15주차 예습

날짜	@2024년 6월 18일
태그	예습

6. 토픽 모델링

- 토픽 모델링: 문서 집합에 숨어 있는 주제를 찾아내는 것
  - 많은 양의 문서가 있을 때 사람이 문서를 읽고 주제를 찾는 데에는 많은 시간이 소요됨
  - 따라서 머신러닝 기반의 토픽 모델링을 적용해 숨어 있는 중요 주제를 효과적으로 찾아낼 수 있음
    : 숨겨진 주제를 효과적으로 표현할 수 있는 중심 단어를 함축적으로 추출함

- 토픽 모델링의 종류
  - LSA: 머신러닝 기반의 토픽 모델링에서 자주 사용
  - LDA: 문서 집합으로부터 어떤 토픽이 존재하는지를 알아내기 위한 알고리즘
    - 수행 과정
      1. 사용자가 알고리즘에게 토픽의 개수(k)를 알려준다
      2. 모든 단어를 k개 중 하나의 토픽에 할당한다
      3. 모든 문서의 모든 단어에 대해 아래의 사항을 반복한다

         어떤 문서의 각 단어 w는 자신은 잘못된 토픽에 할당되어져있지만, 다른 단어들은 모두 올바른 토픽에 할당되어져 있는 상태라고 가정하였을 때 단어 w는 두 가지 기준에 따라 토픽이 재할당된다

         a. 기준1
            p: 문서 d의 단어들 중 토픽 t에 해당하는 단어들의 비율
         b. 기준2
            각 토픽들 t에서 해당 단어 w의 분포

7. 문서 군집화

문서 군집화 개념

- 비슷한 텍스트 구성의 문서를 군집화 하는 것
- 동일한 군집에 속하는 문서를 같은 카테고리 소속으로 분류할 수 있으므로 앞의 텍스트 분류 기반의 문서 분류와 유사

군집별 핵심 단어 추출하기

- 군집화를 진행하면, 각 군집에 속한 문서는 핵심 단어를 주축으로 군집화되어 있음
- KMeans 객체는 각 군집을 구성하는 단어 피처가 군집의 중심을 기준으로 얼마나 가깝게 위치해있는지 clusterscenters라는 속성으로 제공

8. 문서 유사도

코사인 유사도를 이용한 문서 유사도 측정

코사인 유사도: 문서 간 유사도 측정 시 일반적으로 사용

- 벡터와 벡터 간의 유사도를 비교할 때 벡터의 크기보다 벡터의 상호 방향성이 얼마나 유사한지에 기반한다
- 두 벡터 사이의 사잇각을 구해 얼마나 유사한지 수치로 적용한다
    - 사잇각에 따라 상호관계는 유사하거나, 관련이 없거나, 아예 반대가 될 수 있다
    - 두 벡터의 내적을 총 벡터 크기의 합으로 나눠 유사도의 코사인 값을 구한다

9. 한글 텍스트 처리

한글 NLP 처리의 어려움

- 띄어쓰기, 다양한 조사로 인해 한글 언어 처리가 어렵다

- 띄어쓰기를 잘못하면 의미가 왜곡되어 전달될 수 있음
- 조사는 주어나 목적어를 위해 추가되는데 경우의 수가 많아서 어근 추출 등의 전처리 시 제거하기가 까다로움