4. 형태소 분석

한국어는 명사 뒤에 조사, 동사 뒤에 어미(예: -다)가 붙는다. 명사나 동사와 같은 단어로 분석을 하려면 이런 조사나 어미를 떼어낼 필요가 있다. 이를 위한 과정을 형태소 분석이라고 한다. 형태소(morpheme)란 언어에서 의미가 있는 가장 작은 단위를 말한다.

형태소 분석을 해주는 소프트웨어를 형태소 분석기라고 하는데, 공개된 형태소 분석기도 한나눔, 꼬꼬마, 코모란, 트위터 4가지가 있다. 이들은 모두 한국어 문법 규칙을 바탕으로 형태소 분석을 한다. 일본에서 만든 mecab은 기계학습을 통한 형태소 분석기이다. 파이썬에는 KoNLPy 패키지를 통해 이들 형태소 분석기를 모두 사용할 수 있다.

4.1. KoNLP 설치

4.1.1. 리눅스 및 맥

리눅스나 맥의 경우 KoNLPy는 다음과 같은 명령으로 간단히 설치할 수 있다.

```
pip install konlpy
```

4.1.2. 윈도

Java 설치

윈도의 경우 Java설치가 필요하다. 자바는 다음 링크에서 다운받을 수 있다.
https://java.com/ko/download/manual.jsp

Python이 64비트면 Java도 64비트여야 한다. 32비트인 경우도 같다.

환경변수 설정

다음으로 환경변수 설정이 필요하다. 환경변수는 다음과 같이 설정할 수 있다.

1. `Win + Pause` 를 눌러 시스템 메뉴를 연다
2. "고급 시스템 설정"을 연다 (`Win` 키 클릭 후 "고급 시스템 설정"을 타이핑해서도 열 수 있다)
3. "환경변수" 버튼 클릭
4. 상단 "사용자 변수"에서 새로 만들기나 편집으로 추가/수정할 수 있다

Java를 설치하면 `C:\Program Files\Java\jre` 와 같은 폴더에 Java가 설치된다. 해당 폴더의 경로를 `JAVA_HOME` 으로 설정한다.

다음으로 `PATH` 환경변수에 `%JAVA_HOME%\bin;` 을 추가한다.

JPype1 설치

Unofficial Windows Binaries for Python Extension Packages에서 jpype를 설치한다. 파일명에서 `cp36m-win32` 는 Python 3.6 32비트용, `cp36m-win_amd64` 는 Python 3.6 64비트용이다. 자신의 Python에 맞게 다운로드를 받는다.

다운로드 폴더에서 명령창을 띄운 후 다음과 같이 입력한다. JPype1 까지 입력 후 `Tab` 을 누르면 자동 완성이 되므로 파일명은 알맞게 입력한다.

```
pip install JPype1-0.6.2-cp36-cp36m-win_amd64.whl
```

KoNLPy 설치

마지막으로 KoNLPy를 설치한다.

```
pip install konlpy
```

4.2. 형태소 분석

코모란을 이용해 형태소 분석을 실시해보자.

```
from konlpy.tag import Komoran
tagger = Komoran()  # 형태소 분석기
tagger.pos('한국어로 텍스트를 분석해 봅시다.')  # 형태소 분석
```

```
[('한국어', 'NNP'),
 ('로', 'JKB'),
 ('텍스트', 'NNG'),
 ('를', 'JKO'),
 ('분석', 'NNG'),
 ('하', 'XSV'),
 ('아', 'EC'),
 ('보', 'VX'),
 ('ㅂ시다', 'EF'),
 ('.', 'SF')]
```

많은 경우 명사만을 뽑아 텍스트 분석을 실시한다. 코모란의 경우 `.nouns` 메소드로 명사만을 추출할 수 있다.

```
tagger.nouns('한국어로 텍스트를 분석해 봅시다.')  # 명사 추출
['한국어', '텍스트', '분석']
```