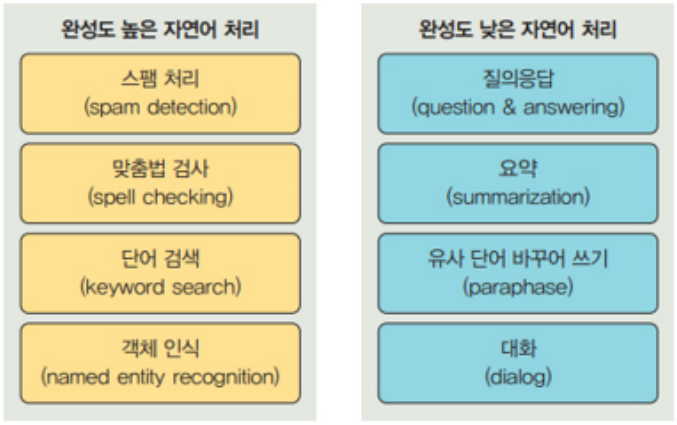
텍스트 분석 이해

1. NLP(National Language Processing)와 텍스트 분석(text analytics)

1) 자연어처리(Natural Language Processing, NLP)

- 머신이 인간의 언어를 이해하고 해석하고 조작(표현)할 수 있도록 연구하고 구현하는 인공지능의 분야
- 1950년대 부터 연구해 옴
- 핵심기술 : 형태소 분석, 구문 분석, 의미분석, 단어 및 문장 생성
- 텍스트분석을 향상하게 하는 기반 기술
- 응용
  - 정보 검색
  - 문서 자동분류
  - 텍스트 요약 (ex: Summly)
  - 대화 시스템 (ex: Apple Siri)
  - 기계 번역 (ex: Google Translate)
  - 챗봇

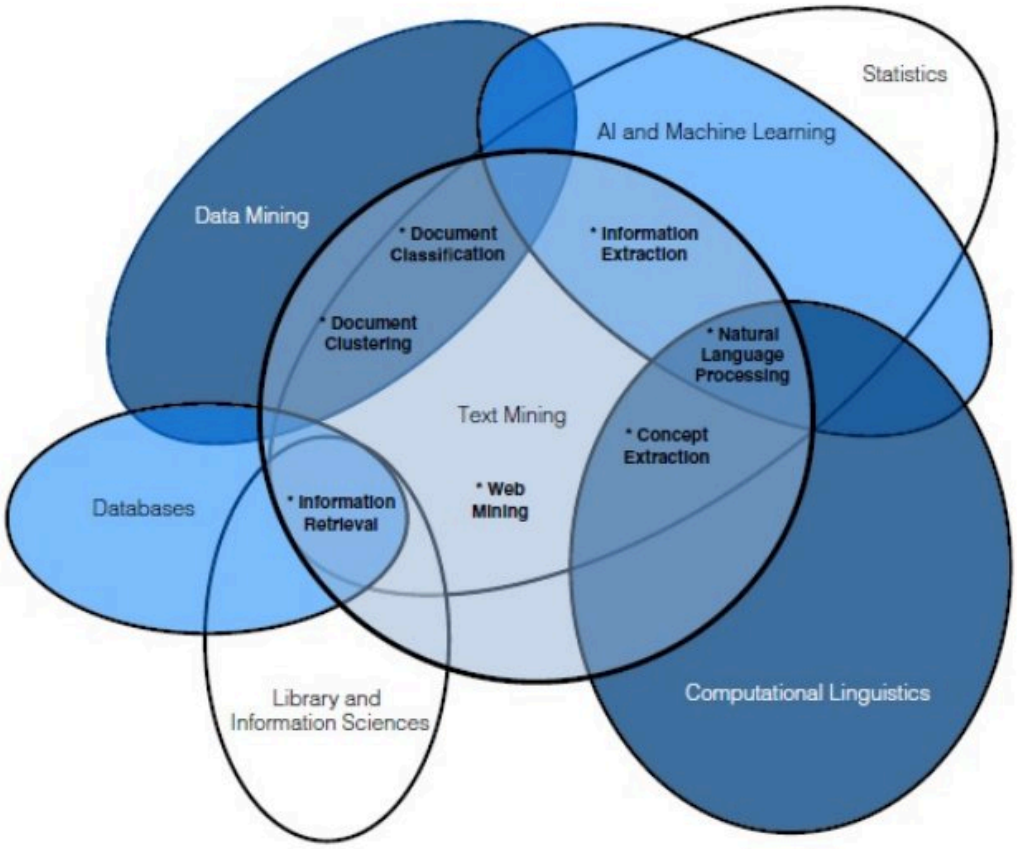
자연어 처리 완성도



2) 텍스트 분석

- 자연어처리 과정을 통해 추출된 단어나 문장 속에서 의미있는 정보(insight)를 추출하는 것에 중점을 둔 영역
- 텍스트 마이닝(text mining)

[그림] 텍스트 마이닝(Text-mining) 과 자연어 처리의 범위 (출처: linguamatics)



https://blog-ko.superb-ai.com/the-hidden-treasures-of-nlp-and-text-mining/

2. 텍스트 분석의 변천

- 과거
  - 텍스트를 구성하는 언어적인 룰이나 업무의 룰에 따라 텍스트를 분석하는 rule-based system
- 현재
  - 머신러닝의 텍스트 데이터를 기반으로 모델을 학습하고 예측

3. 텍스트 분석 기술 영역

: 머신러닝, 언어 이해, 통계 등을 활용해 모델을 수립하고 정보를 추출해 비즈니스 인텔리전스(Business Intelligence)나 예측 분석 등의 분석 작업을 주로 수행함

- 텍스트 분류(Text Classification)
  - Text Categorization
  - 문서가 특정 분류 또는 카테고리에 속하는 것을 예측하는 기법
  - 예.
    - 특정 신문 기사 내용이 어떤 카테고리에 속하는지 자동으로 분류
    - 스팸 메일 검출
  - 지도 학습 적용

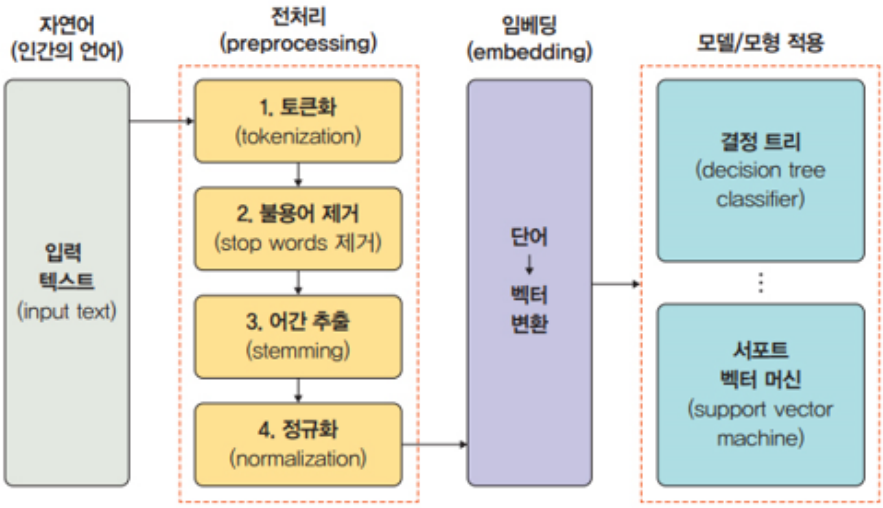
- 감성 분석(Sentiment Analysis)
  - 텍스트에서 나타나는 감정/판단/믿음/의견/기분 등의 주관적인 요소를 분석하는 기법의 총칭
  - 예.

        - 소셜 미디어 감정 분석
        - 영화나 제품에 대한 긍정 또는 리뷰
        - 여론 조사 의견 분석 등
    - 지도/비지도 학습을 이용해 적용

- 텍스트 요약(Summarization)
    - 텍스트 내에서 중요한 주제나 중심 사상을 추출하는 기법
    - 토픽 모델링(Topic Modeling)

- 텍스트 군집화와 유사도 측정
    - 텍스트 군집화
        - 비슷한 유형의 문서에 대해 군집화를 수행하는 기법
        - 텍스트 분류를 비지도학습으로 수행하는 방법의 일환
    - 유사도 측정
        - 문서들간의 유사도를 측정해 비슷한 문서끼리 모을 수 있는 방법

4. 텍스트 분석 수행 프로세스



1단계. 텍스트 사전 준비 작업(텍스트 전처리)

: 텍스트를 피처(feature)로 만들기 전에 텍스트 정규화 작업을 수행하는 것을 통칭

- 대소문자 변경, 특수문자 삭제 등의 클렌징
- 단어 토큰화
- 의미없는 단어(Stop word:불용어) 제거
- 어간 추출(Stemming/Lemmatization)
- 정규화(normalization)

2단계. 피처 벡터화(Feature Vectorization)

- 사전 준비 작업으로 가공된 텍스트에서 피처를 추출하고 여기에 벡터 값을 할당
- 피처 추출(feature extration)이라고도 함
- 텍스트를 word(또는 word의 일부분) 기반의 다수의 피처로 추출하고 이 피처에 단어 빈도수와 같은 숫자 값을 부여하여 단어의 조합인 벡터값으로 표현하는 것
- 피처 벡터화 방법
    - BOW(Bag of Words)

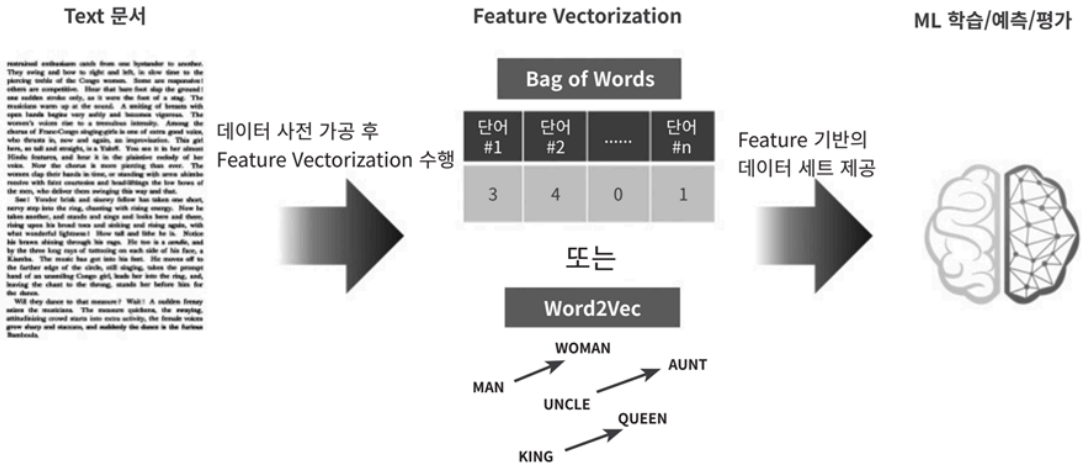
    - Count 기반 벡터화
    - TF-IDF 기반 벡터화
  - Word2Vec

3단계. ML 모델 수립 및 학습/예측/평가

: 피처 벡터화된 데이터 세트에 ML 모델을 적용해 학습/예측 및 평가를 수행

피처 벡터화

- 텍스트 분석은 비정형 텍스트를 분석하는 것
- 텍스트를 머신러닝에 적용하기 위해서 비정형 텍스트 데이터를 어떻게 피처 형태로 추출하고 추출된 피처에 의미있는 값을 부여하는가가 매우 중요한 요소



〈 Text Analytics 수행 프로세스 〉

5. 파이썬 기반의 NLP, 텍스트 분석 패키지

: NLP와 텍스트분석을 쉽고 편하게 하기 위한 라이브러리들

- 텍스트 사전 정제 작업, 피처 벡터화/추출, ML 모델 지원 등

1) 영어기반 라이브러리

- NTLK(Natural Language Toolkit for Python)
  - 파이썬의 가장 대표적인 NLP 패키지(https://www.nltk.org/)
  - 방대한 데이터 세트와 서브 모듈
  - NLP의 거의 모든 영역을 커버
  - 많은 NLP 패키지들이 NLTK의 영향을 받아 작성되고 있음
  - 수행 속도 측면에서 아쉬움
  - 주요 기능 : 말뭉치, 토큰 생성, 형태소 분석, 품사 태깅
  - https://www.nltk.org/book/
- Gensim
  - 토픽 모델링 분야에서 가장 두각을 나타내는 패키지
  - Word2Vec 구현
  - SpaCy 와 함께 가장 많이 사용되는 NLP 패키지
- SpaCy

    - 뛰어난 수행 성능으로 최근 가장 주목을 받는 NLP 패키지
    - 실제 업무에서 자주 활용됨

2) 한글 기반 라이브러리

- KoNLPy(코엔엘파이)
    - 한국어 처리를 위한 파이썬 라이브러리
    - https://konlpy.org/ko/latest/index.html
    - 오픈소스 형태소 분석기
    - 꼬꼬마(Kkma), 코모란(Komoran), 한나눔(Hannanum), 트위터(Twitter), 메카브(Mecab) 분석기 제공

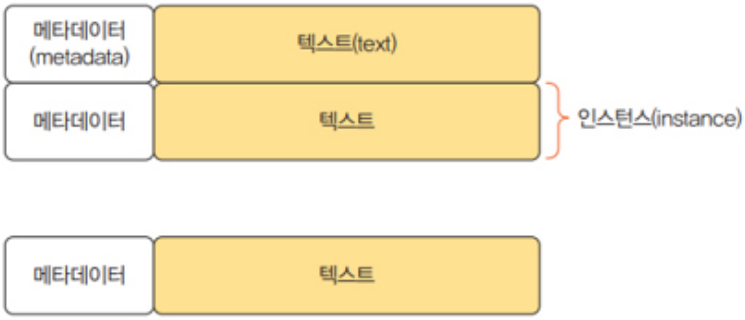
3) 사이킷런(Scikit_Learn)의 머신러닝 알고리즘

- NLP를 위한 어근 처리와 같은 다양한 라이브러리는 가지고 있지 않음
- 텍스트를 일정 수준으로 가공하고 머신러닝 알고리즘에 텍스트 데이터를 피처로 처리하기 위한 편리한 기능 제공
- 텍스트 분석 기능은 충분히 수행 가능
- 주요 기능
    - CountVectorizer : 텍스트에서 단어의 등장 횟수를 기준으로 특성 추출
    - Tfidfvectorizer : TF-IDF 값을 이용해서 텍스트의 특성 추출
    - HashingVectorizer : CountVectorizer의 방법과 동일하나 텍스트 처리시 해시 함수를 사용하여 실행시간이 감소됨

6. 관련 용어들

말뭉치(corpus: 코퍼스)

- 자연어 처리에서 모델을 학습시키기 위한 데이터
- 자연어 연구를 위해 특정한 목적에서 표본을 추출한 집합



토큰(token)

- 문서를 나누는 단위
- 문자열을 토큰으로 나누는 작업을 토큰 생성(tokenizing)이라고 함
- 문자열을 토큰으로 분리하는 함수를 토큰 생성 함수라고 부름

토큰화(tokenization)

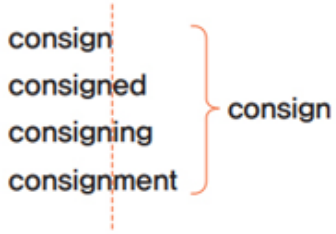
- 텍스트를 문장이나 단어로 분리하는 것
- 토근화 단계를 마치면 텍스트가 단어 단위로 분리됨

불용어(stop words)

- 문장 내에서 많이 등장하는 단어로 분석과 관계없으며 자주 등장하는 빈도로 성능에 영향을 미치는 단어들
- 사전에 제거해 주어야 함
- 예. 'a', 'the', 'she', 'he'

어간 추출(stemming)

- 단어를 기본 형태로 만드는 작업
- 예. 'consign', 'consigned', 'consigning', 'consignment' 등을 기본 단어인 'consign'으로 통일



품사 태깅(part-of-speech tagging)

- 주어진 문장에서 품사를 식별하기 위해 붙여주는 태그(식별 정보)
- 영어 품사의 예.
  - Det : 한정사
  - Noun : 명사
  - Verb : 동사
  - Prep : 전치사