#!/usr/bin/env python

# coding: utf-8

# In[18]:

import numpy as np

import pandas as pd

import re

import os

import pickle

from konlpy.tag import Okt

from gensim import models

from gensim.models import Word2Vec

import tensorflow as tf

from tensorflow import keras

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

def tokenize(sentence):

okt = Okt()

tokenized\_sentence = []

# 우선 단어의 기본형으로 모두 살리고, 명사, 동사, 영어만 담는다.

# 그냥 nouns로 분리하는 것보다 좀 더 정확하고 많은 데이터를 얻을 수 있다.

for line in sentence:

result = []

temp\_sentence = okt.pos(line, norm=True, stem=True) # 먼저 형태소 분리해서 리스트에 담고

for i in temp\_sentence:

if (i[1] == 'Noun' or i[1] == 'Adjective' or i[1] == 'Alpha'):

result.append(i[0])

tokenized\_sentence.append(result)

return tokenized\_sentence