데이터분석캡스톤디자인

8주차 수행보고

Khupid 조

산업경영공학과 김동혁
관광학과 류연주
산업경영공학과 유정수

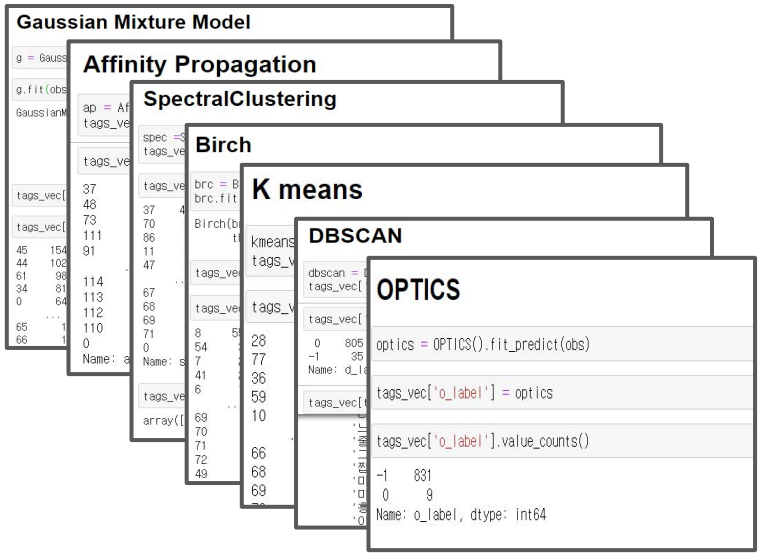
형용사 클러스터링을 위한 연구

지금까지 Fasttext를 기반으로 영어/한국어로 형용사를 추출 후
의미가 비슷한 형용사들을 clustering 기법을 통해 묶어주려했으나, 비슷한 의미가 뜻대로 묶이지 않았다.

Clustering Algorithm의 문제인가?
⇒ 좀 더 많은 Algorithm을 적용시켜본다.*(**FastText** 한국어 모델을 이용한 임베딩 + 클러스터링 기법 적용)*

	tag	0	1	2	3	4	5	6
1	기괴하다	-0.007321	0.012426	0.104259	0.017495	-0.008051	0.037065	0.003022
2	끔찍하다	0.008229	-0.009672	0.008787	0.003563	0.020002	0.011294	0.002887
3	갑작스럽다	-0.005319	-0.000359	-0.009199	0.032676	-0.019576	0.006620	0.006299
4	산만하다	-0.002303	-0.001701	0.056925	0.027435	0.015577	-0.021091	0.015858
5	놀지다	0.054632	0.005100	0.161192	-0.004379	-0.041849	0.054307	-0.028485

형용사 벡터화



GMM
Affinity Propagation
Spectral
Birch
K means
DBSCAN
OPTICS

여러가지 Clustering algorithm을 적용해보았으나 결과는 좋지않았음.

Model 재탐색

1. Word2Vec: 의미론적 word-embedding 기법, Korean/English 둘 다 적용함.

2. ELMO
3. BERT
4. GLoVe

} pre-trained Model이 없어 직접 학습을 시켜줘야함.
목적에 맞도록 labeling을 일일이 하기엔 무리이기때문에 model 후보에서 제외함.

Clustering Algorithm 선정

1. GMM
2. Affinity Propagation
3. Spectral
4. Birch
5. K means
6. DBSCAN
7. OPTICS

} 모두 적용해보고 가장 좋은 알고리즘으로 채택

word2vec 한글모델 기반 embedding
vector size, min_count, sg(skip-gram/cbow)등을 조절해가면서 여러 번 실험
여러 k에 대해 k-means를 사용하여 형용사 벡터들을 클러스터링

```
embedding_model = Word2Vec(token_list, size=300, window = 2, min_count=50, wo

word_vector = embedding_model.wv
vocabs = word_vector.vocab.keys()

word_vectors_list = [word_vector[v] for v in vocabs]

k=20

kmeans = KMeans(k)
idx = kmeans.fit_predict(word_vectors_list)

idx = list(idx)
names = embedding_model.wv.index2word
word_centroid_map = {names[i]: idx[i] for i in range(len(names))}

for c in range(k):
    print("\ncluster {}".format(c))

    words=[]
    cluster_values=list(word_centroid_map.values())
    for i in range(len(cluster_values)):
        if cluster_values[i]==c:
            words.append(list(word_centroid_map.keys())[i])
    print(words)
```

	거대하다	괜찮다	신선하다	이렇다	좋다	따뜻하다	이상하다	사랑스럽다
0	0.324746	0.242504	-0.432430	0.060653	-0.036340	0.124102	0.037602	-0.559983
1	0.320281	0.090737	0.180745	-0.047986	0.176596	-0.186441	-0.342656	-0.151901
2	-0.561919	-0.085515	0.011254	-0.156489	-0.221036	0.416091	0.074069	0.395090
3	-0.723728	0.162096	-0.070018	0.152233	0.146992	0.317003	0.311810	0.713209
4	-0.366435	-0.079968	-0.355087	0.022324	0.355560	0.290566	0.110819	0.775950
...	...	...	...	...	...	...	...	...
295	-0.251156	-0.291988	-0.222897	-0.064508	0.004471	-0.163904	-0.071834	0.172818
296	-0.306532	0.313867	0.085327	0.095472	0.238789	-0.061992	0.264054	-0.227528
297	-0.846977	-0.703526	0.004325	-0.001207	-0.256578	0.479444	0.113657	-0.493363
298	-0.043345	-0.451231	-0.003464	-0.069828	0.042802	0.740236	-0.254832	0.180462
299	0.114376	-0.395718	0.170821	-0.445048	-0.299400	0.155429	0.095275	-0.505252

```
다', '하얗다', '신하다', '싸잇하다', '우엇하다', '좋과하다', '좃좃하다', '필
요없다', '기쁘다', '넓다']

cluster 2
['탄탄하다']

cluster 3
['그렇다', '이렇다', '강하다', '엄청나다', '높다', '가깝다', '잘생기다', '고
맙다', '희다', '괴롭다', '자세하다', '붉다']

cluster 4
['진정하다', '원하다']

cluster 5
['가득하다', '느리다', '놀랍다', '질다']

cluster 6
['예쁘다', '적절하다', '잔잔하다', '씁쓸하다']

cluster 7
['기대하다', '흥미롭다', '분명하다', '지독하다', '찰지다']

cluster 8
['가볍다', '안타깝다', '가능하다', '무겁다', '끊임없다', '섹시하다', '영리하
다', '끔찍하다', '마르다', '심각하다', '담백하다', '실망하다', '멍청하다',
'낮다', '담담하다', '뭉클하다', '빈약하다']

cluster 9
['우아하다']

cluster 10
['재밌다', '슬프다', '행복하다', '부족하다', '넘치다', '따뜻하다', '재미있
다', '당하다', '어색하다', '이쁘다', '길다', '성공하다', '뜨겁다', '다양하
다', '그러하다', '딱하다', '편하다', '촌스럽다', '불쌍하다', '철저하다', '친
절하다', '쉽다', '화끈하다', '처절하다', '쓸쓸하다', '선하다', '잔혹하다',
'달콤하다']
```

skip-gram 기반의 fast text보다 훨씬 나아진 모습
형용사 셋 처리 및 튜닝을 좀 더 해주면 괜찮아 질 것 같다.

영어 리뷰: Word2Vec + K-Means 결과

clustering 개수를 20, 30개로 조정해가며 결과를 보고
의미가 없거나 명사인데 형용사로 구분되어 나타난 클러스터링에 속한 단어들을 제거해가며 약 10번 이상의 클러스터링 작업 진행

가격	부정적인 감정	물리적인 장소	분위기	색상	자연/환경	나라별/국제
group_0	**group_1**	**group_8**	**group_9**	**group_16**	**group_19**	**group_21**
expensive	different	places	chic	black	environmental	nan
convenient	burdensome	outdoor	trendy	silver	visual	chinese
cheap	ambiguous	nearby	luxurious	blue	acid	indian
affordable	easy	indoor	oriental	red	scenic	b
0	unfriendly	interior	tropical	yellow	facial	shit
0	unlikely	dusty	nightlife	fluorescent	mechanical	pong
0	strange	spacious	sensuous	green	human	korean
0	improbable	upstairs	pickled		nature	etc
0	uncomfortable	abandoned	fermented		plastic	
0	dangerous	octagonal	romantic		organic	
0	unpleasant	observatory	exotic		pure	

다음주 할 것

1. 갈/볼/먹 부분에서 각각 형용사 Matrix 생성
2. 협업 필터링 모델 탐색
3. 수집한 Matrix 및 User 정보를 협업 필터링 모델에 적용

	cluster1	cluster2	cluster3	cluster4	cluster5
장소1	0.6	0.1	0	0	0.3
장소2	0.4	0.2	0.2	0.2	0
장소3	0.5	0.3	0.2	0	0

	cluster1	cluster2	cluster3	cluster4	cluster5
사람1	0	0	0.8	0.1	0.1
사람2	0.5	0.2	0.3	0	0
사람3	0.3	0.2	0.1	0.2	0.2