1주차 패키지 – 크롤링 및 텍스트 전처리

주제: 신문기사 자동 분류

우리는 인터넷 신문기사 분류의 번거로움을 덜기 위해 자동으로 신문기사를 분류시켜주는 모델을 구현하고자 합니다. 이번주는 데이터 수집 및 전처리를 진행해보겠습니다.

0. 제공한 메모장의 함수들을 실행하세요.

1. 모델을 훈련시키기 위해 아래 사이트에서 음악(music), 정치(politics), 영화(film), 경제(economy)의 기사를 각각 30개씩 최근 일자부터 크롤링하는 코드를 짜고 직접 크롤링을 해보세요.

크롤링 홈페이지: http://www.koreaherald.com/

제출시 클롤링한 데이터를 첨부하고 일시를 적어주세요.

2. 데이터 확인

2-1. type 함수를 이용하여 데이터의 타입을 확인하세요.

2-2. head 함수를 이용하여 상위 5개의 데이터를 확인하세요.

2-3. tail 함수를 이용하여 하위 5개의 데이터를 확인하세요.

3. 전처리

3-1. 영어가 아닌 텍스트는 모두 삭제하세요.

힌트: 텍스트를 제거하기 위해 df.str.replace() 함수와 정규식을 사용해보세요

3-2. 글자가 3자이하인 단어들을 모두 삭제하세요.

힌트: split함수를 이용하여 문장을 단어로 쪼개준 후 3개 이하인 단어들을 제거한 후 join함수를 이용해 다시 문장으로 만들어주세요.

3-3. 모든 단어들을 소문자로 바꿔주세요.

4. **전처리를 완료한 데이터로 각 주제별 워드 클라우드를 진행하세요.**

   필요하다면 데이터를 class에 따라 나눠주세요.

   전처리와 워드클라우드 이미지는 자유.