WEEK13_예습과제_정윤지

토픽 모델링 – 20 뉴스그룹

토픽 모델링

- 문서 집합에 숨어 있는 주제를 찾아내는 것
- 숨겨진 주제를 효과적으로 표현할 수 있는 중심 단어를 함축적으로 추출
- LSA, LDA
- 실습

문서 군집화 소개와 실습

문서 군집화 개념

- 비슷한 텍스트 구성의 문서를 군집화하는 것
- 동일한 군집에 속하는 문서를 같은 카테고리 소속으로 분류할 수 있음
- 실습

군집별 핵심 단어 추출하기

- 군집에 속한 문서는 핵심 단어를 주축으로 군집화
- KMeans: 각 군집을 구성하는 단어 피처가 군집의 중심을 기준으로 얼마나 가깝게 위치해 있는지 clusters_centers_ 속성으로 제공
- clusters_centers_는 배열 값으로 제공, 행은 개별 군집, 열은 개별 피처를 의미함
- 각 배열 내의 값은 개별 군집 내의 상대 위치를 숫자 값으로 표현한 일종의 좌표 값
- 실습

문서 유사도

문서 유사도 측정 방법 – 코사인 유사도

- 문서와 문서 간의 유사도 비교는 일반적으로 코사인 유사도를 사용
- 벡터와 벡터 간의 유사도를 비교할 때 벡터의 크기보다 벡터의 상호 방향성이 얼마나 유산한지에 기반

- 코사인 유사도는 두 벡터 사이의 사잇각을 구해, 얼마나 유사한지 수치로 적용

두 벡터 사잇각

- 두 벡터 사잇각에 따라 상호 관계느 유사하거나, 관련이 없거나, 아예 반대 관계가 될 수 있음

$$A * B = \| A \| \| B \| cos\theta$$

-

$$\text{similarity} = cos\theta = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

-
- 문서를 피처 벡터화 변환을 하면, 차원이 매우 많은 희소 행렬이 되기 쉬움
- 희소 행렬 기반에서 문서와 문서 벡터 간의 크기에 기반한 유사도 지표는 정확도가 떨어지기 쉬움
- 실습

Opinion Review 데이터 세트를 이용한 문서 유사도 측정

- 실습

한글 텍스트 처리 – 네이버 영화 평점 감성 분석

한글 NLP 처리의 어려움

- 한글 언어 처리는 띄어쓰기와 다양한 조사 때문에 처리가 어려움
- 띄어쓰기에 따라 의미가 달라질 수 있음
- 조사는 어근 추출 등의 전처리 시 제거하기 까다로움

KoNLPy

- 파이썬의 대표적인 한글 형태소 패키지
- 설치 방법
- 실습