8장. 텍스트 분석

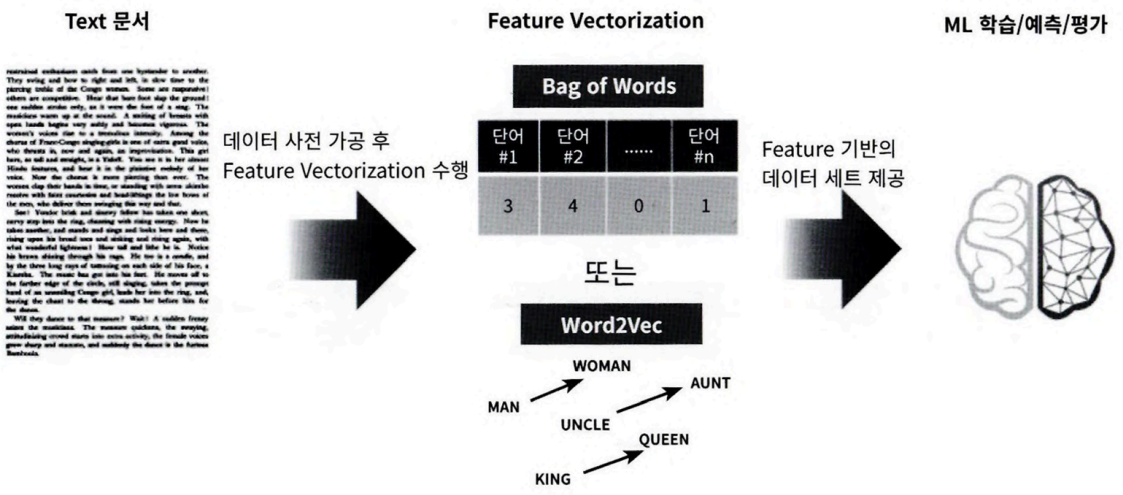
01 텍스트 분석 이해

텍스트 분석 개요

- 비정형 데이터인 텍스트를 분석하는 것
- 피처 백터화(피터 추출): 텍스트를 word 기반의 다수의 피처로 추출하고 단어 빈도수와 같은 값을 부여해 텍스트를 단어의 조합인 벡터값으로 표현하는 것

텍스트 분석 수행 프로세스

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



〈 Text Analytics 수행 프로세스 〉

파이썬 기반의 NLP, 텍스트 분석 패키지

- NLTK: 파이썬의 가장 대표적인 NLP 패키지, NLP의 거의 모든 영역을 커버, 수행 속도 측면의 문제로 대량의 데이터 기반에서는 제대로 활용되지 못함
- Gensim: 토픽 모델링 분야에서 가장 두각을 드러내는 패키지
- SpaCY: 가장 뛰어난 수행 성능을 가진 패키지

02 텍스트 사전 준비 작업(텍스트 전처리)-텍스트 정규화

개요

- 텍스트를 가공하는 준비 작업
- 크게 클렌징(Cleansing)과 토큰화Tokenization)로 분류됨

클렌징

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

텍스트 토큰화

- 문장 토큰화: 문서에서 문장을 분리

  문장의 마침표(.)나 기행문자(\n)로 분리하거나 정규 표현식을 이용

- 단어 토큰화: 문장에서 단어를 분리

  공백, 콤마(,), 마침표(.) 기행문자로 분리하거나 정규 표현식 이용

스톱 워드 제거

- 스톱 워드: 큰 의미가 없는 단어

Stemming과 Lemmatization

- 문법적 또는 의미적으로 변화하는 단어의 원형을 찾는 것

03 Bag of Words - BOW

개요

- 문서가 가지는 모든 단어를 문맥이나 순서를 무시하고 단어의 빈도 값을 부여해 피처 값을 추출하는 모델
- 장점: 쉽고 빠른 구축

단점: 문맥 의미 반영 부족, 희소 행렬 문제

BOW 피처 벡터화

- 텍스트를 특정 의미를 가지는 숫자형 값인 벡터 값으로 변환하는 것
- 텍스트 데이터를 또 다른 형태의 피처의 조합으로 변경한다는 점에서 넓은 범위의 피처 추출에 포함됨

BOW의 피처 벡터화	설명
카운트 기반	문서에서 해당 단어가 나타나는 횟수(Count)를 부여함 카운트 값이 높을수록 중요한 단어
TF-IDF 기반	개별 문서에서 자주 나타나는 단어에 높은 가중치를 주되, 모든 문서에서 전반적으로 자주 나타나는 단어에 대해 패널티를 줌

⇒ 텍스트가 길고, 문서의 개수가 많은 경우 TF-IDF 방식을 사용하는 것이 더 좋은 예측 성능을 보장

BOW 벡터화를 위한 희소행렬

- 희소 행렬: 대규모 행렬의 대부분의 값을 0이 차지하는 행렬

  BOW 형태를 가진 언어 모델의 피처 벡터화는 대부분 희소 행렬
- 희소 행렬이 물리적으로 적은 메모리 공간을 차지하도록 변환하는 방법에는 COO와 CSR 형식이 있음

희소 행렬 - COO 형식

- 0이 아닌 데이터만 별도의 데이터 배열에 저장하고, 그 데이터가 가르키는 행과 열의 위치를 별도의 배열로 저장
- 파이썬에서 Scipy(사이파이)의 sparse 패키지를 이용해 수행 가능

희소 행렬 - CSR 형식

- 행 위치 배열 고유한 값의 시작 위치만 표기하는 방식
- COO 형식의 행과 열의 위치를 나타내기 위한 반복적인 위치 데이터 사용의 문제점을 해결
- 사이파이의 csr_matrix 클래스를 이용해 수행 가능

04 텍스트 분류 실습 - 20 뉴스그룹 분류

05 감성 분석

감성 분석 소개

- 문서의 주관적인 감성/의견/감정/기분 등을 파악하기 위한 방법
- 문서 내 텍스트가 나타내는 여러 주관적인 단어와 문맥을 기반으로 감성 수치를 계산
- 감성 지수(긍정/부정 감성 지수)를 합산해 긍정 감성 또는 부정 감성 결정
- 머신러닝 관점에서 지도학습과 비지도학습 방식으로 나눌 수 있음

지도학습	학습 데이터와 타깃 레이블 값을 기반으로 감성 분석 학습을 수행한 뒤, 이를 기반으로 다른 데이터의 감성 분석 예측
비지도학습	감성 어휘 사전 'Lexicon'을 이용해 문서의 긍정적, 부정적 감성 여부 판단

지도학습 기반 감성 분석 실습 - IMDB 영화평

비지도학습 기반 감성 분석 소개

- 감성 분석용 데이터가 결정된 레이블 값을 가지고 있지 않을 때 Lexcion을 이용해 감성 분석 수행
- Lexcion(감성 사전): 긍정 감성과 부정 감성의 정도를 의미하는 감성 지수를 가지고 있음
- NLTK 패키지를 통해 구현함
- NLP 패키지는 시맨틱(문맥상 의미)을 프로그램적으로 인터페이스할 수 있는 다양한 방법 제공
- 대표적인 감성 사전

SentiWordNet을 이용한 감성 분석

1. 문서를 문장 단위로 분해
2. 문장을 단어 단위로 토큰화하고 품사 태깅
3. 품사 태깅된 단어 기반으로 객체 생성

4. 객체에서 긍정/부정 감성 지수 구하고, 이를 합산해 긍정/부정 감성 결정

VADER를 이용한 감성 분석

- SentimentIntensityAnalyer 클래스를 이용해 감성 분석 제공