import json

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

from nltk.stem import WordNetLemmatizer

from nltk.tag import pos\_tag

import re

n=WordNetLemmatizer()

tx1 = ''

tx2 = ''

#print('\*\*\*original text\*\*\*\n')

with open('final\_review.json', 'rt', encoding='UTF-8') as f:

for line in f:

lineobj = json.loads(line)

tx1 = lineobj['text']

tx3 = tx1.lower()

word\_tokens = word\_tokenize(tx3)

result = []

words = []

#print(word\_tokens)

#print('\n\n\*\*\*lemmatize\*\*\*\n')

tag\_l = pos\_tag(word\_tokens)

for w in tag\_l:

if w[1][:2] == 'VB':

words.append(n.lemmatize(w[0], 'v'))

else:

words.append(n.lemmatize(w[0]))

#print(words)

stop\_words = set(stopwords.words('english'))

stop\_words.update(['\'re', '\'d', '\'t', '\'ll', '\'ve', '\'s','\'m', '!', '.', ',', '/', '?', '\"', '@', '%', '&', '\*', '=','(',')','{','}', '-', '--', 'u', '...','$', '#', '\*', '@', ':', ';', '[',']','~'])

stop\_words.remove('not')

for w in words:

if w == 'n\'t':

w = 'not'

if w not in stop\_words:

a = re.sub('[^a-zA-Z ]', '', w)

if a != '':

result.append(a) #result에 불용어 제거된 tokens 존재

#print('\n\n\*\*\*result\*\*\*\n', result)

with open('final\_review\_tokens.txt', 'a', encoding='UTF-8') as mf:

mf.write('[\n')

mf.write('\n'.join(result))

mf.write('\n]\n')