# -\*- coding: utf8 -\*-  
  
from konlpy.tag import Okt  
from hanspell import spell\_checker  
from konlpy.tag import Kkma  
from soynlp import DoublespaceLineCorpus  
from soynlp.word import WordExtractor  
from soynlp.tokenizer import LTokenizer  
import kss  
from kss import split\_sentences  
from wordcloud import WordCloud  
from konlpy.tag import Twitter  
from collections import Counter  
import matplotlib.pyplot as plt  
  
  
texts = [ "조용하고 공부에 집중하기 좋은 곳 이예요 !!!!!", "서현 스터디 카페중에 젤 조음", "진짜 분위기 좋다 알바생분들도 친절 ㅠㅠ", "편안한 분위기여서 공부 잘되요", "내부도 적당히 크고 자리마다 무선 충전기 설치 되어 있는 등 시설이 되게 좋아요!!제공되는 것도 많구요 건물이 낡아서 화장실이 후지긴 한데 어쩔 수 없죠ㅠ음료 포함 정기권 끊었는데 음식 종류도 많구 아직 많이 먹어보진 않았지만 맛있어요!볶음밥이랑 베이커리류도 팔아서 식사 해결도 가능해요ㅎㅎ다만 음료는 만들어주시는 알바 분들에 따라 맛 차이가 큰거 같아요. 그리고 창가 쪽에 앉으면 비둘기 등등의 각종 새들이 자꾸 와서 달려 있는 간판 위에 앉고 부딪히고 막 그래서 소리가 좀 나요ㅠㅠ구구구하고 우는 소리도 들리고 날아가는 소리,부딪히는 소리 전부 잘 들리는게 조금 거슬린다는 점 감안해야 해요!" ]  
  
i=0  
for x in texts :  
 i=i+1  
 print("<" ,i,">\n")  
 print(" 원래 문장 : ", x)  
  
 t= kss.split\_sentences(x)  
 print("문장토큰화 : ",t)  
  
 spelled\_sent = spell\_checker.check(x) #t(토큰화된 문장)넣으면 에러  
 hanspell\_sent = spelled\_sent.checked  
 print(" 맞춤법 검사 후 : ", hanspell\_sent)  
  
 okt = Okt()  
  
 print("\*\*konlpy\_okt\*\*")  
 print(" <형태소 추출> ")  
 print(okt.morphs(hanspell\_sent))  
 print(" <품사 태깅> ")  
 print(okt.pos(hanspell\_sent))  
 print(" <명사 추출> ")  
 print(okt.nouns(hanspell\_sent))  
  
 # kkma = Kkma()  
 #  
 # print("\*\*konlpy\_kkma\*\*")  
 # print(" <형태소 추출> ")  
 # print(kkma.morphs(hanspell\_sent))  
 # print(" <품사 태깅> ")  
 # print(kkma.pos(hanspell\_sent))  
 # print(" <명사 추출> ")  
 # print(kkma.nouns(hanspell\_sent))  
 #  
 # print("\*\* soynlp\_ LTokenizer \*\*")  
 # corpus = DoublespaceLineCorpus("study\_cafe\_review.csv") # 학습할 문서  
 #  
 # word\_extractor = WordExtractor()  
 # word\_extractor.train(corpus)  
 # word\_score\_table = word\_extractor.extract() # 학습 완료  
 #  
 # scores = {word: score.cohesion\_forward for word, score in word\_score\_table.items()}  
 # l\_tokenizer = LTokenizer(scores=scores)  
 # print(l\_tokenizer.tokenize(hanspell\_sent, flatten=False))  
 #  
 # print("\n\n")  
  
  
  
noun\_list=okt.nouns(hanspell\_sent)  
  
counts = Counter(noun\_list)  
tags = counts.most\_common(40)  
  
wc = WordCloud(font\_path="malgun.ttf",background\_color="white", max\_font\_size=60)  
cloud = wc.generate\_from\_frequencies(dict(tags))  
  
cloud.to\_file('test.jpg')  
  
plt.figure(figsize=(10, 8))  
plt.axis('off')  
plt.imshow(cloud)  
plt.show()

텍스트이(가) 표시된 사진

자동 생성된 설명