워드클라우드

1. 데이타 가져오기

In [ ]:

```
f=open('[박경리]_토지1.txt','r')
story=f.read()
#print(story)
print(type(story))
```

2. 워드클라우드 그리기

In [ ]:

```
!pip list
```

In [ ]:

```
!pip install wordcloud
```

In [ ]:

```
!pip show wordcloud
```

- 워드클라우드에서 제외하고 싶은 단어

In [ ]:

```
stopwords_kr = ['하지만', '그리고', '그런데', '저는','제가','그럼', '이런', '저런', '합니다',
                '많은', '많이', '정말', '너무','[',']','것으로','했습니다','했다','있다']
```

- 워드클라우스 설정

In [ ]:

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
%matplotlib inline

def displayWordCloud(data = None, backgroundcolor = 'white', width=800, height=600 ):
    wordcloud = WordCloud(
                          #font_path = '/Library/Fonts/NanumBarunGothic.ttf',
                          font_path = '/Library/Fonts/NanumPen.ttf',
                          stopwords = stopwords_kr,
                          background_color = backgroundcolor,
                          width = width, height = height).generate(data)
    print(wordcloud.words_)
    plt.figure(figsize = (15 , 10))
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.show()
```

In [ ]:

```
# 결과를 출력해 보면 불용어(STOPWORD)가 너무 많습니다.
%time displayWordCloud(story)
```

3. 불용어 제거

- 문장에서 명사를 추출합니다.
- https://github.com/lovit/soynlp (https://github.com/lovit/soynlp)

In [ ]:

```
!pip install soynlp
!pip show soynlp
```

In [ ]:

```
from soynlp.noun import NewsNounExtractor
```

In [ ]:

```
%%time
noun_extractor = NewsNounExtractor()
nouns = noun_extractor.train_extract([story])
```

- list = str.split() : 문자열 => 리스트, 공백시 스페이스 기준
- ' '.join(list) : 리스트에서 문자열으로

In [ ]:

```
print(type(nouns))
```

In [ ]:

```
# 추출된 명사를 찍어봅니다.
%time displayWordCloud(' '.join(nouns))
```

4. 특정 이미지 형태로 워드 클라우드 그리기

In [ ]:

```
# 이미지 파일위에 출력하기
from PIL import Image
import numpy
```

In [ ]:

```
img = Image.open('cloud.png')
img_array=numpy.array(img)
```

In [ ]:

```
wordcloud = WordCloud( font_path = '/Library/Fonts/CookieRun Black.ttf',
                       stopwords = stopwords_kr,
                       background_color = 'white',
                       mask=img_array,
                       width = 800, height = 600).generate(' '.join(nouns))
plt.figure(figsize = (15 , 10))
plt.imshow(wordcloud)
plt.axis("off")
plt.show()

# 이미지로 결과 저장
wordcloud.to_file("simple.png")
```