

13. 자연어 처리와 문자 추출

3강. 토픽 모델링과 문서 군집화

학습내용

- 토픽 모델링
- LDA

학습목표

- 모델링의 대표적인 알고리즘인 LDA에 대해 학습하고 머신러닝에 적용할 수 있다.

1. 토픽 모델링과 문서 군집화

1) 기본 개념

가. 토픽 모델링

- 텍스트 데이터에 자주 적용하는 특별한 기법
- 비지도 학습으로 문서를 하나 또는 그 이상의 토픽으로 할당하는 작업을 통칭
- 토픽 모델링의 예) 정치, 스포츠, 금융 등의 토픽으로 묶을 수 있는 뉴스 데이터
- 학습된 각 성분은 하나의 토픽에 해당하며 문서를 표현한 성분의 계수는 문서가 어떤 토픽에 얼마만큼 연관되어 있는지를 말해줌
- 토픽은 일상에서의 "f주제"를 의미하기 보다는 PCA나 NMP로 추출한 어떤 성분에 가까움
- 토픽 모델링의 대표적 알고리즘 : LDA (잠재 디리클레 할당, Latent Dirichlet Allocation)

- 문서 하나에 여러가지 토픽이 혼합 가능
- 토픽에는 여러 개의 단어가 포함 가능
- 문서에 사용한 각각의 단어는 각각의 토픽에 포함
- 문서 안에 있는 단어로 잠재 주제를 만들어 문서 각각이 어떤 내용을 구성되었는지 표현

2) LDA

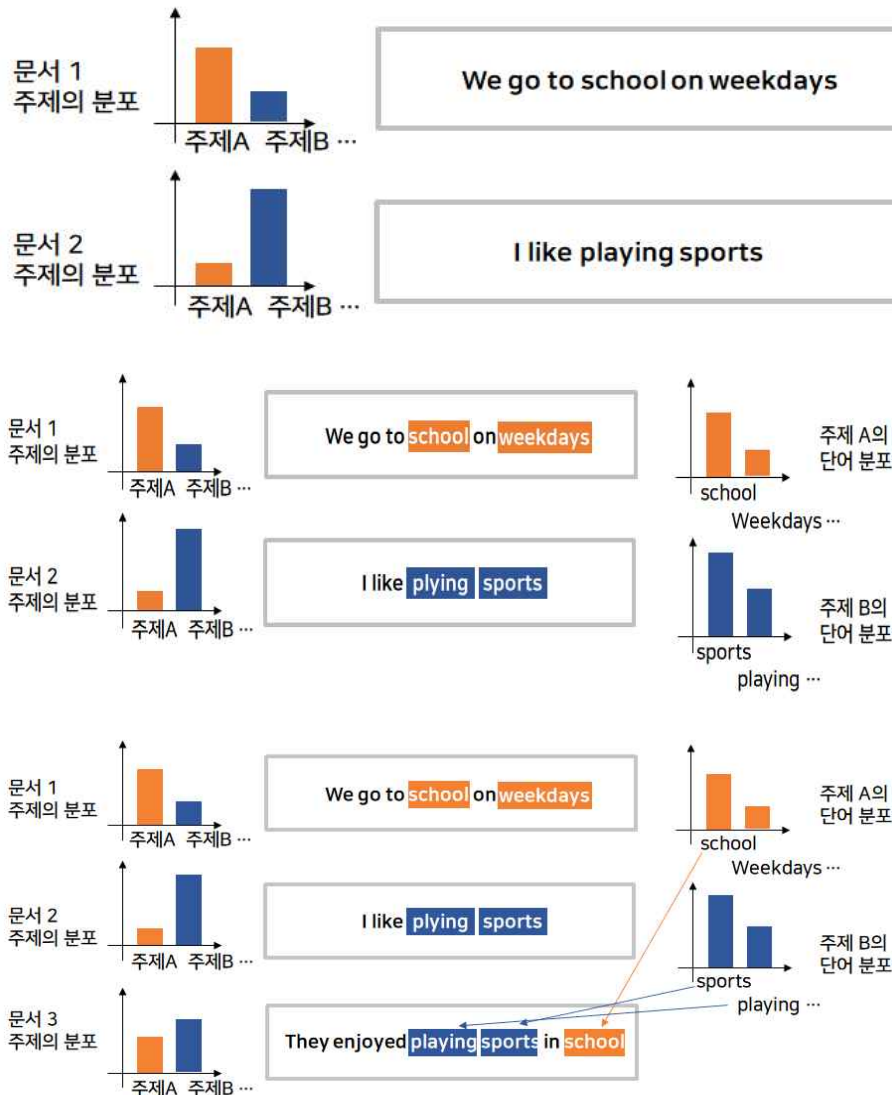
가. 알고리즘 과정

- ① 문서 각각의 단어에 무작위로 주제를 할당
 - ② 단어에 할당한 주제에서 문서 각각의 주제별 확률을 계산
 - ③ 단어에 할당한 주제에서 주제별 단어 확률을 계산
 - ④ ②와③을 곱해서 계산하는 확률로 문서 각각의 단어에 주제를 다시 할당
 - ⑤ 과정 ② ~ ④를 수렴 조건까지 반복
- 과정 ④에서 계산한 확률에 따라 문서 각각의 단어에 주제를 할당하며 문서 각각의 주제 확률은 과정 ②로 정하므로, 같은 문서 안에서 특정 주제를 선택하기 쉬움. 즉, 같은 문서의 단어는 같은 주제를 할당하는 경향이 있음
 - 계산을 반복하면 문서에 특정 주제를 할당할 확률이 커짐 또한, 특정 주제와 관련 있는 단어를 선택하기 쉬우므로 해당 주제를 대표하는 단어를 선택할 확률이 커짐

나. 예제 학습

- 5개의 영어문장에 잠재 다리클레 할당 적용
- 주제의 개수 : 2개

1. We go to school on weekdays.
2. I like playing sports.
3. They enjoyed playing sports in school.
4. Did she go there after school?
5. He read the sports columns yesterday.



평가하기

1. 토픽 모델링은 지도 학습으로 문서를 하나 또는 그 이상의 토픽으로 할당하는 작업을 통칭한다.

O X

- 정답 : X

해설 : 토픽 모델링은 비지도 학습에 속합니다.

학습정리

1. 토픽 모델링

- 비지도 학습으로 문서를 하나 또는 그 이상의 토픽으로 할당하는 작업을 통칭
- 토픽은 일상에서의 '주제'를 의미하기 보다는 PCA나 NMP로 추출한 어떤 성분
에 가까움

2. LDA

문서 안에 있는 단어로 잠재 주제를 만들어 문서 각각이 어떤 내용을 구성되었는지 표현

다음 주 예고

“14. 딥러닝과 심층 신경망”에 대해 학습하겠습니다.