2020 데이터 크리에이터 캠프

데이터사나이
이재상
김윤환, 박준민, 박대한, 정규형

NIA 한국정보화진흥원

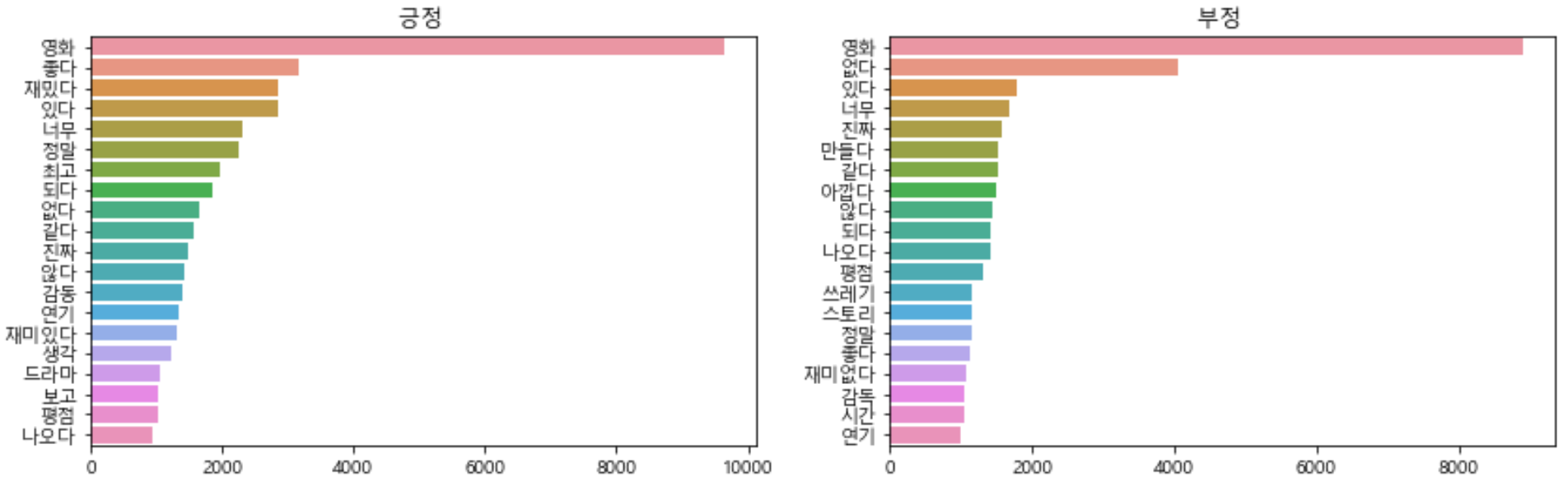
Ⅰ. 소개

1. 문제 분석
2. 모델 소개

Ⅰ. 소개

1 문제 분석

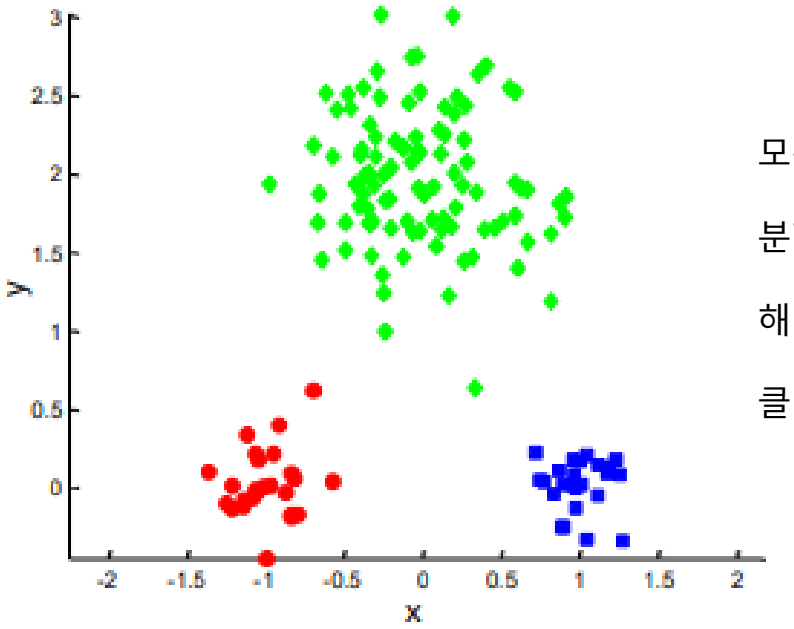
긍정적 / 부정적 영화 리뷰들을 군집화하고,
각 군집 별 특징을 해석하여 고객층을 세분화하여
마케팅 전략팀에게 전달할 수 있도록 분석



기존 코드를 활용하여 긍정/부정 리뷰 형태소 빈도수

2 모델 소개

① Kmeans-clustering



모든 객체들을 주어진 k개의 그룹으로 분할

분할한 클러스터 내의 객체들로부터 새로운 seed 객체를 탐색

해당 클러스터의 평균값을 새로운 중심으로 할당

클러스터가 변화하지 않을 때까지 반복

Ⅱ. 데이터 전처리 과정

1. 전처리 방법
2. 데이터 전처리

Ⅱ. 데이터 전처리 과정

1 전처리 방법

형태소 분석기 선택

빠른 속도와 보통의 정확도를 원한다면 "Komoran“

속도는 느리더라도 정확하고 상세한 품사 정보를 원한다면 "Kkma"

어느 정도의 띄어쓰기 되어 있는 "인터넷" 영화평/상품명을 처리할 땐 "Okt"

⇒ 시간적 효율성과 도메인에 대한 기존 분석들을 참고하여,
“Okt” 형태소 분리기를 이용하기로 결정

<참고>
https://m.blog.naver.com/PostView.nhn?blogId=wideeyed&logNo=221337575742&proxyReferer=https:%2F%2Fwww.google.com%2F

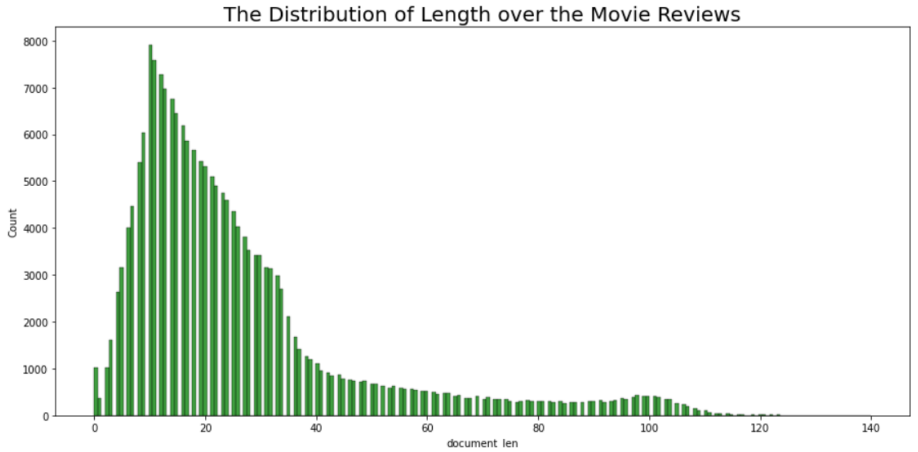
Ⅱ. 데이터 전처리 과정

2 데이터 전처리

① 사전 처리

형태소 분리기 처리 전 데이터 전처리

– 띄어쓰기 제거 한 데이터를 바탕으로 글자 수 추출 후, 범위 설정(10–17) : 53490개의 리뷰를 분석




```
subset_df = unique_df1[unique_df1["document_len"] >= 10] #
subset_df = subset_df[subset_df["document_len"] <= 17] # 'd
subset_df.shape # 'document_len' 값이 1분위수 (12) 이상, 15
```

```
(53490, 5)
```

Ⅱ. 데이터 전처리 과정

2 데이터 전처리

② Okt 형태소 분리기

- 정규 표현식을 이용, 띄어쓰기를 포함 시킨 데이터를 이용하여 형태소 분리
- 분리 결과가 명사, 동사, 형용사, 부사만 포함하도록 설정

	id	document	label	preprocessed_document	document_len	v_n_ad
0	8112052	어릴때보고 지금다시봐도 재밌어요ㅋㅋ	1	어릴때보고 지금다시봐도 재밌어요ㅋㅋ	17.0	[어리다, 보고, 지금, 다시, 재밌다, ㅋㅋ]
1	9279041	완전 감동입니다 다시봐도 감동	1	완전 감동입니다 다시봐도 감동	13.0	[완전, 감동, 다시, 감동]
2	9250537	바보가 아니라 병 쉰 인듯	1	바보가 아니라 병 쉰 인듯	10.0	[바보, 병, 쉰]
3	9537008	고질라니무 귀엽다능ㅋㅋ	1	고질라니무 귀엽다능ㅋㅋ	11.0	[고질, 귀엽다, ㅋㅋ]
4	8703997	가면 갈수록 더욱 빠져드네요 밀회 화이팅!!	1	가면 갈수록 더욱 빠져드네요 밀회 화이팅	17.0	[가면, 갈수록, 더욱, 빠지다, 밀회, 화이팅]

- Stop words list를 간이로 만들어서 결과에 반영

```
# stopwords 설정
stopword = ['중세시대', '다년', '롤라', '거더', '델과킴', '일리', '와대', 'MBC', '더빨', '크슈', '홖', '도난', '리좀', '말타', '카스
            '야할량', '랑', '마테', '가희', '넝', '삼겹살', '게왜점', '조도', '로프', '스리', '넥션', '와쿠', '김영', '갠', '이아쉬
            '내릴', '스내치', '년됫', '년됐', '헤드', '게토', '개토', '때', '이다', '보다', '아니다', '능', '무', '드', '알다', '하다
```

Ⅲ. 분석결과

1. 긍정 그룹 분석
2. 부정 그룹 분석

Ⅲ. 분석결과

1 긍정 그룹 분석

> 공통된 특성으로는 재밌다, 영화, 최고 등의 단어가 많이 노출되었다

Cluster 0 : 너무, 드라마, 감동 ,연기 등의 키워드가 많이 노출
-> 영화의 감성적 특징(배경, 미장셴, 배우들의 연기톤, 감동적 포인트)

Cluster 1 : 영화, 재밌다, 만들다, 생각, 감독
-> 영화를 만드는 과정, 제작, 감독의 역량등이 위주

Cluster 2 : 좋다, 나오다, 마지막, 배우, 스토리, 평점, 작품
-> 결말의 완결성, 스토리의 완전성 등이 중요한 파트,

Ⅲ. 분석결과

2 부정 그룹 분석

> 영화, 쓰레기가 공통된 특성

Cluster 0 : 진짜 나오다 없다 재미없다 영화 같다 있다 아깝다 쓰레기 최악
-> 영화의 감성적 특징(배경, 미장셴, 배우들의 연기톤, 감동적 포인트)

Cluster 1 : 영화 쓰레기 만들다 없다 되다 사람 감독 너무
--> 영화를 만드는 과정, 제작, 감독의 역량등이 위주

Cluster 2 : 없다 같다 않다 되다 좋다 스토리
-> 결말의 완결성, 스토리의 완전성 등이 중요한 파트,

Ⅳ. 향후 발전 방향

Ⅳ. 향후 발전 방향

- dbscan, hdbscan,kmedoids 등 다른 클러스터링 방법들을 시간 제약, 컴퓨터 성능 등의 제약으로 시도해보지 못한 점.

- 중복 의미되는 단어 처리

Q & A