WEEK12_예습과제_정윤지

08 텍스트 분석

텍스트 분석

- 머신러닝, 언어 이해, 통계 등을 활용해 모델을 수립하고 정보를 추출해 비즈니스 인텔리전스나 예측 분석 등의 분석 작업을 주로 수행
- 텍스트 분류, 감성 분석, 텍스트 요약, 텍스트 군집화 등의 기술 영역에 집중

텍스트 분석 이해

- 비정형 데이터인 텍스트를 분석
- 텍스트를 word 기반의 다수의 피처로 추추하고 이 피처에 단어 빈도수와 같은 숫자 값을 부여하면 텍스트는 단어의 조합인 벡터 값으로 표현될 수 있는데, 이렇게 텍스트를 변환하는 것을 피처 백터화 또는 피처 추출이라고 함
- 텍스트를 피처 백터화해서 변환하는 방법에는 BOW와 Word2Vec 방법이 있음

텍스트 분석 수행 프로세스

- 텍스트 사전 준비 작업
- 피처 백터화/추출
- ML 모델 수립 및 학습/예측/평가

파이썬 기반의 NLP, 텍스트 분석 패키지

- NLTK는 방대한 데이터 세트와 서브 모듈, 다양한 데이터 세트를 지원해 오래전부터 대표적인 파이썬 NLP 패키지였지만 수행 성능과 정확도, 신기술, 엔터프라이즈한 기능 지원 등의 측면에서 부족한 부분이 있음
- Genism, SpaCy는 이런 부분을 보완하면서 실제 업무에서 자주 활용되는 패키지
- NLTK: 파이썬의 가장 대표적인 NLP 패키지, 방대한 데이터 세트와 서브 모듈을 보유, NLP의 거의 모든 영역을 커버
- Gensim: 토픽 모델링 분야에서 가장 두각을 나타내는 패키지
- SpaCy: 뛰어난 수행 성능으로 최근 가장 주목을 받는 NLP 패키지

텍스트 사전 준비 작업(텍스트 전처리) – 텍스트 정규화

- 텍스트 정규화는 텍스트를 머신러닝 알고리즘이나 NLP 애플리케이션에 입력 데이터로 사용하기 위해 클렌징, 정제, 토큰화, 어근화 등의 다양한 텍스트 데이터의 사전 작업을 수행하는 것을 의미
- 클렌징, 토큰화, 필터링/스톱 워드 제거/철자 수정, stemming, lemmatization

클렌징

- 텍스트에서 분석에 방해되는 불필요 문자, 기호 등을 사전에 제거하는 작업

텍스트 토큰화

- 토큰화의 유형은 문서에서 문장을 분리하는 문장 토큰화와 문장에서 단어를 토큰으로 분리하는 단어 토큰화로 나눌 수 있음

문장 토큰화

- 문장의 마침표, 개행문자 등 문자으이 마지막을 뜻하는 기호에 따라 분리하는 것이 일반적

단어 토큰화

- 문장을 단어로 토큰화하는 것
- 정규 표현식을 이용해 다양한 유형으로 토큰화를 수행할 수 있음

스톱 워드 제거

- 분석에 큰 의미가 없는 단어를 지칭

Stemming 과 lemmatization

- 문법적 또는 의미적으로 변화하는 단어의 원형을 찾는 것
- Lemmatization이 Stemming 보다 정교하며 의미론적인 기반에서 단어의 원형을 찾음

Bag of words – BOW

- 문서가 가지는 모든 단어를 문맥이나 순서를 무시하고 일괄적으로 단어에 대해 빈도 값을 부여해 피처 값을 추출하는 모델

BOW 피처 벡터화

- 텍스트는 특정 의미를 가지는 숫자형 값인 벡터 값으로 변환해야 하는데, 이런 변환을 피처 벡터화라고 함

사이킷런의 count 및 TF-IDf 벡터화 구현: countvectorizer, tfidfvestorizer

- Countvectorizer는 카운트 기반의 벡터화를 구현한 클래스
- 소문자 일관 변환, 토큰화, 스톱 워드 필터링 등의 텍스트 전처리도 함께 수행
- 텍스트 전처리 및 피처 벡터화를 위한 입력 파라미터를 설정해 동작
- 사이킷런에서 TF-IDF 벡터화는 Tfidfvestorizer 클래스를 이용(파라미터 변환 방법은 동일)

BOW 벡터화를 위한 희소 행렬

- 사이킷런의 countvectorizer, tfidfvestorizer 를 이용해 텍스트를 피처 단위로 벡터화해 변환하고 CSR 형태의 희소 행렬을 반환

희소 행렬 – COO 형식

- 0이 아닌 데이터만 별도의 데이터 배열에 저장하고 그 데이터가 가리키는 행과 열의 위치를 별도의 배열로 저장하는 방식
- 파이썬에서는 희소 행렬 변환을 위해 사이파이를 사용함

희소 행렬 – CSR 형식

- COO 형식이 행과 열의 위치를 나타내기 위해서 반복적인 위치 데이터를 사용해야 하는 문제점을 해결한 방식

감성 분석 소개

- 문서의 주관적인 감성/의견/감정/기분 등을 파악하기 위한 방법으로 소셜 미디어, 여론조사, 온라인 리뷰, 피드백 등 다양한 분야에서 활용
- 문서 내 텍스트가 나타내는 여러 가지 주관적인 단어와 문맥을 기반으로 감성 수치를 계산하는 방법을 이용

지도학습 기반 감성 분석 실습 – IMDB 영화평

- 실습

비지도학습 기반 감성 분석 소개

- Lexicon을 기반으로 함
- 감성을 분석하기 위해 지원하는 감성 어휘 사전으로 감성 사전으로 표현
- 긍정 감성 또는 부정 감성의 정도를 의미하는 수치를 가짐. 이를 감성 지수라고 함

Sentiwardnet

- Synset을 이용
- 실습

Sentiwordnet을 이용한 형화 감상평 감성 분석

- 실습
- Wordnet을 이용해 문서를 다시 단어로 토큰화한 뒤 어근 추출과 품사 태킹을 적용
- 생성한 swn_polarity(text) 함수를 IMDB 감상평의 개별 문서에 적용해 긍정 및 부정 감성을 예측

VADER를 이용한 감성 분석

- VADER는 소셜 미디어의 감성 분석 용도로 만들어진 룰 기반의 lexicon
- 실습