: 19,
 12: 11,
 9: 7,
 0: 16,
 8: 3,
 13: 13,
 10: 5}
```

긴밀하다
부패하다
뻣뻣하다
집요하다
황당하다
친밀하다
청아하다
경솔하다
복잡다단하다
즐비하다
걸걸하다
부지런하다
반하다
충분하다
무능하다
탁월하다
경건하다
올바르다
정통하다
묘연하다
덜하다
타락하다
무뚝뚝하다
긴급하다

이번주 문제점

1. 장소마다 다른 형식의 data(ex. 매월 첫,셋째주 화요일 휴무 등)로 인한 전처리의 어려움

2. 리뷰 data가 없거나 부족한 장소들이 다수 발생
   a. 다른 플랫폼에서 추가적인 크롤링
   b. 그래도 데이터가 없으면 그 장소는 제거

3. 한국어 -> 영어로 바꾸는 과정에서 품사 태깅이 깨지는 문제
   a. 영어 리뷰 전처리

4. 같은 모양으로 클러스터링 되는 문제 (sleep, deep / nasty, wasty 등등)
   a. 전처리 후 재확인

다음주에 해야할 일

1. 한국어 -> 영어 리뷰 번역 과정에서
2. 영어 리뷰 전처리 후 한국어와 비교해서 클러스터링 성능 비교
3. 태그 사전(matrix) 구축
4. Collaborative Filtering 모델 스터디