다음에 제시된 웹 페이지는 다음뉴스의 랭킹페이지이다.

(http://media.daum.net/ranking/popular/)



이 페이지에서 각 뉴스의 제목과 이 뉴스를 올린 신문사명칭을 스크래핑하여

뉴스제목(newstitle), 신문사명(newspapername) 형식의 데이터프레임을 만들고 CSV 파일로 저장하려고 한다.

이 페이지를 요청하고 앞에서부터 5개의 데이터만 데이터프레임으로 생성하여 리턴하는 것은 R 로 구현하고 R이 생성한 데이터 프레임을 받아와서 파일(daumnews\_schedule.csv)에 저장하는 것은 Java 로 구현한다. 5개 행의 출력 형식은 다음과 같다.

newstitle,newspapername,datetime

뉴스제목,신문사명,YYYY-MM-dd hh:mm

:

또한 가장 중요한 것은 위의 기능을 10분에 한 번씩 수행하도록 스프링의 스케쥴링 컨포넌트와 연계시켜본다. csv 파일에 저장하는 것은 append 모드로 처리해서 이전에 저장된 내용에 추가되게 한다. 즉, SpringSchedulerTest 와 비슷한 역할의 클래스(ScrapingScheduler)를 하나 만들고 10분에 한번씩 수행하도록 설정한다.

제출 소스 : ScrapingScheduler.java