2021.03.26

[한글 형태소 분석]

1. 자연어(사람들이 일상적으로 사용하는 언어) 처리 절차

: 전처리(단어, 어절 추출) -> 분석 후보 생성 - > 제약조건 규칙 확인 - > 분석

2. 한글 형태소 분석 엔진

- **KoNLPy** : 파이썬용 자연어 처리기 (JPype1 패키지를 의존)
  - **KOMORAN** : 자바로 만든 형태소 분석기 (JAVA_HOME 시스템 변수 필요)
  - **HanNanum** : 자바로 만든 형태소 분석기 (JAVA_HOME 시스템 변수 필요)
  - **Kkma** : 서울대학교 연구실에서 만든 형태소 분석기
- **KoNLP** : R 용 자연어 처리기

ex) HanNanum 분석기로 단어 분석하기

```
from konlpy.tag import Hannanum      # Hannanum을 사용하기 위해 필요한 패키지 import
han = Hannanum( )                    # Hannanum Class 객체 생성
han.analyze( )                       # 텍스트 분석
han.morphs( )                        # 형태소 분석
han.pos( )                           # 품사 태깅
han.nouns( )                         # 명사만 추출하기
```

3. 대표적인 형태소 분석 관련 용어들과 기법들

(1) 말뭉치: 컴퓨터를 이용해 자연어 분석 작업을 할 수 있도록 만든 문서의 집합

(2) 불용어(stepword): 필요한 단어를 추출하는 작업에서 제거해도되는 불필요한 단어들

(3) 워드 클라우드: 단어를 출현 빈도에 비례하는 크기로 단어의 빈도수를 시각화하는 기법

(4) 워드 임베딩(Word Embedding): 단어를 벡터로 표현하는 대표적인 방법

ex) 워드 클라우드

```
from wordcloud import WordCloud              # 워드클라우드를 위한 패키지 import
from wordcloud import STOPWORDS              # 불용어 설정을 위한 패키지 import
# 워드클라우드 생성
word = WordCloud(background_color='white'    # 배경색 지정
                 font_path='c:/Windows/Fonts/H2PORM.TTF', # 파일에서 글씨체 불러오기
                 max_words=20,               # 들어갈 최대 단어 수
                 stopwords = 불용어변수명)   # 설정한 불용어 추가
word.generate( )                             # 워드 클라우드에 적용할 텍스트 불러오기
plt.figure( )
plt.imshow(word, interpolation = 'bilinear')  # 'bilinear': 보강법
plt.axis('off')                              # 축 숫자 없애기
```