인공지능 API

[한글 데이터 분석]
텍스트 마이닝

한국폴리텍대학 성남캠퍼스 인공지능소트트웨어과
이혜정 교수

1

문서 분석을 위한 전처리

종류	설명
정제cleaning	불필요한 기호나 문자를 제거하는 작업으로 주로 정규식을 이용하여 수행한다.
정규화normalization	정제와 같은 의미지만 형태가 다른 단어를 하나의 형태로 통합하는 작업으로 대/소문자 통합, 유사 의미의 단어 통합 등이 있다.
토큰화tokenization	데이터를 토큰으로 정한 기본 단위로 분리하는 작업이다. 문장을 기준으로 분리하는 문장 토큰화, 단어를 기준으로 분리하는 단어 토큰화 등이 있다.
불용어stopword 제거	의미가 있는 토큰을 선별하기 위해 조사, 관사, 접미사처럼 분석할 의미가 없는 토큰인 불용어stopword를 제거한다.
어간 추출semming	단어에서 시제, 단/복수, 진행형 등을 나타내는 다양한 어간stem을 잘라내어 단어의 형태를 일반화한다.
표제어 추출lemmatization	단어에서 시제, 단/복수, 진행형 등을 나타내는 다양한 표제어lemma를 추출하여 단어의 형태를 일반화한다. 품사를 지정하여 표제어를 추출하는 것이 가능하다.

2

한글 데이터 분석을 위한 기본 개념

▪ 형태소와 형태소 분석

- 언어에서 의미가 있는 가장 작은 단위
- 단어는 의미를 갖는 문장의 가장 작은 단일 요소로, 문장에서 분리될 수 있는 부분
- 독립형 형태소인 단어도 있지만, 대부분의 단어는 형태소와 접사로 구성
- 형태소 분석: 형태소, 어근, 접두사/접미사, 품사 등 다양한 언어학적 속성으로 구조를 파악하는 것

▪ 품사 태깅

- 형태소의 뜻과 문맥을 고려하여 품사를 붙이는 것
  - 예) 가방에 들어가신다 → 가방/NNG + 에/JKM + 들어가/VV + 시/EPH + ㄴ다/EFN

3

한글 데이터 분석 라이브러리 : KoNLPy

▪ KoNLPy에서 사용 가능한 품사 태깅 패키지

- Hannanum, Kkma, Komoran, Mecab, Okt(Twitter) 등

Hannanum	Kkma	Komoran	Mecab	Okt(Twitter)
아버지가방에들어가 / N	아버지 / NNG	아버지가방에들어가신다 / NNP	아버지 / NNG	아버지 / Noun
이 / J	가방 / NNG		가 / JKS	가방 / Noun
시ㄴ다 / E	에 / JKM		방 / NNG	에 / Josa
	들어가 / VV		에 / JKB	들어가신 / Verb
	시 / EPH		들어가 / VV	다 / Eomi
	ㄴ다 / EFN		신다 / EP+EC	

4

텍스트 마이닝 (Text Mining)

- **비정형의 텍스트 데이터로부터 패턴을 찾아내어 의미 있는 정보를 추출하는 분석 과정 또는 기법**
  - 데이터 마이닝과 자연어 처리, 정보 검색 등의 분야가 결합된 분석 기법을 사용
- **텍스트 마이닝의 프로세스**
  - 텍스트 전처리 → 특징 벡터화 → 머신러닝 모델 구축 및 학습/평가 프로세스 수행

5

텍스트의 특징 추출 (특징 벡터화)

- **머신러닝 알고리즘을 적용하여 문서 분석**
  → 문서를 구성하는 단어 기반의 특징 추출
  → 숫자형 값을 갖는 벡터로 특징 표현
  → 특징 벡터 (Feature Vector)
- **특성 벡터화의 대표적인 방법 : BoW와 Word2vec**
  - BOW (Bag of Word)
    - 문서가 가지고 있는 모든 단어에 대해 순서는 무시한 채 빈도만 고려하여 단어가 얼마나 자주 등장하는지로 특성 벡터를 만드는 방법
    - 방식
      - 카운트 기반 벡터화
      - TF-IDF 기반 벡터화

6

카운트 기반 벡터화

- **단어 피처에 숫자형 값을 할당할 때 각 문서에서 해당 단어가 등장하는 횟수(단어 빈도)를 부여하는 벡터화 방식**
  - 문서별 단어의 빈도를 정리
  - 문서 단어 행렬(DTM, Document-Term matrix)을 구성
    - 문서 d에 등장한 단어 t의 횟수는 tf(t,d)로 표현
  - 단어 출현 빈도가 높을수록 중요한 단어로 다루어짐

	그래서	데이터	분석	…	이다	한다
doc#1	13	20	16	…	65	71
doc#2	11	15	32	…	69	81

그림 13-1 카운트 기반 벡터화의 DTM 예: tf("데이터", doc#1) = 20

- **사이킷런의 CountVectorizer 모듈에서 제공**

7

TF-IDF 기반 벡터화

- **Term Frequency – Inverse Document Frequency**
  - 특정 문서에 많이 나타나는 단어는 해당 문서의 단어 벡터에 가중치를 높임
  - 모든 문서에 많이 나타나는 단어는 범용적으로 사용하는 단어로 취급하여 가중치를 낮추는 방식
  - d에 등장한 단어 t의 TF-IDF

$$tf-idf(t,\ d) = tf(t,d) \times idf(t,d)$$

$$idf(t,\ d) = log\frac{n_d}{1+df(d,\ t)}$$

    - (역 문서 빈도)idf(t,d)
    - 는 : 전체 문서의 개수
    - df(d,t)는 단어 t가 포함된 문서 d의 개수

	그래서	데이터	분석	…	이다	한다
doc#1	0.12	0.52	0.42	…	0.19	0.20
doc#2	0.13	0.48	0.67	…	0.18	0.22

그림 13-2 TF-IDF 기반 벡터화의 DTM 예: tf-idf("데이터", doc#1) = 0.52

8

감성 분석 = 오피니언 마이닝

- Sentiment Analysis, Opinion Mining
  - 텍스트에서 사용자의 주관적인 의견이나 감성, 태도를 분석하는 텍스트 마이닝의 핵심 분석 기법 중 하나
  - 텍스트에서 감성을 나타내는 단어를 기반으로 긍정 또는 부정의 감성을 결정
  - 감성 사전 기반의 감성 분석은 감성 단어에 대한 사전을 가진 상태에서 단어를 검색하여 점수를 계산
  - 최근에는 머신러닝 기반의 감성 분석이 늘어나고 있음

9