주제분석 1주차 패키지

1. 패키지 불러오기

In [10]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
from konlpy.tag import Okt
from collections import Counter
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
import os
import datetime
```

2. 뉴스 크롤링

1) 뉴스 목록의 링크 크롤링

In [23]:
```
href_list = [] #### 링크를 담을 리스트를 만듭니다.
date_list = [] #### 뉴스 발행 날짜를 담을 리스트를 만듭니다.

for i in range(0, 2):
    url = 'https://www.mk.co.kr/news/stock/?page=' + str(i) #### 페이지 수를 돌아가며
    response = requests.get(url)
    html = response.text
    soup = BeautifulSoup(html, 'html.parser')
    a = soup.select('#container_left > div.list_area > dl > dt > a') #### 뉴스 기사의
    for j in range(0, len(a)):
        href = a[j]['href']
        href_list.append(href)
    b = soup.select('#container_left > div.list_area > dl > dd.desc > span.date') ###
    for k in range(0, len(b)):
        date = b[k].get_text()
        date_list.append(date)
```

In [24]:
```
df = pd.DataFrame({'date':date_list, 'href':href_list}) #### 크롤링한 결과를 데이터프
```

In [25]:
```
df.head()
```

Out[25]:

	date	href
0	2021.10.16 15:10	https://www.mk.co.kr/news/stock/view/2021/10/9...
1	2021.10.16 05:32	https://www.mk.co.kr/news/stock/view/2021/10/9...
2	2021.10.15 20:50	https://www.mk.co.kr/news/stock/view/2021/10/9...
3	2021.10.15 20:22	https://www.mk.co.kr/news/stock/view/2021/10/9...
4	2021.10.15 17:30	https://www.mk.co.kr/news/stock/view/2021/10/9...

2) 뉴스 기사 제목, 본문 크롤링

In [26]:
```
title_list = [] #### 뉴스기사의 제목 리스트를 만듭니다.
txt_list = [] #### 뉴스기사의 내용을 담을 리스트를 만듭니다.
for i in range(0, 50):
    try :
        url = df.href[i]
```

```
        response = requests.get(url)
        html = response.text
        soup = BeautifulSoup(response.content.decode('euc-kr', 'replace'), "html.pars
        txt = soup.select('#article_body > div')[0].get_text() #### 뉴스기사의 본문을
        title = soup.select('#top_header > div > div > h1')[0].get_text() ### 뉴스기사

        title_list.append(title)
        txt_list.append(txt)
    except :
        title_list.append('error')
        txt_list.append('error')
```

In [27]:
```
df['title'] = title_list
df['txt'] = txt_list
```

In [28]:
```
df['date'] = [i[:10] for i in df['date']]
df['date'] = pd.to_datetime(df['date'])
```

In [29]:
```
df.head()
```

Out[29]:

	date	href	title	txt
0	2021-10-16	https://www.mk.co.kr/news/stock/view/2021/10/9...	위드 코로나 기대감·인플레이션 우려에 횡보세 지속 [주간 증시전망]	\n\n\n코스피가 전날보다 23.98p(0.80%) 오른 3,012.62로 시작한...
1	2021-10-16	https://www.mk.co.kr/news/stock/view/2021/10/9...	뉴욕증시, 소매판매·실적 호조에 상승... 다우 1.09%↑ 마감	\n\n\n\n다우 1.09%↑ 마감 (PG)\n▶ 여기를 누르시면 크게 보실 수 ...
2	2021-10-15	https://www.mk.co.kr/news/stock/view/2021/10/9...	"묻고 더블, 아니 3배로 가"...초고위험 ETF에 몰리는 불개미들	\n\n\n[UPI = 연합뉴스]\n 미국 증시가 최근 주춤한 모습을 보이는 가운데...
3	2021-10-15	https://www.mk.co.kr/news/stock/view/2021/10/9...	"카카오 12만원 회복했지만 고민되네"...증권사 목표주가 줄줄이 하향	\n\n\n코스피가 3,000선을 회복한 15일 오후 서울 중구 하나은행 딜링룸 전...
4	2021-10-15	https://www.mk.co.kr/news/stock/view/2021/10/9...	코스피 한숨 돌렸지만...추가반등은 글쎄	\n\n\n 14일 코스피가 8거래일 만에 3000선을 회복했지만 전문가들은 "유의...

In [30]:
```
df.to_csv("C:/Users/User/Desktop/3주차 패키지/결과물예시파일.csv", index=False, encoc
```

3. 워드클라우드

1) 워드클라우드 생성

In [31]:
```
from wordcloud import WordCloud
```

In [32]:
```
wc = WordCloud(font_path="C:/Users/User/Downloads/RixYeoljeongdo_Regular/RixYeoljeongd
               background_color="white",
               width=500,
               height=500,
               colormap='PuBu',
               max_font_size=250)
```

In [38]:
```
cloud = wc.generate(" ".join(df.title))
```

In [39]:
```
plt.figure(figsize=(10, 8))
plt.axis('off')
plt.imshow(cloud)
plt.show()
```

2) 형태소 분석기 사용

In [40]:
```
okt = Okt()
```

In [65]:
```
sentences_tag = []
sentences_tag = okt.pos(" ".join(df.title))
noun_list = []
```

In [86]:
```
stopwords=['기자', '아이', '표', '로봇', '넷', '수도', '종목', '주간']
```

In [87]:
```
# tag가 명사인 단어이고, 불용어가 아닌 단어만 noun_adj_list에 넣어준다.
for word, tag in sentences_tag:
    if tag in ['Noun'] and word not in stopwords :
        noun_list.append(word)
```

In [102]:
```
from PIL import Image
mask = np.array(Image.open("C:/Users/User/Downloads/cloud.JPG"))
```

In [113]:
```
wc = WordCloud(font_path="C:/Users/User/Downloads/RixYeoljeongdo_Regular/RixYeoljeongd
               background_color="white",
               width=1500,
               height=1500,
               colormap='seismic',
               mask = mask,
               max_font_size=100)
```

In [114]:
```
cloud = wc.generate(" ".join(noun_list))
```

In [115]:
```
plt.figure(figsize=(10, 8))
plt.axis('off')
plt.imshow(cloud)
plt.show()
```