<<<<파일의 자연어처리 건>>>>

**목표: 파일 하나의 자연어를 읽어 거기서 유의미한 일정 추출하기!!!!**

1. **주피터 자연어처리 파일을 파이참으로 옮겨서….**
2. **ㅅㅂ 그럼 나중에 어떻게 자바로 옮기냐 시발ㅗ’**

Work flow:

Data 구하기(한글로 된 강의계획서 3000개 모아오기. -> DB 저장)

전처리(konlpy 패키지의 twitter형태소 분석기를 이용해 stemming하기 -> 인코딩하기 -> 어절 N gram 이용)????

분석(속성별 키워드 선정 by LDA -> 키워드 포함된 N gram 추출 -> countVectorizer, anova 분석이용해서 속성별 단어 분류)

검증(수동으로 단어 분류해서 결과와 비교하여 성능평가)

자연어 처리 단계:

전처리

Tokenizing(어절 단위, 형태소 단위, n gram 단위, wordPiece 단위)

* 영어라면 문자 모두 대문자 / 소문자로 변환하기 + 글의 공백 제거 + 공백 기준으로 분리, 탭 기준 분리, 라인 단위 분리 / 숫자 여부 판별, 영어 여부 판별, 대소문자 여부 판별, 공백 여부 판별 등.

Lexical analysis(어휘 분석, 형태소 분석, 개체명 인식, 상호 참조)

Syntactic analysis: 구문 분석

Semantic analysis: 의미 분석

* 한국어는 조사 어미를 붙여 말을 만드는 교착어. -> 조사 어미를 분리하는 형태소 분석을 통해 토큰화 하기.

전처리: 긁어온 언어를 용도에 맞게 토큰화하고 정제하기.

* 단어 토큰화:

1. 영어: New York같은 합성어나 it’s 같은 줄임말만 예외처리하면 띄어쓰기 기준으로 잘 동작한다.

**from** nltk.tokenize **import** word\_tokenize

print(word\_tokenize("Don't be fooled by the dark sounding name, Mr. Jone's Orphanage is as cheery as cheery goes for a pastry shop."))

혹은

**from** nltk.tokenize **import** WordPunctTokenizer (이 방식은 구두점도 분리함.)

print(WordPunctTokenizer().tokenize("Don't be fooled by the dark sounding name, Mr. Jone's Orphanage is as cheery as cheery goes for a pastry shop."))

혹은

**from** tensorflow.keras.preprocessing.text **import** text\_to\_word\_sequence (이렇게 케라스를 쓰면 아포스트로피 보존하면서 다른 마침표, 컴마 등의 구두점은 제거.)

print(text\_to\_word\_sequence("Don't be fooled by the dark sounding name, Mr. Jone's Orphanage is as cheery as cheery goes for a pastry shop."))

아래는 문장 단위로 잘라내는 방식.

**from** nltk.tokenize **import** sent\_tokenize

text = "His barber kept his word. But keeping such a huge secret to himself was driving him crazy. Finally, the barber went up a mountain and almost to the edge of a cliff. He dug a hole in the midst of some reeds. He looked about, to make sure no one was near."

print(sent\_tokenize(text))

1. 한국어 처리하기.

문장별로 잘라내는 건 아래.

**import** kss text = '딥 러닝 자연어 처리가 재미있기는 합니다. 그런데 문제는 영어보다 한국어로 할 때 너무 어렵습니다. 이제 해보면 알걸요?'

print(kss.split\_sentences(text))

**from** konlpy.tag **import** Okt

okt = Okt()

print(okt.morphs("열심히 코딩한 당신, 연휴에는 여행을 가봐요")) (단어 별로 잘라낸다.)

**print**(okt.**pos**("열심히 코딩한 당신, 연휴에는 여행을 가봐요")) (단어별로 자르고 품사 구분해줌.)

**print**(okt.nouns("열심히 코딩한 당신, 연휴에는 여행을 가봐요")) (이근 예시인데 noun 품사만 추출해서 보여줌.)