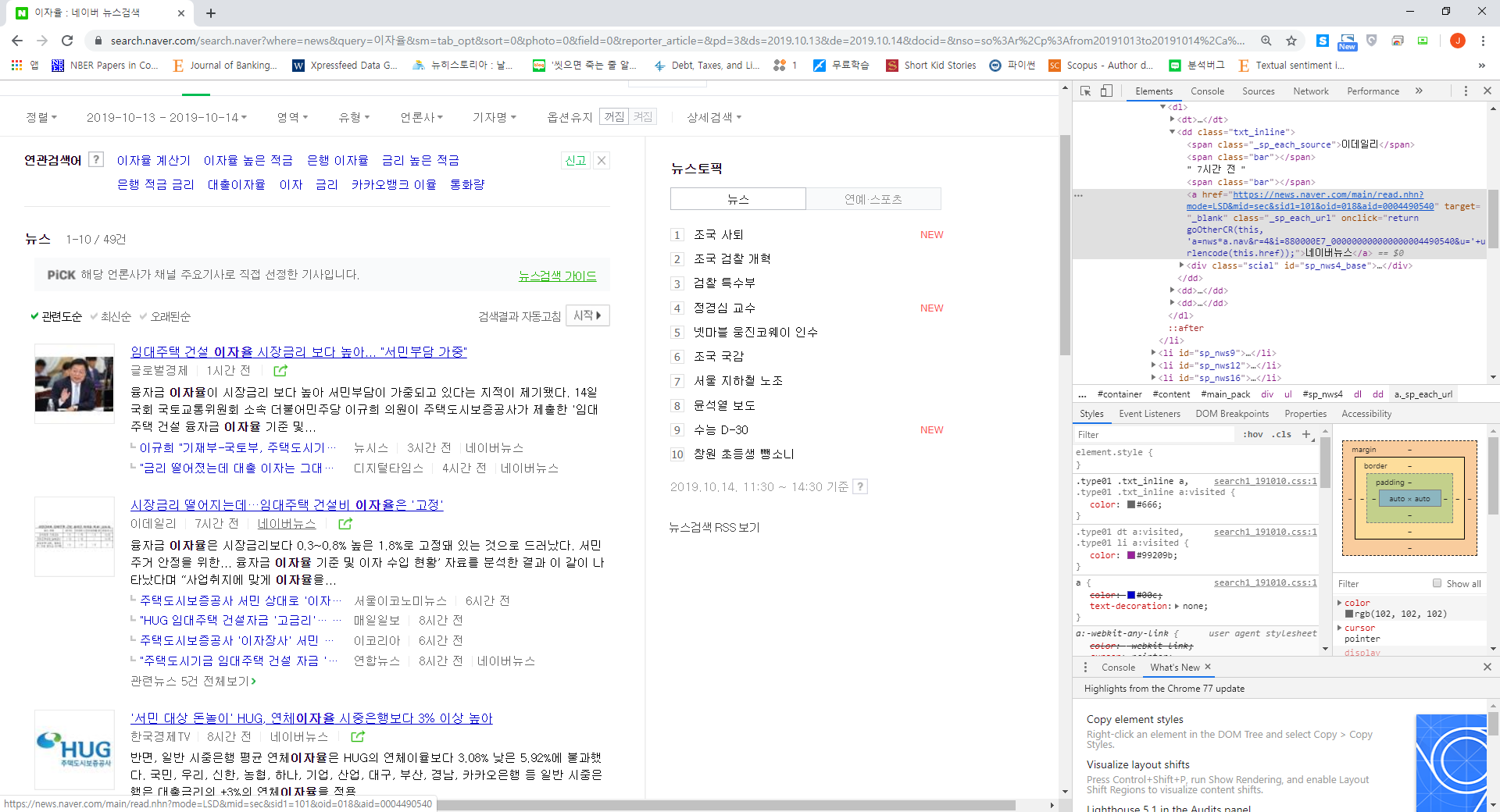
**R – Py 컴퓨팅: Homework 1**

**Due: 2019-10-30, 블랙보드 제출, 파일 형식은 pdf로, 코드는 부록으로 첨부**

**Part 1. 네이버 뉴스 크롤러 만들기 및 검색**

* **네이버 검색창에 검색어를 통해 관련 기사들을 눌러보면, 각기 다른 신문사 사이트로 들어가서 기사를 확인할 수 있다.**
* **예를 들어 조선일보 , 데일리팜 , 한겨례 등등 링크가 연결되있다. 하지만 그들은 완전 별개의 신문사들이고 홈페이지 html구조도 다 다른 구조로 되어 있다.**
* **이에 따라 각 신문사 마다 html tag의 id값이나 class값이 달라서 크롤러를 만들려면 신문사 마다 크롤러를 변경해줘야 한다.**
* **하지만 네이버 포맷으로 바뀐 뉴스에 대해서는 이와 같은 제공 신문사에 상관없이 내용을 크롤링 하는 것이 가능할수 있다.**
* **네이버 뉴스에서 이자율을 검색하고, 검색 기간을 2019년 10월 13일부터 2019년 10월 14일 까지 선택하면 다음과 같은 형태의 url을 얻을 수 있음**
  + **https://search.naver.com/search.naver?where=news&query=이자율&sm=tab\_opt&sort=0&photo=0&field=0&reporter\_article=&pd=3&ds=2019.10.13&de=2019.10.14**

**참고: 확인 방법**

* **한편 네이버에 제공되는 웹페이지 뉴스 링크는 anchor tag (<a>) 이며 href 속성이 "https://news.naver.com/main/read.nhn” 로 시작함을 확인할 수 있음.**

**Q1. 본 질문에서는 뉴스를 키워드로 하여 뉴스 검색 첫번째 페이지에서 네이버 플랫폼에서포맷팅된 뉴스 링크 즉 “https://news.naver.com/main/read.nhn” 로 시작하는 모든 링크를 t수집해 보려고 한다.**

1. **urllib.request.urliib를 활용하여 “금리”를 키워드로 동일한 기간의 뉴스페이지를 열고 req란 이름의 객체로 저장한다.**
   * **참고로 검색어가 한국어로 표시되기 때문에 우선 key\_words=urllib.parse.quote("금리") 하고, url=”**[**https://search.naver.com/search.naver?where=news&query**](https://search.naver.com/search.naver?where=news&query)**=”+ key\_words+”&sm=tab\_opt&sort=0&photo=0&field=0&reporter\_article=&pd=3&ds=2019.10.13&de=2019.10.14” 형태가 되어야 한다.**
2. **이 req객체에 read() 메소드를 사용하여 data 객체를 만든다.**
3. **BeautifulSoup을 사용하여 data 객체를 soup이라는 beautifulsoup 함수를 통해서 생성된 객체로 전환시킨다. 옵션은 html.parser를 이용하면 된다.**
4. **이 soup 객체에 findAll 메소드를 사용하여 tag값이 a 인것을 모두 수집하여 anchor\_set이라는 객체를 만든다.**
5. **각 링크 주소를 저장할 수 있는 news\_link란 리스트를 생성한다.**
6. **이 anchor\_set 에 for loop를 사용하여, 이 anchor\_set에서 하나의 요소들에 대해 만약 href 속성이 'https://news.naver.com/main/read.nhn'으로 시작한다면 news\_link란 객체에 이 링크 주소를 받아서 이 링크의 주소를 받아서 append 메소드를 통해 주소들을 저장한다.**

**참고로 사용되는 Library는 다음과 같다.**

**import urllib.request**

**import urllib.parse**

**from bs4 import BeautifulSoup**

**import re**

**다음 참고 웹페이지를 미리 읽어 구체적인 내용에 우선적으로 대해서 파악한다.**

[**https://m.blog.naver.com/cbs3455/221242251959**](https://m.blog.naver.com/cbs3455/221242251959)

[**https://bumcrush.tistory.com/155?category=697185**](https://bumcrush.tistory.com/155?category=697185)

[**https://rednooby.tistory.com/5**](https://rednooby.tistory.com/5)

**Q1에서 정의된 soup 객체에서 우리가 뉴스 검색과 연관된 총 기사 수를 확인하는 방법은 다음과 같은 방법을 사용하면 같다.**

**count\_tag=soup.find("div",{"class","title\_desc all\_my"})**

**count\_text=count\_tag.find("span").get\_text().split()**

**total\_num=count\_text[-1][0:-1].replace(",","")**

* **위와 같은 코드를 사용하면 total\_num에는 뉴스 검색에 나온 총 기사수가 저장이 되게 된다.**
* **위와 같은 총 기사수와 네이버의 검색 페이지가 10개를 한 페이지 단위로 뉴스 검색 자료를 제공한다는 사실을 활용한다면 우리는 전체 뉴스 중 네이버 플랫폼에서 포맷팅 된 뉴스들에 대한 링크를 모두 찾아낼 수 있다.**
* **이러한 크롤링은 다음과 같은 네이버뉴스 URL 특성에 의존하여 가능하게 된다.** 
  + **연관뉴스 11-20까지 검색 링크:**
    - **https://search.naver.com/search.naver?where=news&query=이자율&sm=tab\_opt&sort=0&photo=0&field=0&reporter\_article=&pd=3&ds=2019.10.13&de=2019.10.14&docid=&nso=so:r,p:from20191013to20191014,a:all&mynews=0&cluster\_rank=26&start=11&refresh\_start=0**
  + **연관뉴스 21-30까지 검색 링크:**
    - **https://search.naver.com/search.naver?where=news&query=이자율&sm=tab\_opt&sort=0&photo=0&field=0&reporter\_article=&pd=3&ds=2019.10.13&de=2019.10.14&docid=&nso=so:r,p:from20191013to20191014,a:all&mynews=0&cluster\_rank=26&start=21&refresh\_start=0**

**Q2. 위에서 제시된 total\_num을 이용해서 “이자율” 검색 관련해서 전체 기사에 대해 네이버 플랫폼에서포맷팅된 뉴스 링크 즉 “https://news.naver.com/main/read.nhn” 로 시작하는 모든 링크를 찾아서 new\_link 리스트에 저장하다.**

* Hint: 다음과 같은 for loop로 시작하는 파이썬 구문을 사용할 수 있다.

for val in range(int(total\_num)//10):

start\_val=str(val\*10)+"1"

* 파이썬의 set과 list 성격을 파악해서 이미 획득한 리스트에 대해서 중복을 제거 할 수도 있을까?

**Q3. 위에서 구성한 news\_link 리스트를 이용해서 뉴스 제목과 뉴스 본문 내용을 각각 title\_list와 text\_list로 저장한다.**

* For 구문을 news\_list 요소들에 대해서 구성한다.
* Q1에서와 같이 이 각 news\_list에서 제공되는 링크를 방문하고 데이터를 얻고 parsing 한다.
* 뉴스 제목의 태그는 h3이고 속성 id 값은 articleTitle 이다
* 뉴스 본문의 태그는 div이고 articleBodyContents이다.
* find("h3",{"id": "articleTitle"}).get\_text()과 같은 메소드를 통해 특정 속성을 가진 특정 태그를 찾을 수 있고 그 내용을 확보할 수 있다.

**Q4. title\_list에 저장된 뉴스 제목에 대해 “금리”로 시작하고, 중간에 모든 문자가 여러 번 나와도 상관없으며, “인하”로 끝나는 문자열이 있는지 정규표현식을 활용하여 검색하고, 이 검색이 맞으면 제목을 출력하는 프로그램을 구현하라.**

**Part 2 보스턴 주택가격 데이터 분석하기: boston\_csv.csv 파일**

* **보스턴 주택 가격 데이터의 변수명은 다음과 같다.**

|  |  |
| --- | --- |
| CRIM | 자치시(town) 별 1인당 범죄율 |
| ZN | 25,000 평방피트를 초과하는 거주지역의 비율 |
| INDUS | 비소매상업지역이 점유하고 있는 토지의 비율 |
| CHAS | 찰스강에 대한 더미변수  (강의 경계에 위치한 경우는 1, 아니면 0) |
| NOX | 10ppm 당 농축 일산화질소 |
| RM | 주택 1가구당 평균 방의 개수 |
| AGE | 1940년 이전에 건축된 소유주택의 비율 |
| DIS | 5개의 보스턴 직업센터까지의 접근성 지수 |
| RAD | 방사형 도로까지의 접근성 지수 |
| TAX | 10,000 달러 당 재산세율 |
| PTRATIO | 자치시(town)별 학생/교사 비율 |
| B | 1000(Bk-0.63)^2,   여기서 Bk는 자치시별 흑인의 비율을 말함. |
| LSTAT | 모집단의 하위계층의 비율(%) |
| MEDV | 본인 소유의 주택가격(중앙값) (단위: $1,000) |
| CAT.MEDV | MEDV 가 $30,000 을 넘는지에 대한 변수  (넘는 경우 1, 아닌 경우 0) |

* **보스턴 주택 가격 데이터에서 결측치는 na와 NaN으로 표시되어 있다.**
* **다음과 같은 라이브러리를 불러와야 프로그램이 구현될 것이다.**

**from sklearn.linear\_model import LinearRegression**

**from sklearn.metrics import mean\_squared\_error**

**import numpy as np**

**import matplotlib.pyplot as plt**

**import pandas as pd**

**import seaborn as sns**

**Q5 다음과 같은 데이터 전처리를 시행하라.**

1. **제공된 boston\_csv.csv 파일을 사용하여 pandas 데이터 프레임 객체를 만든다, 결측치 코드인 na와 NaN이 모두 실제 결측치로 되도록 한다.**
2. **이와 같은 결측치가 있는 관측치를 모두 제거한다.**

**Q6 다음과 같은 요약 통계를 구하라.**

1. **describe 메소드를 적용해서 각 변수병 요약 통계를 구한다.**
2. **상관관계를 구한후 seaborn 라이브러리의 heatmap 을 구현한다.**

**Q7 다음과 같은 단순회귀분석 모형을 Training Set과 Test Set을 통해 구현하라.**

* 모집단의 하위계층의 비율(LSTAT)이 독립변수
* 본인 소유의 주택가격(중앙값)인 MEDV가 종속변수
* Training set이 표본의 75%를 차지한다.
* Training Set에대해서는 회귀 분석 추정 계수 및 R2 값 mean squared error 값을 보고한다.
* Training Set에대해서는 회귀 분석 추정 계수 값을 바탕으로 Test Set에서 예측한 후 mean squared error 값을 보고한다.

**Q8 다음과 같은 다중회귀분석 모형을 Training Set과 Test Set을 통해 구현하라.**

* 모집단의 하위계층의 비율(LSTAT)과 10,000 달러 당 재산세율(TAX)가 독립변수
* 본인 소유의 주택가격(중앙값)인 MEDV가 종속변수
* Training set이 표본의 75%를 차지한다.
* Training Set에대해서는 회귀 분석 추정 계수 및 R2 값 mean squared error 값을 보고한다.
* Training Set에대해서는 회귀 분석 추정 계수 값을 바탕으로 Test Set에서 예측한 후 mean squared error 값을 보고한다.