한글, 영어 텍스트 전처리

In [ ]:

```
import konlpy
import jamo
import nltk
```

한글 전처리

- Tokenize : konlpy.tag.Okt(), Hannanum(), Kkma() 등 사용가능.
- 자모 분리 : jamo.h2j()

In [ ]:

```
kor_string = "아버지가 방에 들어 가신닭ㅋ"
okt = konlpy.tag.Okt()
print(okt.morphs(kor_string))

print(okt.nouns(kor_string))

print(okt.pos(kor_string))
print(okt.pos(kor_string, norm=True))
print(okt.pos(kor_string, stem=True))
```

In [ ]:

```
han = konlpy.tag.Hannanum()
print(han.analyze(kor_string))

print(han.morphs(kor_string))

print(han.nouns(kor_string))

print(han.pos(kor_string))
```

In [ ]:

```
kkma = konlpy.tag.Kkma()
print(kkma.morphs(kor_string))

print(kkma.nouns(kor_string))

print(kkma.pos(kor_string))

kor_string_2 = '아버지 방에 들어 가신다. 그러니까 TV나 계속 보자.'
print(kkma.sentences(kor_string_2))
```

In [ ]:

```
words = okt.morphs(kor_string)
print(words)
jm = [jamo.h2j(word) for word in words]
print(jm)
```

영어 전처리

```
* Tokenize: nltk.word_tokenize()
* Stemming: nltk.stem.PorterStemmer()
* Lemmatize:nltk.stem.WordNetLemmatizer()
* Stop words removal: nltk.corpus.stopwords
* Others: nltk.pos_tag()
```

In [ ]:

```
eng_string = 'Ten years had passed since the showery day in late summer when Lord Moping had been ta

words = nltk.word_tokenize(eng_string)
print(words)

for word in words:
    print('Stem: ', nltk.stem.PorterStemmer().stem(word))
    print('Lemm: ', nltk.stem.WordNetLemmatizer().lemmatize(word))

stop_words = list(nltk.corpus.stopwords.words('english'))
print(len(stop_words))
print(stop_words[:5])
print(stop_words[-5:])

print(nltk.pos_tag(words))
```