Week 12_예습과제_김정은

텍스트 분석

- 비정형 데이터인 텍스트를 분석하는 것
- 비정형인 텍스트 데이터를 어떻게 피처 형태로 추출하고 추출된 피처에 의미 있는 값을 부여하는가 하는 것이 매우 중요한 요소
- 텍스트를 word 기반의 다수의 피처로 추출하고 이 피처에 단어 빈도수와 같은 숫자 값을 부여하면 텍스트는 단어의 조합인 벡터값으로 표현 가능 → 피처 벡터화, 피처 추출
- 텍스트를 피처 벡터화해서 변환하는 방법 → BOW(Bag of Words)

텍스트 분석 수행 프로세스

1. 텍스트 사전 준비작업(텍스트 전처리)
2. 피처 벡터화/추출
3. ML 모델 수립 및 학습/예측/평가

텍스트 사전 준비 작업(텍스트 전처리) - 텍스트 정규화

텍스트 정규화 작업

- 클렌징
- 토큰화
- 필터링/스톱워드 제거/철자 수정
- Stemming
- Lemmatization

클렌징

- 텍스트에서 분석에 방해가 되는 불필요한 문자, 기호 등을 사전에 제거

텍스트 토큰화

- 문서에서 문장을 분리하는 문장 토큰화 / 문장에서 단어를 토큰으로 분리하는 단어 토큰화

문장 토큰화

- 문장의 마침표, 개행문자 등 문장의 마지막을 뜻하는 기호에 따라 분리하는 것이 일반적.
- 정규 표현식에 따른 문장 토큰화도 가능.

단어 토큰화

- 문장을 단어로 토큰화
- 기본적으로 공백, 콤마, 마침표, 개행문자 등으로 단어를 분리하지만 정규 표현식을 이용해 다양한 유형으로 토큰화를 수행

스톱 워드 제거

- 스톱 워드는 분석에 큰 의미가 없는 단어를 지칭
- 문맥적으로 큰 의미가 없는 단어가 해당, 사전에 제거 → 전처리 작업
- 언어별로 이런 스톱 워드가 목록화되어 있음
- NLTK의 경우 가장 다양한 언어의 스톱워드를 제공

Stemming과 Lemmatization

- Lemmatization이 Stemming보다 정교하며 의미론적인 기반에서 단어의 원형을 찾음
- Stemming은 원형 다어로 변환 시 일반적인 방법을 적용하거나 더 단순화된 방법을 적용해 원래 단어에서 일부 철자가 훼손된 어근 단어를 추출하는 경향

- Lemmatization은 품사와 같은 문법적인 요소와 더 의미적인 부분을 감안해 정확한 철자로 된 어근 단어를 찾음, 더 오랜 시간 필요

Back of Words - BOW

- 문서가 가지는 모든 단어를 문맥이나 순서를 무시하고 일괄적으로 단어에 대해 빈도값을 부여해 피처 값을 추출하는 모델.
- 문서 내 모든 단어를 한꺼번에 봉투에 넣은 뒤 흔들어서 섞는다는 의미로 BOW 모델이라고 함
- 장점 : 쉽고 빠른 구축
- 단어의 발생 횟수에 기반하지만 예상보다 문서의 특징을 잘 나타내는 모델이므로 전통적으로 여러 분야에 활용도가 높음
- 단점 : 문맥 의미 반영 부족, 희소 행렬 문제

BOW의 피처 벡터화 방식

- 카운트 기반의 벡터화
- TF-IDF 기반의 벡터화

BOW 벡터화를 위한 희소 행렬

- 희소 행렬 : 대규모 행렬의 대부분의 값을 0이 차지하는 행렬을 가리켜 희소 행렬이라고 한다.
- BOW 형태를 가진 언어 모델의 피처 벡터화는 대부분 희소 행렬
- 너무 많은 불필요한 0값이 메모리 공간에 할당되어 메모리 공간이 많이 필요하며, 행렬의 크기가 커서 연산 시에도 데이터 액세스를 위한 시간이 많이 소모됨. → 적은 메모리 공간을 차지할 수 있도록 변환해야 함
    - → 대표적 방법 : COO 형식 / CSR 형식
    - 일반적으로 큰 희소 행렬을 저장하고 계산을 수행하는 능력이 CSR 형식이 더 뛰어남

감성 분석

- 문서의 주관적인 감성/의견/감정/기분 등을 파악하기 위한 방법으로 소셜 미디어, 여론 조사, 온라인 리뷰, 피드백 등 다양한 분야에서 활용됨
- 문서 내 텍스트가 나타내는 여러 가지 주관적인 단어와 문맥을 기반으로 감성 수치를 계산하는 방법을 이용
- 긍정 감성 지수와 부정 감성 지수로 구성되며 이들 지수를 합산해 긍정 감성 또는 부정 감성을 결정
- 머신러닝 관점에서 지도학습 / 비지도학습 방식으로 나눌 수 있음