2021.03.25

[NLTK(영어 자연어 처리 패키지)]

1. NLTK 패키지 불러오기

```
import nltk          # nltk 패키지 불러오기
nltk.download()      # nltk의 파일들 다운로드하기(실행 후 새 창이 열리면 경로를 설정한다.)
```

2. 텍스트 분리하기(tokenizing)

문장 단위로 분리	from nltk.tokenize import sent_tokenize sent_tokenize( )
단어 단위로 분리	from nltk.tokenize import word_tokenize word_tokenize( )
정규 표현식을 이용한 분리	from nltk.tokenize import RegexpTokenizer ret = RegexpTokenizer("정규표현식") ret.tokenize( )

3. 형태소(의미가 있는 가장 작은 말의 단위)를 분석하는 방법

어간 추출(Stemming)	- 방법1: PorterStemmer 클래스 사용 - 방법2: LancasterStemmer 클래스 사용 - 방법3: RegexpStemmer 클래스 사용 (특정 어미를 설정함)
원형 복원(Lemmatizing)	어간 추출을 하면 의미가 달라질 수 있으므로 단어를 원래의 형태로 복원한다.
품사 태깅 (Part of Speech Tagging)	단어를 분리하여 각 단어마다 해당하는 품사를 표기하여 출력한다.

ex) – 품사 태깅

```
from nltk.tag import pos_tag                          # 품사 태깅을 위해 필요한 패키지 import
tag_list = pos_tag(word_tokenize(sent_tokens[10])) # tag_list변수에 자료의 단어를 분리하여 품사 태깅
print(tag_list) # 결과 출력  => [('단어', '품사')]의 형식으로 출력됨
```

```
품사 태깅 결과
[('It', 'PRP'), ('was', 'VBD'), ('on', 'IN'), ('the', 'DT'), ('wedding-da
y', 'NN'), ('of', 'IN'), ('this', 'DT'), ('beloved', 'VBN'), ('friend', 'N
N'), ('that', 'WDT'), ('Emma', 'NNP'), ('first', 'RB'), ('sat', 'VBD'),
('in', 'IN'), ('mournful', 'JJ'), ('thought', 'NN'), ('of', 'IN'), ('any',
'DT'), ('continuance', 'NN'), ('.', '.')]
```