

빅데이터분석및실험 중간시험 문제

<2020. 10. 20 14:00 ~ 15:50>

<open books/internet, close mind!!!>

<주의: 교수 PC에서 여러분들의 화면을 monitoring 하고 있습니다. 메신저/메일 프로그램이 작동되면 즉시 해당 PC의 전원을 끄겠습니다>

제출물

- ULMS 의 8주차 <과제물>에 등록
- 제출물: python(ipython notebook) 프로그램, 결과물(생성파일 등)
- 가급적 압축파일(학번_이름) 하나로 제출

1. 공정거래위원회 가맹사업거래

(<https://franchise.ftc.go.kr/user/extra/main/70/firHope/listBrand/jsp/LayOutPage.do>)

브랜드별 비교에서 {2019년 외식, 커피, 가맹점 현황 정보}를 검색/추출하여 csv 파일로 저장하는 프로그램을 작성하시오. 단, {가맹점수, 신규개점, 계약종료, 계약해지, 명의변경, 가맹점 평균매출액, 가맹점 면적당 평균매출액}가 모두 0인 경우는 제외하고 저장하시오. (20)

생활커피	(주)데일리리비어	0	0	0	0	0	0 (0.00)	0 (0.00)
반할커피	반할커피	0	0	0	0	0	0 (0.00)	0 (0.00)
두이모	두이모	0	0	0	0	0	0 (0.00)	0 (0.00)

[그림 1. 제외할 데이터의 예시]

2. 투썸플레이스 울산 매장을 추출하여 아래와 같이 출력하라(구, 매장명, 전화번호 순). (20)

- selenium 사용
- 구/군 전체 보기가 아닌, 각 구(ex. 남구)를 순서대로 선택하여, 해당 구에 존재하는 매장명과 전화번호를 모두 추출하는 방식 사용.

```
[['남구', '투썸울산태화강역', '052-267-2321'],  
 ['남구', '투썸울산무거', '052-277-5055'],  
 ['남구', '투썸울산달동대영', '052-271-2800'],
```

3. 네이버 뉴스 페이지를 활용하여 프로그램 작성하라. 단, 셀레니움을 사용하지 마시오.

(https://news.naver.com/)

(1) 언론사별 많이 본 뉴스를 모두 추출하라.(제목, 신문사)

- 더보기기를 통해 모든 언론사에 대한 정보를 추출할 필요 X



result

[[['매일경제', '오뚜기 3세' 합연지 "유튜브로 친구 30만명 생겼어요...'],
 ['중앙일보', '아들 친권 잃은 고유정, 현남편과도 이혼...'위자료 3000 ...'],
 ['SBS', '공항 화장실에 조산아 카타르, 여성 승객 강제 자궁 검 ...'],
 ['월간 산', '[10월 넷째 주 추천산행지] 강천산, 기암괴석에 계곡 단 ...'],
 ['경향신문', '추미애, 윤석열에 "선 넘은 발언, 부적절" 비판']]]

[그림 1. 2020년 10월 26일, 오후 3시 기준으로 추출한 예시]

(2) 3-(1)에서 추출했던 언론사별 많이 본 뉴스의 원문을 찾아서 제목과 URL을 새로 추출하여 신문사, 제목, url 순으로 **csv 파일**로 저장하시오. 추출은 아래 설명을 참고하여 순서대로 구현하시오.

① 네이버 언론사 뉴스(<https://news.naver.com/main/officeList.nhn>)에서 '페이지 소스 보기'를 통해 데이터를 추출하여 언론사 이름(key)에 대한 oid(value)를 가지는 dictionary를 직접 생성하시오.

언론사 뉴스

 신문 게재된 기사 정보 표시합니다.

종합	경향신문 서울신문 한겨레	국민일보 세계일보 한국일보	동아일보 조선일보	문화일보 중앙일보
방송/통신	뉴스1 재능A MBC TV조선	뉴스S 한국경제TV MBN YTN	연합뉴스 JTBC SBS	연합뉴스TV KBS SBS CNBC
경제	매일경제 이데일리 한국경제	머니투데이 조선비즈 해럴드경제	서울경제 조세일보	아시아경제 파이낸셜뉴스
인터넷	노컷뉴스 미디어오늘	더팩트 아이뉴스24	데일리안 오마이뉴스	머니S 프레시안
IT	디지털이데일리 ZDNet Korea	디지털타임스	블로터	전자신문
매거진	레이디경향 신동아 주간동아 한경비즈니스	매경이코노미 월간 산 주간조선	시사IN 이코노미스트 중앙SUNDAY	시사저널 주간경향 한겨레21
전문지	기자협회보 일다 코메디닷컴	뉴스타파 장세상 월스조선	동아사이언스 코리아중앙데일리	여성신문 코리아재팬드
지역	강원일보	매일신문	부산일보	
포토	신화사 연합뉴스	AP연합뉴스	EPA연합뉴스	

생성한 dictionary 예시

```
media_dictionary
{
  '경향신문': '032',
  '국민일보': '005',
  '동아일보': '020',
  '문화일보': '021',
  '서울신문': '081',
  '세계일보': '022',
  '조선일보': '023',
  '중앙일보': '025',
  '한겨레': '028',
  '한국일보': '469',
  '뉴스1': '421',
  '뉴스S': '003',
}
```

[그림 2. 언론사 dictionary 생성]

② 네이버 뉴스 페이지의 url의 parameter로 oid(dictionary 사용), date(오늘 날짜), page를 사용하여 3-1에서 추출한 정보의 original 제목과 url을 추출하시오. (7주차 과제 참고)

- 3-1에서 추출했던 제목을 이용하여 같은 제목의 기사를 찾는 방식 사용
- 문자열의 처음 10음절이 같다면 같은 기사라고 판단해도 됨(아래와 같은 경우때문)

3-(1)에서 추출했던 제목	`오투기 3세` 함연지 "유튜브로 친구 30만명 생겼어요... .."
original 제목	`오투기 3세` 함연지 "유튜브로 친구 30만명 생겼어요...굉장한 힐링"



[그림 3. 2020-10-26, 오후 3시 50분을 기준으로 추출한 예시]

A	B	C	D	E	F	G	H	I	J	K	L
연론사	제목	원문 링크									
0 매일경제	'오투기 3세' 함연지 "유튜브로 친구 30만명 생겼어요...굉장한 힐링"	https://news.naver.com/main/read.nhn?mode=LPOD&mid=sec&oid=0098&aid=0004681893									
1 중앙일보	아들 친권 잃은 고유정, 현남편과도 이혼..."위자료 3000만원"	https://news.naver.com/main/read.nhn?mode=LPOD&mid=sec&oid=025&aid=0003046675									
2 SBS	'공항 화장실에 조산아' 카타르, 여성 승객 강제 자궁 검사 ...	https://news.naver.com/main/read.nhn?mode=LPOD&mid=sec&oid=055&aid=0000850331									
3 월간 산	[10월 넷째 주 추천산행지] 강천산, 기암괴석에 계곡·단풍 절경	https://news.naver.com/main/read.nhn?mode=LPOD&mid=sec&oid=094&aid=0000008536									
4 경향신문	추미애, 윤석열에 "선 넘은 발언, 부적절" 비판	https://news.naver.com/main/read.nhn?mode=LPOD&mid=sec&oid=032&aid=0003039575									

[그림 4. 정답 파일의 예시]

4. 공공데이터(data.go.kr)에서 <국민권익위원회_민원빅데이터_분석정보> (30)

(1) 일반 인증키를 발급받으세요.

(발급받지 못할 경우 아래 것(옥철영교수) 사용

KoQvU3n5iWPhwcIIlNlXksch6O5FI9%2Bpo7v5X4I2OTuEx5WmdZwPEdGh9zynJQMw0H

H4NM0KC%2BcmrZv4z%2Br6QQ%3D%3D

(2) 참고문서 API기술문서 - 민원분석정보조회 (ver 1.7).docx 참조 혹은

<https://www.data.go.kr/tcs/dss/selectApiDataDetailView.do?publicDataPk=15040459>

에서 목록 “TOP 키워드 정보” 참조

(3) 2019.01.01. ~ 2019.12.31. 사이의 주별(weekly) 국가인권위원회의 일반민원, 고충민원을 csv 파일로 저장