

13. 텍스트 마이닝(Text Mining)

1강. 웹 크롤링으로 기초 데이터 수집하기

학습내용

- 웹 크롤링으로 기초 데이터 수집하기
- 나무위키 최근 변경 페이지 키워드 분석

학습목표

- 웹 크롤링으로 기초 데이터를 수집하고 분석할 수 있다.
- 나무위키에서 최근 변경이 일어난 페이지들의 키워드를 이용하여 분석할 수 있다.

1. 웹 크롤링으로 기초 데이터 수집하기

1) 텍스트 마이닝

- 비정형 데이터, 텍스트 데이터로부터 유의미한 정보를 추출하는 데이터 분석
- 비정형 데이터는 이미지 데이터나 음성 데이터와 같은 정해진 형태가 없고 연산이 불가능한 데이터를 의미

2) 텍스트 마이닝의 예

- 구글, 네이버 번역기, 챗봇
- 텍스트 임베딩(자연어 처리)
- 검색, 추천시스템
- 자연어 처리는 텍스트가 기계적으로 어떤 의미를 가지고 있는지를 추출하는 것

3) 웹 크롤링

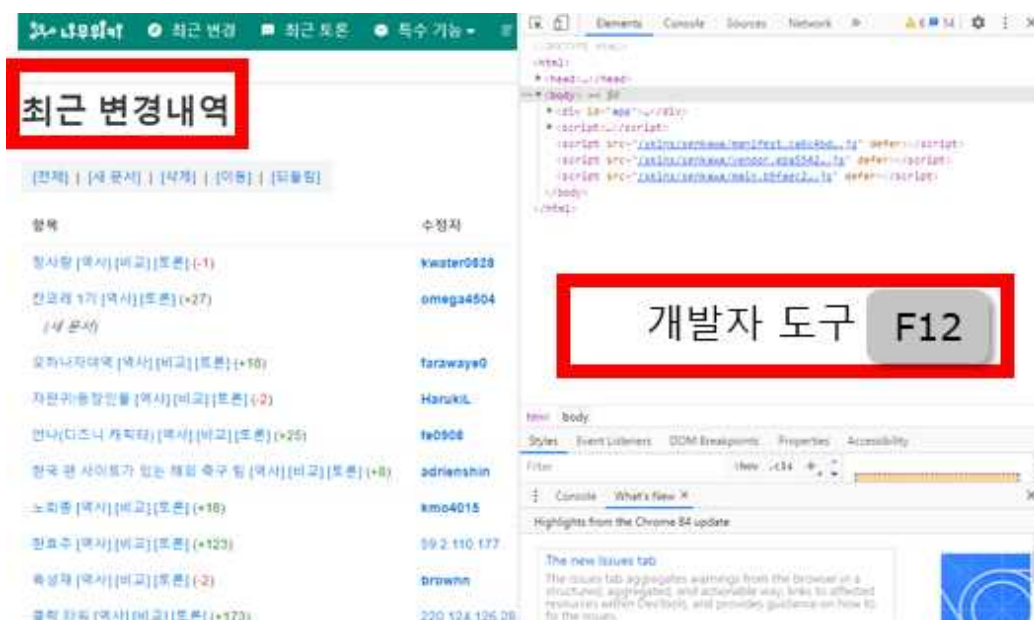
- 웹 스크래핑(Web Scraping)이라고도 하며 인터넷에 있는 웹 페이지를 방문해서 페이지의 자료를 자동으로 수집하는 작업
- 텍스트 데이터를 웹 크롤링으로 수집한 다음 데이터 내에서 등장한 키워드의 출현 빈도를 분석할 수 있음

가. 웹 크롤링으로 기초 데이터 수집하기

- 크롤링을 위한 첫 번째 단계는 인터넷 익스플로러, 크롬 등의 웹 브라우저를 실행하여 크롤링의 대상이 될 페이지 구조를 살펴보는 것



- [도구 더보기] - [개발자 도구]를 순서대로 클릭하여 개발자 도구를 실행



나. 웹 크롤링 라이브러리

- BeautifulSoup과 requests 라는 라이브러리로 웹 크롤러를 만들 수 있음
- requests는 특정 URL로부터 HTML 문서를 가져오는 작업 수행
- BeautifulSoup 모듈은 HTML 문서에서 데이터를 추출하는 작업 수행

4) 페이지의 URL 정보 추출하기

- requests . get () 함수로 URL의 HTML 문서 가져옴

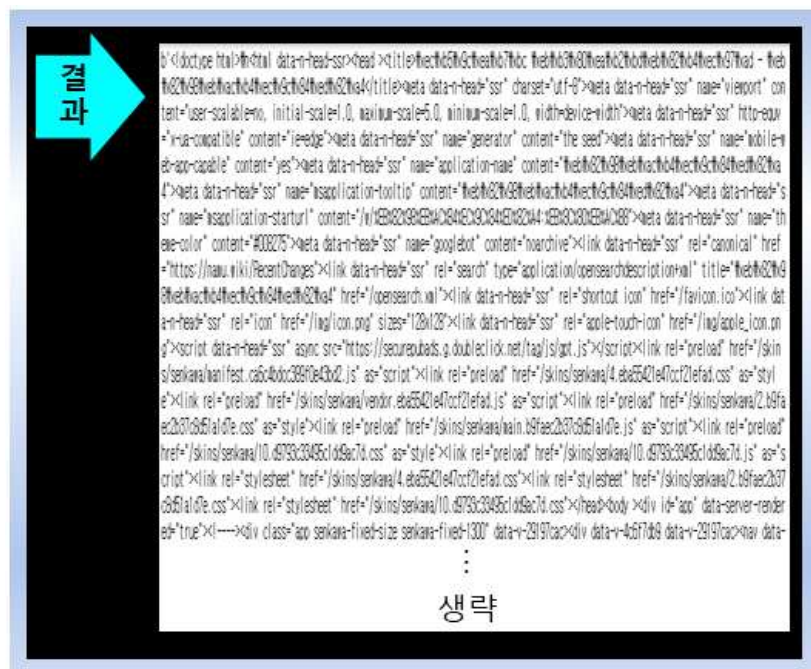
```
import requests
from bs4 import BeautifulSoup
import re

# 크롤링할 사이트 주소를 정의합니다.
source_url = "https://namu.wiki/RecentChanges"

# 사이트의 html 구조에 기반하여 크롤링을 수행합니다.
req = requests.get(source_url)
html = req.content

html
```

- html 코드를 모두 가져옴



- 파싱(Parsing)은 어떤 페이지(문서, html)에서 내가 원하는 데이터를 특정 패턴이나 순서로 추출하여 정보로 가공하는 것

```
soup = BeautifulSoup(html, 'lxml')
contents_table = soup.find(name="table",
                             attrs={"class": "table-hover"})
table_body = contents_table.find(name="tbody")
table_rows = table_body.find_all(name="tr")
```

- HTML 구조에 기반하여 table → tbody → tr → td → a 태그 순으로 HTML 계층 구조를 좁혀나가는 과정
- 목표 태그에 도달했을 때 get(href)함수로 태그의 속성 정보를 추출
- get()함수는 해당 태그가 가지고 있는 특정한 속성을 추출

5) 텍스트 정보 수집하기

- 추출한 웹 페이지들의 URL을 방문하여 HTML 구조를 개발자 도구로 살펴 본 것

The screenshot shows a web browser with a page titled '사설' (Private). The page content includes a search bar, a title '사설', a date '최근 수정 시각: 2020-08-18 23:40:03', and two sections: '1. 개요' and '2. 생애'. The developer tools on the right show the HTML structure, including the head and body tags, and the table structure used for parsing.

- get() 함수 대신 text() 함수를 사용하여 태그의 텍스트 정보만 추출

```
req = requests.get(page_urls[0])
html = req.content
soup = BeautifulSoup(html, 'lxml')
contents_table = soup.find(name="article")
title = contents_table.find_all('h1')[0]
category = contents_table.find_all('ul')[0]
content_paragraphs =
contents_table.find_all(name="div",
attrs={"class":"wiki-paragraph"})
content_corpus_list = []

print(title.text)
print(category.text)
print(content_corpus)
```

결과

사설

한나라의 관료동오의 인물베트남의 군주추존된 왕137년 출생226년 사망

사설 가문의 역대 군주 베트남 종 제대 - 초대 사설 - 2대 사설 陳朝 - 胡朝 전 왕조 - 대무 황제 [굴치기 · 겹기] 陳朝전 왕조제1대
 단종제2대 성종제3대 인종제4대 영종제5대 명종제6대 현종제7대 유종제8대 전폐제제9대 예종종무제10대 예종종무제 陳朝 전 왕조 胡朝 호 왕조(대무)제11대 후
 폐제제12대 순종제13대 소제제14대 국조제20대 호한왕 陳朝 후 전 왕조제14대 간정제제15대 중왕제제16대 진고추존 영무왕 · 목조 · 영조 · 원조 ·
 단조 · 충무왕 · 송도왕 · 황대백 비정통 진도애 · 진익직둘러보기 陳朝전 왕조제11대 후폐제제12대 순종제13대 소제제14대 국조제20대 호한왕 陳朝
 7대 유종제8대 전폐제제9대 예종종무제10대 예종종무제 陳朝 전 왕조 胡朝 호 왕조(대무)제11대 후폐제제12대 순종제13대 소제제14대 국조제20대 호한왕 陳朝
 후 전 왕조제14대 간정제제15대 중왕제제16대 진고추존 영무왕 · 목조 · 영조 · 원조 · 단조 · 충무왕 · 송도왕 · 황대백 비정통 진도애 ·
 진익직 시호 선강가음령무대왕(恭敬嘉應肅武大王) 왕호 사왕(土王, 시보엡/St Vong) 성씨 새(士) 휘 설(奘) 베트남식 이름 시 니엡(St Nh
 iệp) 생몰 년도 137년 - 226년 재위 년도 187년 - 226년1. 개요2. 생애3. 사후4. 평가5. 미디어 믹스5.1. 삼국지 시리즈5.2. 삼국전투기5.3. 토
 말 워: 삼국6. 둘러보기후한 말, 삼국시대 오나라의 인물. 후한 말부터 삼국시대에 교주를 다스린 인물. 자는 위연(威彦), 뒤에 베트남의 전 왕
 조(陳朝)는 사설을 선강가음령무대왕(恭敬嘉應肅武大王)으로 추존하였다. 연의에는 언급조차 없고, 정사에만 등장한다. 묘하게도 반삼국지에는 등
 장한다. 교주(交州) 창오군(蒼梧郡) 광신현(廣信縣) 사람. 교주라는 중국 최남단 지방(현재의 광둥 성, 광시 치구, 베트남 북부하고 중부) 주자사
 였다. 사실 한나라가 베트남을 제대로 지배한 시기가 거의 없다시피하다. 교지는 북파장군 마윈에 쫓겨(흥박(11), 쯔니)의 변란을 진압한 이후
 허해 속해제 되었지만 한성 베트남은 패한적이었고, 결국 무능한 인물 하 도명애 의존하는 황태로 변방을 놓려왔는데 사설 역시 그 중의 한 명에

2. 나무위키 최근 변경 페이지 키워드 분석하기

1) 나무위키의 최근 데이터 크롤링하기

- 크롤링한 데이터를 데이터 프레임으로 만들기 위해 준비

```
columns = ['title', 'category', 'content_text']
df = pd.DataFrame(columns=columns)

# 각 페이지별 '제목', '카테고리', '본문' 정보를 데이터 프레임으로 만듭니다.
for page_url in page_urls:

    # 사이트의 html 구조에 기반하여 크롤링을 수행합니다.
    req = requests.get(page_url)
    html = req.content
    soup = BeautifulSoup(html, 'lxml')
    :
df.head(5)
```

- 각 페이지별 제목, 카테고리, 본문 정보를 데이터 프레임으로 생성

결과

	title	category	content_text
0	루도 유세이	유희왕 5D's 등장인물유희왕:주인공공작자 캐릭터	유희왕 시리즈의 주인공 [글치기 - 접기] 토에이판, DMGXSD's 루토우 유우기...
1	RPG 인전장	RPG 인	상위 문서: RPG 인. 개요2. 일판2.1 수련장2.2 야산2.2.1...
2	블루터가	블루터가	1. 개요2. 가계도2.1. 1남 신격포 (롯데)2.2. 2남 신필포2.3. 1녀...
3	나츠(유희왕)	나츠(유희왕)	닉네임 나츠이름 김가현성년출일1993년 5월 20일 (27세)거주지출산장역시전공화...
4	MAP OF THE SOUL : 7 - THE JOURNEY -	방탄소년단/음반2020년 음반일본의 음반	전승가게이홈RM지민북장국Discography [한국음반]SKOOL Series ...

2) 키워드 정보 추출

- 수집한 텍스트 데이터에서 키워드 정보를 추출하기 위해 텍스트 전처리 작업이 필요
- 텍스트 전처리는 특수문자나 외국어를 제거하는 등의 과정을 포함
- 키워드 정보 추출은 좁은 의미에서
- 명사 혹은 형태소 단위의 문자열을 추출하는 것

가. 키워드 정보 추출에서 언어와 상황마다 다른 경우

- 스팸메일을 분류하는 텍스트 마이닝의 경우 특수문자나 외국어가 분석의 중요한 힌트가 될 수 있기 때문에 이를 제거하지 않는 편임
- 키워드 분석처럼 ‘단어’를 추출하는 것이 목적이라면 특정 언어의 글자만을 추출하기도 함

나. 정규표현식

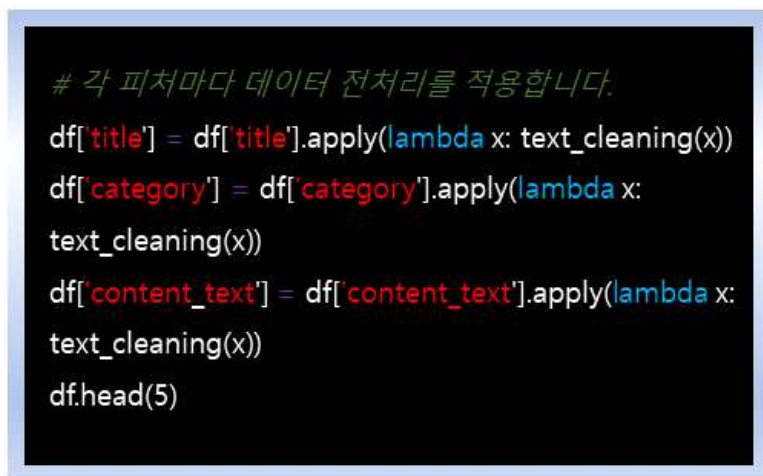
- 정규표현식이란 특정 규칙을 가진 문자열의 집합을 표현하는 형식
- 파이썬에서는 ‘re’라는 모듈을 통해 정규표현식을 사용
- ‘[[^] ^ㄱ - ^ㅣ 가 - ^ㅎ]+’ 이라는 코드로 한글에 대한 정규표현식을 정의하면 대상이 되는 텍스트 데이터에서 한글만 추출할 수 있음

3) 텍스트 데이터 전처리하기

```
# 텍스트 정제 함수 : 한글 이외의 문자는 전부 제거합니다.
def text_cleaning(text):
    hangul = re.compile("[^ ㄱ - ㅣ 가 - ㅎ]+")
    # 한글의 정규표현식을 나타냅니다.
    result = hangul.sub("", text)
    return result
print(text_cleaning(df['content_text'][0]))
```



4) 모든 데이터에 전처리하기



- 모든 데이터에 전처리를 적용하기 위해서는 apply() 함수 사용

	title	category	content_text
0	루카 생크	인물성독일의 축구선수프로파 영역	헤르타 시즌 선수명단 스포츠 칼리크 라치 슈파르트 다리아...
1	장국영인수신 단	방탄소년단 멤버장국영인수신 단인물아이들 단서아이들 단서유종고인아이들 보컬 단	가일 후 일기 지니어 문장 가능성은 문서입니다전송가게이름지은박경숙 한국...
2	주인 아이들	의 프로그램	대한민국의 주요 예술 프로그램은 문화가 많이 볼 수...
3	김민	종영이전	가오 유종성제로 김 이름으로 전을 사용 중인 사람들의 이름지은 박경숙...
4	문득있는 그 내	드라마로로 드라마인 드라마	문득 드라마를 많이 보기 위해 재방송을 하기로 했다고...

평가하기

1. 인터넷에 있는 웹 페이지를 방문해서 페이지의 자료를 자동으로 수집하는 작업은?

- ① 찾기
- ② 리스트 URL
- ③ 웹 크롤링
- ④ 웹 판다스

- 정답 : ③번

해설 : 인터넷에 있는 웹 페이지를 방문해서 페이지의 자료를 자동으로 수집하는 작업을 의미합니다.

2. 웹 크롤링 라이브러리 중 특정 URL로부터 HTML 문서를 가져오는 작업을 수행하는 것은?

- ① BeautifulSoup
- ② Requests
- ③ Numpy
- ④ Pandas

- 정답 : ②번

해설 : Requests는 특정 URL로부터 HTML 문서를 가져오는 작업을 수행합니다.

학습정리

1. 웹 크롤링으로 기초 데이터 수집하기

- 텍스트 마이닝
- 웹 크롤링의 개념, 라이브러리

2. 나무위키 최근 변경 페이지 키워드 분석하기

- 웹 데이터 가져오기
- 키워드 정보 추출하기