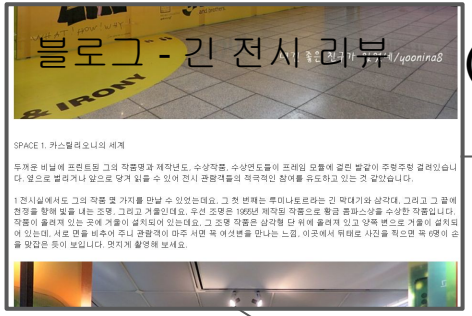
데이터분석캡스톤디자인

6주차 수행보고

Khupid 조

산업경영공학과 김동혁
관광학과 류연주
산업경영공학과 유정수

형태소 분석



(2) 중요 문장 계산 후 요약

38	0.05179020833613878	그럼전시장보다 훨씬 더 생생한 감이 있는 곳이 디자인 전시장 같습니다 .
13	0.05128247629951667	동시대 활동한 작가들의 작품도 일부 전시되어 있습니다 .
19	0.04251123001765623	전시장 가에 있는 철골에는 그가 디자인한 작품들에 대한 설명과 실물을 , 중앙에 있는 철골에는 붉은색 상자 안에 호기심이 많아 남의 집 서랍을 열어다 보는 듯한 느낌의 작품들이 전시되어 있어요 .
3	0.039387673568825436	전시장 구성은 SPACE 1~5 까지 다섯 섹션으로 구분되어 있는데 , 구역마다 그의 작품의 특색이 명확하게 표현되어 있어 전시되어 있는 점이 좋았습니다 .
0	0.03821818353286199	예술의 전당 한가람미술관에서 열리고 있는 이탈리아 디자인의 거장 카스틸리오니 전을 드디어 다녀왔습니다 .
22	0.037945625459128 15	그도 뒤상과 같은 레디메이드 작품을 여럿 선보이고 있습니다 .
31	0.03742567135649523	34 점의 작품을 사방의 거울을 통하여 보니 정말 나무가 빽빽한 숲속에 있는 느낌이 들죠 ?
27	0.03421513498038972	포스터의 숲 이번 전시회가 특별한 두 가지 의미를 가지고 있는데 , 하나는 아시아 최초의 카스틸리오니 단독전이라는 점이고 .
23	0.0337524825511475 86	아이콘 이 전시실에서는 카스틸리오니 형제에게 국제적 명성을 안겨준 제품들이 전시되어 있습니다 .
34	0.03234672270929389 6	세상은 그의 작품을 놓고 ' 평범함의 승리 " 라고 한다고 합니다 .
1	0.03173824892473695	이번 달 26 일까지이니 관심 있는 분 서두르세요 .
20	0.03084813364330009	눈 표시된 상자 안의 세계는 어쩌면 이번 전시장에서 상상으로 볼 수 있는 그의 건축물 디자인을 감상할 수 있는 코너였을 겁니다 .
18	0.03024444291892078 3	그래도 전시니 약간의 디자인 요소를 가미했는데 , 그것이 뭔지 알아맞혀 보라고 도슨트는 얘기합니다 .
2	0.02907994515746934	도슨트 해설도 놓치지 마시고요

(1) Okt (Open Korean text) 형태소 분석

```
In [6]: def get_tags_to_json(res):
            review_str = ''

            for review in review_df[review_df['res_name']==res]['review']:
                review_str+=str(review)

            okt = Okt()
            tags_dump = okt.pos(review_str, stem=True)
            tags = []

            for word in tags_dump:
                if word[1] == 'Adjective':
                    tags.append(word[0])

            tags = '/'.join(tags)
            # text 파일로 저장
            with open('../data/review_tags/'+res+'.txt','w') as make_file:
                make_file.write(tags)
```

형용사 추출

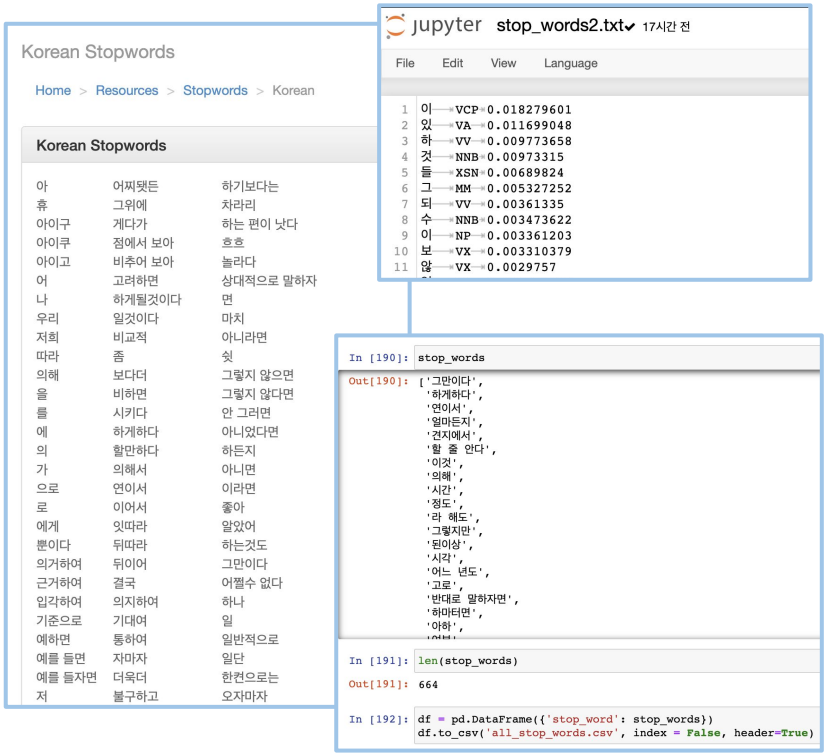
<맛집>	<전시>
'건강하다',	'다르다',
'두드러지다',	'이렇다',
'무관심하다',	'궁금하다',
'서툴다',	'좋다',
'처참하다',	'붉다',
'흡사하다',	'많다',
'소란하다',	'새롭다',
'필요없다',	'소소하다',
'아담하다',	'어떻다',
'야박하다',	'야하다',
'쉽다',	'선하다',
'독하다',	'특별하다',
'듬뿍하다',	'빽빽하다',
'적당하다',	'평범하다',
'분잡하다',	'자세하다',
'얇다',	'평범하다',
'닙데데하다',	'생생하다',
'토실하다',	'평범하다',

중요 문장에서 형용사 추출

(('좋았습니다', 'Adjective'),)	0.193432
(('두꺼운', 'Adjective'),)	0.193432
(('같았습니다', 'Adjective'),)	0.193432
(('같으면', 'Adjective'),)	0.193432
(('어려웠을', 'Adjective'),)	0.193432

불용어 처리 & 영어 번역 시도

(1) 한글 불용어 사전 구축



https://www.ranks.nl/stopwords/korean
https://bab2min.tistory.com/544

(2) 한글 -> 영어 (papago 번역)

review_en	adv_en
It's the best. It's comfortable. He's gone. Th...	[best, comfortable, only, favorite]
Heesuya Hyosun king pork cutlet Ssangmun No.1	[]
It was a little duhol. It's a little weak. Don...	[little, little, weak, first, wrong, female, o...
A place where you can drink and enjoy bowling ...	[delicious, good, cheap, better, good, many, c...
If you find a way to walk, you will be guided ...	[prettier, better, best, beautiful]
...	...
There are a lot of fun and diverse board games...	[diverse, great]
The cattle garden in the city center...Hidden ...	[best]
It was clean and nice on the weekend! Chicken...	[clean, nice, delicious, many, best, clean, de...
NaN	[]
Older people, dirty because they have a lot of...	[Older]

(3) Tokenizing -> 형용사 추출

```
('clean', 199)
('comic', 98)
('bad', 90)
('quiet', 78)
('new', 71)
('comfortable', 70)
('cheap', 70)
('expensive', 67)
('hard', 62)
('neat', 61)
```

Fasttext를 이용한 단어 벡터화

***pre-trained** 한국어 모델*

```
In [9]: ko_model = models.fasttext.load_facebook_model('cc.ko.300.bin.gz')
```

```
In [11]: ko_model.wv.word_vec('운치있는')
```

```
Out[11]: array([ 3.23515316e-03, -1.44122355e-03, -8.17747638e-02,  4.61231880e-02,
                 6.49930760e-02, -1.86859816e-03,  3.08580901e-02,  3.63530256e-02,
                 4.65812832e-02,  3.09575372e-03, -4.47942875e-02, -1.39526548e-02,
                -9.68437642e-04,  5.35678901e-02,  1.20583080e-01,  5.72555773e-02,
                 3.45472582e-02, -2.87518445e-02,  1.74929295e-02, -2.02273875e-02,
                -4.26698886e-02,  2.56524589e-02,  1.53578512e-04,  1.47120044e-01,
                -9.23347250e-02,  2.25054383e-01,  1.42958090e-02, -3.28313969e-02,
                 7.17691844e-03, -3.52043323e-02, -8.20247531e-02,  9.34616756e-03,
                -1.27324024e-02, -3.66074927e-02, -1.87187623e-02, -8.31611380e-02,
                 3.57123390e-02, -1.97619572e-02,  3.68778743e-02,  9.15034022e-03,
                -2.39420845e-03,  1.74316410e-02, -3.91573645e-02, -1.21175172e-02,
                -1.50920032e-02,  1.48614153e-01, -2.31579691e-02, -6.09047071e-04,
                 2.43108217e-02, -5.14785685e-02, -2.24889815e-02,  7.58593110e-03,
                -9.17353705e-02,  3.79088591e-03,  8.14315006e-02, -5.26090898e-02,
                 4.30432940e-03, -1.62116084e-02, -1.06273042e-02, -1.22643486e-02,
                 4.80931215e-02, -2.30879411e-02, -1.24667808e-02, -7.11133704e-02,
                 9.66753662e-02,  4.45196740e-02,  6.21299148e-02,  5.14498651e-02,
                -5.90744428e-02, -6.24508969e-02,  3.96463200e-02,  5.72216101e-02,
                -8.12430158e-02, -5.77316098e-02, -1.04852393e-03,  5.44773452e-02,
```

```
In [46]: ko_model.wv.word_vec('운치있는').shape
```

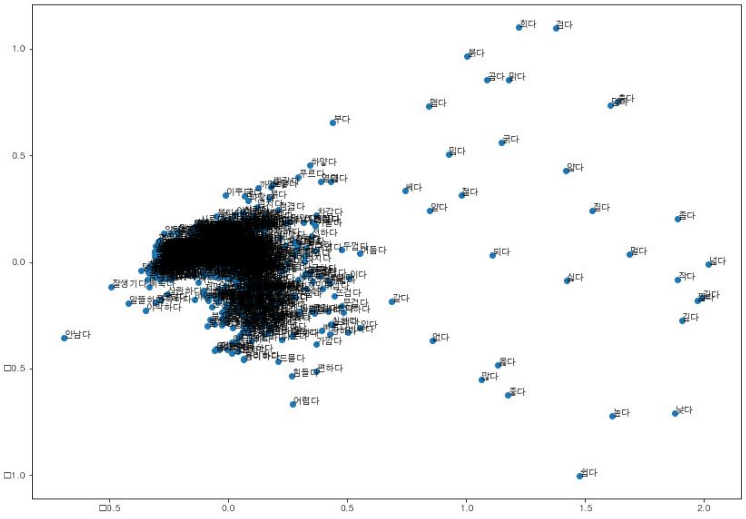
```
Out[46]: (300,)
```

```
tags_vec.head()
```

	tag	0	1	2	3	4	5	6	7	8	...	290	291	292	293	
1	건강하다	-0.033358	-0.107392	0.079212	-0.001186	-0.048449	-0.030069	-0.007499	0.001004	0.133595	...	0.013249	0.028847	0.045904	-0.028165	0.13
2	두드러지다	0.018185	0.010256	-0.047716	-0.012996	0.004557	-0.013141	-0.018327	-0.006541	0.017292	...	0.010328	-0.005689	0.014510	0.014435	-0.00
3	무관심하다	0.001222	0.031817	0.083566	0.017415	-0.005664	-0.016022	-0.024136	-0.002087	-0.029156	...	0.014161	0.028465	0.032042	-0.075265	-0.01
4	서툴다	-0.035465	0.106854	0.156820	-0.054557	0.041650	0.000100	-0.079635	-0.158097	-0.081634	...	0.072447	-0.029971	0.035325	-0.101966	-0.05
5	저조하다	-0.076719	0.009150	0.046816	-0.033846	0.011217	-0.013326	0.062515	-0.017747	-0.009819	...	0.055262	-0.029961	-0.042899	0.036459	0.00

5 rows × 301 columns

PCA로 2차원으로 줄인 후 시각화한 모습



어떤 모델을 어떻게 쓸지?

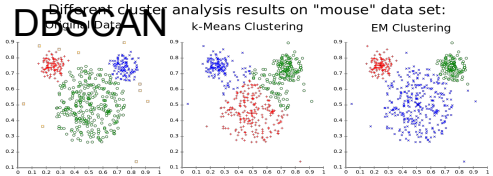
1) 한국어 리뷰 -> 영어로 바꿔서도 지 해

review	review_en
최고예요. 편안하게 있다 갔어요. 분위기 작살!!. 분위기조아요 연남동에서 유일하게...	It's the best. It's comfortable. He's gone. Th...
희수야효선왕돈까스 쌍문1줄번게	Heesuya Hyosun king pork cutlet Ssangmun No.1

2) 단어 벡터화 - FastText

```
Out[11]: array([ 3.23515316e-03, -1.44122355e-03, -8.17747638e-02,  4.61231880e-02,
                 6.49930760e-02, -1.86859816e-03,  3.08580901e-02,  3.63530256e-02,
                 4.65812832e-02,  3.09575372e-03, -4.47942875e-02, -1.39526548e-02,
                -9.68437642e-04,  5.35678901e-02,  1.20583080e-01,  5.72555773e-02,
                 3.45472582e-02, -2.87518445e-02,  1.74929295e-02, -2.02273875e-02,
```

3) 클러스터링 - Kmeans, GMM, DBSCAN



다음주 계획

1) 전처리 마무리 및 필드 통일 : 동혁

2) 클러스터링 결과 최적화 : 연주, 정수

3) 클러스터링 기반으로 매트릭스 생성

DataSet5 <형용사 태그 데이터셋>

ID	떠들썩한	조용한	운치있는	가족끼리	...
	1	0	0	1	
	0	1	1	0	
	0	1	1	1	