7 문서 군집화 소개와 실습

문서군집화 개념

문서 군집화: 비슷한 텍스트 구성의 문서를 군집화
동일한 군집에 속하는 문서를 같은 카테고리 소속으로 분류할 수 있으므로 텍스트 분류 기반의 문서 분류와 유사
비지도 학습 기반으로 동작

opinion review 데이터 세트를 이용한 문서 군집화 수행하기

개별 문서 텍스트에 대해 TF-IDF 변환된 피처 벡터화된 행렬 구함
군집화 기법은 k-means 이용

군집별 핵심 단어 추출하기

+ 코드 + 텍스트

각 군집에 속한 문서는 핵심 단어를 주축으로 군집화
kmeans 객체는 각 군집을 구성하는 단어 피처가 군집의 중심을 기준으로 얼마나 가깝게 위치해 있는지 cluster_centers_라는 속성으로 제공
0에서 1까지의 값을 가질 수 있으며 1에 가까울 수록 중심과 가까운 값을 의미
cluster_centers_ 속성값을 이용해 각 군집별 핵심 단어 추출
ndarray의 argsort()[:,::-1]를 이용하면 cluster_centers 배열 내 값이 큰 순으로 정렬된 위치 인덱스 값 반환

8 문서 유사도

문서 유사도 측정 방법- 코사인 유사도

문서와 문서 간의 유사도 비교는 일반적으로 코사인 유사도를 사용
벡터와 벡터 간의 유사도를 비교할 때 벡터의 크기보다는 벡터의 상호 방향성이 얼마나 유사한지에 기반
두 벡터 사이의 사잇각을 구해서 얼마나 유사한지 수치로 적용한 것

두 벡터 사잇각

두 벡터 사잇각에 따라 상호 관계는 유사하거나 관련이 없거나 아예 반대 관계가 될 수 있음
유사도는 두 벡터의 내적을 총 벡터 크기의 합으로 나눈 것

코사인 유사도가 문서의 유사도 비교에 가장 많이 사용되는 이유: 문서를 피처 벡터화 변환하면 차원이 매우 많은 희소 행렬이 되기 쉬움, 희소 행렬 기반에서 문서와 문서 벡터간의 크기에 기반한 유사도 지표는 정확도가 떨어지기 쉬움
문서가 매우 긴 경우 단어의 빈도수도 더 많을 것이기 때문에 이러한 빈도수에만 기반해서는 공정한 비교를 할 수 없음

opinion review 데이터 세트를 이용한 문서 유사도 측정

9 한글 텍스트 처리 - 네이버 영화 평점 감성 분석

한글 NLP 처리의 어려움

띄어쓰기와 다양한 조사 때문

KoNLPy 소개

파이썬의 대표적인 한글 형태소 패키지
형태소 분석이란 말뭉치를 이러한 형태소 어근 단위로 쪼개고 각 형태소에 품사 태깅을 부착하는 작업을 지칭

데이터 로딩

각 문장을 한글 형태소 분석을 통해 형태소 단어로 토큰화
TF-IDF 피처 모델을 생성
로지스틱 회귀를 이용해 분류 기반의 감성 분석 수행
최종 감성 분석 예측 수행

✓ 0초 오후 4:57에 완료됨