주제분석 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 주제분석 1-3주차 패키지 문제의 조건 및 힌트는 Python을 기준으로 하지만, R을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 5시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요.

Chapter 1 크롤링

이번주에는 크롤링을 해보겠습니다. 크롤링을 통해 인터넷에 있는 다양한 데이터를 수집할 수 있습니다. 이번주에 크롤링할 사이트는 '매일경제'(https://www.mk.co.kr/news/stock/)사이트의 뉴스기사입니다. 뉴스기사를 크롤링하여 데이터를 수집해봅시다!

첨부된 결과파일과 최대한 비슷하게 크롤링을 통하여 결과물을 만들어 주시면 됩니다. 크롤링을 통해 만든 결과파일(csv형식)과 소스코드(HTML, PDF형식)을 제출해주세요.

문제0. (기본 세팅) 크롬 드라이버를 설치하고, Beautiful Soup, Selenium 라이브러리를 다운받으세요.

문제1. 매일경제(https://www.mk.co.kr/news/stock/) 사이트에 뉴스기사를 크롤링하여 첨부된 결과파일과 최대한 비슷하게 데이터를 만들어 주세요.

- 크롤링 항목 : '날짜', '뉴스기사제목', '뉴스기사본문'
- 데이터 개수 : 최근 뉴스기사 500개

HINT1. 페이지를 뒤로 넘기면서 URL이 어떻게 변화하는지 확인하고, 한 페이지에 몇 개의 뉴스기사가 있는지 확인하세요. 여러 뉴스기사를 보면서 공통된 구조를 가지고 있는지 확인해보세요.

HINT2. 뉴스목록 페이지를 바꾸어 가면서 최근 뉴스기사 500개의 URL과 날짜를 먼저 추출하세요.

HINT3. 앞에서 추출한 뉴스기사 URL을 이용하여 뉴스기사에 접근하여 '뉴스기사 제목'과 '뉴스기사 본문'을 추출하여 데이터 프레임 형태로 저장하세요.

Chapter 2 워드클라우드

앞에서 만든 데이터를 활용하여 워드클라우드를 만들어 봅시다.

문제1. 앞에서 크롤링한 데이터의 '뉴스기사 제목' 데이터를 활용하여 워드클라우드를 만들어보세요.

(색상, 모양, 폰트 등 디자인은 자유롭게 해주시면 됩니다!! *멋진 워드클라우드 기대할게요~-~*)

문제2. 워드클라우드를 통해 파악할 수 있는 인사이트를 간단히 적어주세요.

HINT1. WordCloud 라이브러리 사용하시면 됩니다.

HINT2. 형태소 분석기를 사용하여 명사만 추출하여 워드클라우드를 만들면 더 깔끔한 워드클라우드를 만들 수 있습니다. (Okt, Mecab, Komoran, Kkma 등) (필수는 아닙니다)

HINT3. 의미가 없다고 판단되는 단어들을 불용어로 지정하여 워드클라우드를 만들면 더 깔끔한 워드클라우드를 만들 수 있습니다.) (필수는 아닙니다)

(워드클라우드 예시)