[Code-Korean] Tokenize Sentences

Code Snippet

```
# package import
import pandas as pd
import re
from konlpy.tag import Mecab # Korean Tokenizer, Do not import at other languages

# Basic Cleaning Text Function
def CleanText(readData):

    # Remove Retweets
    text = re.sub('RT @[\w_]+: ', '', readData)

    # Remove Mentions
    text = re.sub('@[\w_]+', '', text)

    # Remove or Replace URL
    # text = url_re.sub('URL', text)
    text = re.sub(r"http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+", ' ', text) # start with http
    text = re.sub(r"[-a-zA-Z0-9@:%._\+~#=]{1,256}\.[a-zA-Z0-9()]{2,6}\b([-a-zA-Z0-9()@:%_\+.~#?&//=]*)", ' ', text) # Don't start wi

    # Remove Hashtag
    text = re.sub('[#]+[0-9a-zA-Z_]+', ' ', text)

    # Remove Garbage Words (ex. &lt, &gt, etc)
    text = re.sub('[&]+[a-z]+', ' ', text)

    # Remove Special Characters
    text = re.sub('[^0-9a-zA-Zㄱ-ㅎ가-힣]', ' ', text)

    # Remove Numbers (If you want, activate the code)
    # text = re.sub(r'\d+',' ',text)

    # Remove English (If you want, activate the code)
    # text = re.sub('[a-zA-Z]' , ' ', text)

    # Remove newline
    text = text.replace('\n',' ')

    # Remove multi spacing & Reform sentence
    text = ' '.join(text.split())

    # If you want to normalize Korean text, activate code below:
    # from konlpy.tag import Okt # Must use Konlpy ver 0.5.2 above
    # okt = Okt()
```

```
    # if len(text) != 0:
      # text = okt.normalize(text)

    return text

# Korean Stopwords Load
korean_stopwords = pd.read_csv("./korean_stopwords.txt", delimiter='\t', names=["형태", "품사", "비율"])

# Add Custom Korean Stopwords
my_data = [["님", "NNG"], ["들", "XSN"], ["ㅋㅋㄱㅋㄱㅋ", "NNG"],
           ["오늘", "NNG"], ["얘기", "NNG"], ["ㅠㅠ", "NNG"], ["없이", ], ["딱히", ],
           ['ㅋㅋ', ], ['ㅋㅋㅋ', ], ["그런데", ], ["누구", ], ["여기저기", ]]
my_korean_stopwords = pd.DataFrame(my_data, columns = ['형태', '품사'])

# Main
mecab = Mecab(dicpath="/usr/local/lib/mecab/dic/mecab-ko-dic") # Mecab Dictionary Path

def preprocessing_mecab(readData):

    #### Clean text
    sentence = CleanText(readData)

    #### Tokenize
    morphs = mecab.pos(sentence)

    JOSA = ["JKS", "JKC", "JKG", "JKO", "JKB", "JKV", "JKQ", "JX", "JC"]
    SIGN = ["SF", "SE", "SSO", "SSC", "SC", "SY"]
    TERMINATION = ["EP", "EF", "EC", "ETN", "ETM"] # 어미
    SUPPORT_VERB = ["VX"]
    NUMBER = ["SN"]

    # Remove JOSA, EOMI, etc
    morphs[:] = (morph for morph in morphs if morph[1] not in JOSA+SIGN+TERMINATION+SUPPORT_VERB)

    # If you want to save only Nouns:
    # morphs = mecab.nouns(sentence)

    # Remove Stopwords
    morphs[:] = (morph for morph in morphs if morph[0] not in korean_stopwords["형태"].tolist())
    morphs[:] = (morph for morph in morphs if morph[0] not in my_korean_stopwords["형태"].tolist())

    # Remove length-1 words
    morphs[:] = (morph for morph in morphs if not (len(morph[0]) == 1))

    # Remove Numbers
    morphs[:] = (morph for morph in morphs if morph[1] not in NUMBER)

    # Result pop-up
    result = []
    for morph in morphs:
        result.append(morph[0])

    return result

print(f"Before preprocessing : {SAMPLE_TEXT}")
>>> Before preprocessing : RT @boxplus01: 美언론, '한국 코로나 확산주범은 신천지와 보수세력' https://t.co/Phq0l48aUm

print(f"After preprocessing : {preprocessing_mecab(SAMPLE_TEXT)}")
>>> After preprocessing : ['언론', '한국', '코로', '확산', '주범', '신천지', '보수세력']
```

- For Korean, MeCab-Ko was used as a morpheme analyzer (issue: the MeCab tokenizer is highly dependent on the dictionaries)
- Added Twitter-speciic functions (RT, Mention, Hashtag, and so on)
- Used stopwords from dictionaries and customed ones
- Added text normalization codes