웹 크롤링

강사소개

Lecturer

유 연 휘

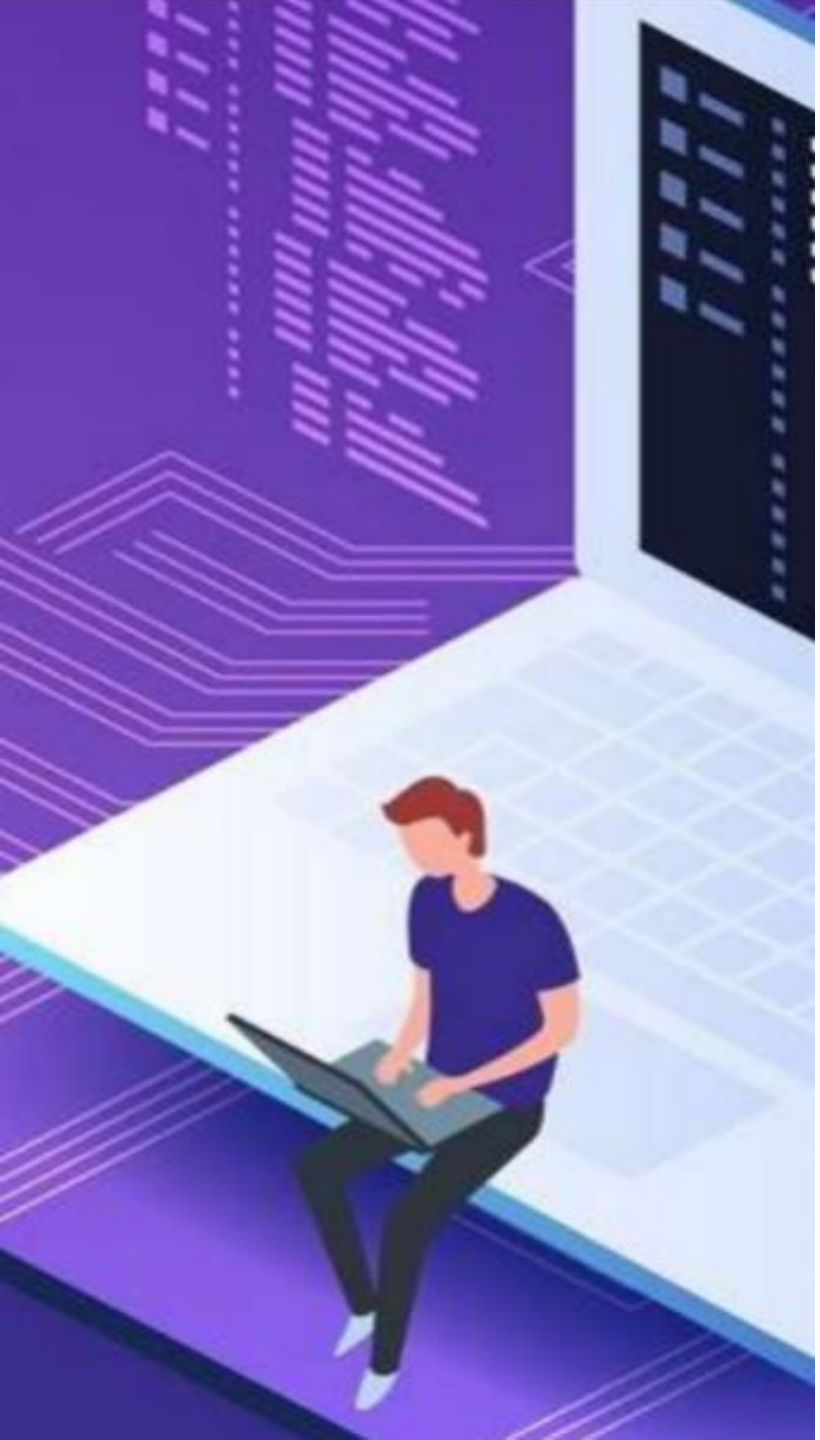
약력	▪ (現) 팀 에뚜 공동대표 ▪ (前) 복자여자고등학교 테크메니저 ▪ 한국기술교육대학교 공학 석사 (연구 분야 : 추천시스템, 생성형 AI)
강의 분야	▪ 생성형 AI, 데이터분석, Python 프레임워크 등
강의이력	▪ [NIA] 우크라이나 학생 대상 Python 언어 교육 ▪ [삼성전자] 파이썬 강의 ▪ [삼성전자 Citizen Developer] Pandas 데이터 전처리 강의 ▪ [한국장애인고용공단] 파이썬 및 신기술 강의, OA 강의 ▪ [천안 복자여자고등학교] 파이썬 및 인공지능입문 강의 ▪ [천안 오성고등학교] 정보교과 강사 ▪ 그 외 인공지능 기초 및 생성형 AI 특강 다수
프로젝트 이력	▪ 문경새재 게임 개발 (팀 에뚜) ▪ 메타데이터 추출과 추천시스템 (한국연구재단) ▪ 모빌리티 센싱, 자율주행 알고리즘 구축 (RIS)

WEB

인터넷에 연결된 컴퓨터를 통해 사람들이 정보를 공유할 수 있는 전세계적인 정보 공간을 의미

CRAWLER

유용한 정보를 찾아 특정 저장소로 수집해 오는 프로그램

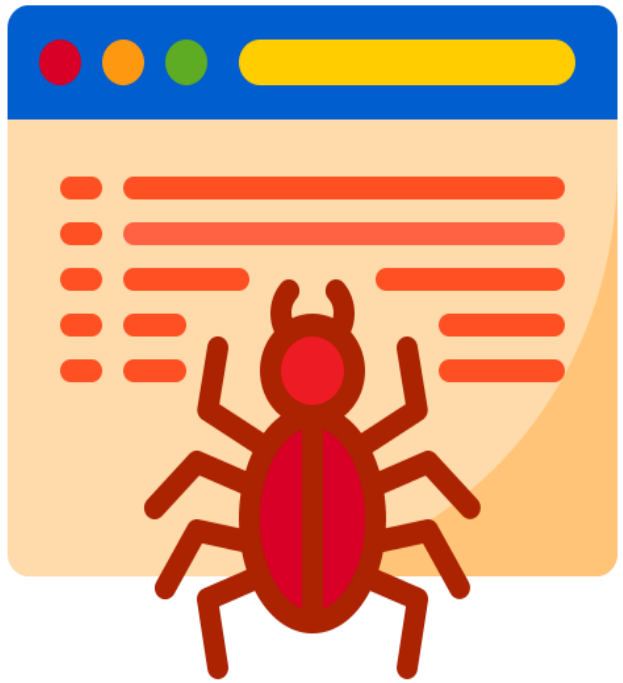


1. 웹 상의 다양한 정보를 자동으로 검색하고 색인하기 위해 검색 엔진을 운영하는 사이트에서 사용하는 소프트웨어 [한국정보통신기술협회]

2. 웹 페이지의 하이퍼링크를 순회하면서 웹 페이지를 다운로드 하는 작업

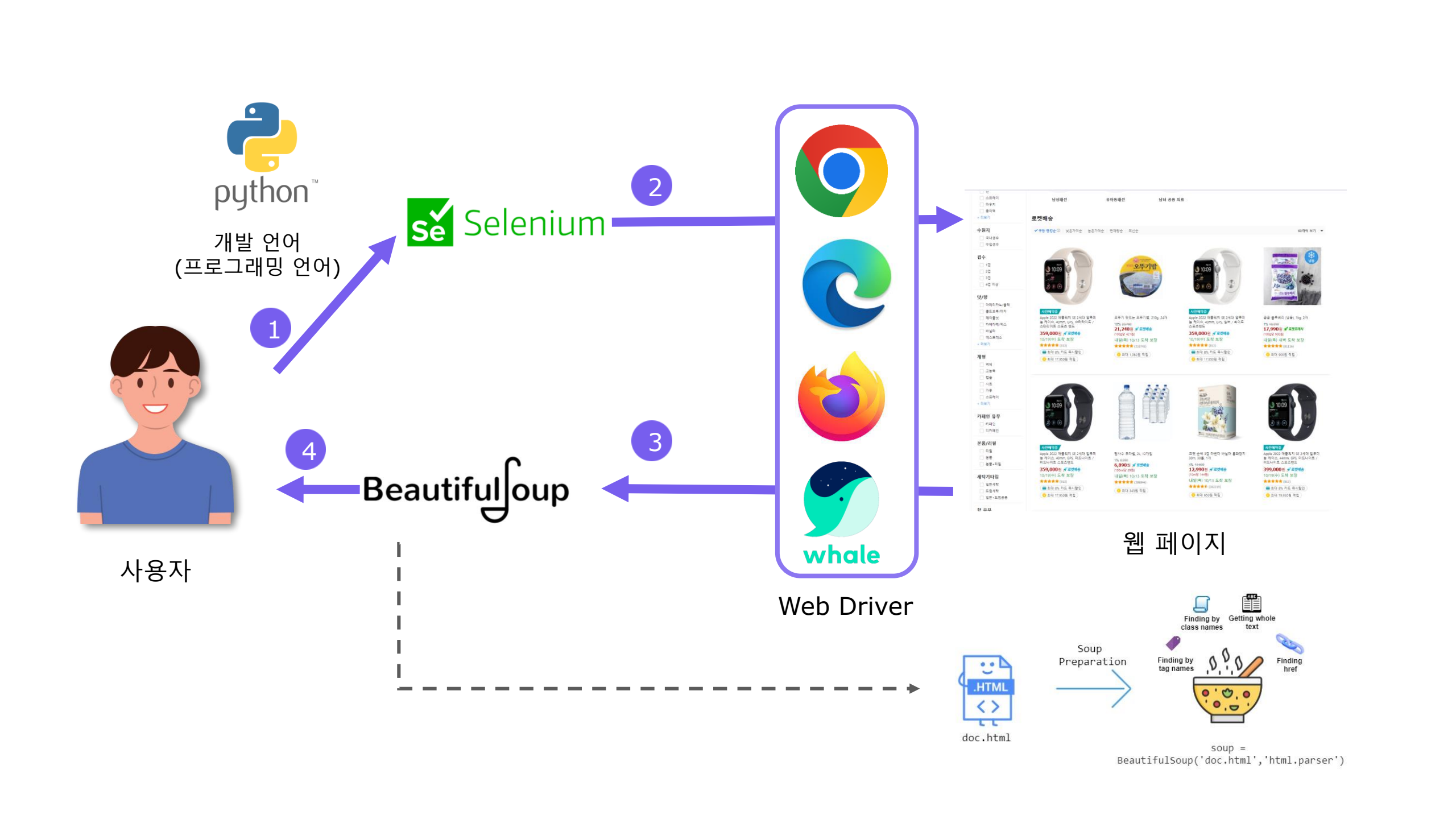
Web Crawler

웹 페이지의 하이퍼링크를 순회하면서 웹 페이지를 다운로드 하는 작업



Scraping

다운로드한 웹 페이지에서 필요한 정보를 추출하는 작업


python
개발 언어
(프로그래밍 언어)
Selenium
1
2
3
4
BeautifulSoup
사용자
whale
Web Driver
웹 페이지
.HTML
doc.html
Soup
Preparation
Finding by
class names
Getting whole
text
Finding by
tag names
Finding
href
soup =
BeautifulSoup('doc.html','html.parser')

Google

facebook

twitter

coupang

amazon

삼성생명, 웹 크롤링 활용 '보험사기' 적발

이해리 기자 | 입력 2022.07.11 10:08 | 댓글 0

2014년 부당청구방지시스템 도입...'고객 보호' 노력

[뉴스포스트=이해리 기자] 삼성생명이 웹 크롤링(Web Crawling)을 통해 보험 사기 방지에 힘쓴다.

(사진=삼성생명)

11일 삼성생명에 따르면 보험사기 특별조사팀(SIU)은 웹 크롤링을 통해 인터넷 커뮤니티, 블로그 등을 통해 홍보되고 있는 백내장 관련 게시글 504개를 올 상반기에 확보했다. 진료비 할인, 이벤트 등의 내용을 담고 있는 게시글들은 모두 보험 사기와 연계된 브로커가 올린 광고다.

삼성생명은 이 중 4개 병원을 '보험 사기 외 브로커 연루 환자 유인, 알선 행위'로 수사의뢰 했다고 밝혔다.

https://www.newspost.kr/news/articleView.html?idxno=100279

앞서 삼성생명은 2014년 9월 보험 업계 최초로 '부당청구방지시스템(FDS)'을 도입한 바 있으며, 최근에는 웹 크롤링을 활용해 보험 사기를 적발하고 있다.

사회일반

‘야놀자 vs 여기어때’ 숙박업체 소송…2심도 ‘크롤링’ 불법 판정

2022.08.26 11:05

여기어때, 경쟁사 야놀자 숙박정보 무단수집
부정경쟁행위 영업손실 인정…10억원 배상
숙박업체 정보 그대로 가져와 제공 ‘불법’
웹사이트 소스 차용 ‘크롤링’ 분쟁 첫 사건

야놀자 서울 사옥[야놀자 제공]

[헤럴드경제=유동현 기자] 여행·숙박 어플리케이션 ‘여기어때’가 경쟁사 ‘야놀자’와의 소송 항소심에서도 일부 승소했다. 자동으로 타사 웹 사이트 정보를 복제해 활용하는 이른바 ‘크롤링’이 불법이라는 법원 판단이 그대로 유지돼 향후 비슷한 소송이 이어질 전망이다.

서울고법 민사4부(부장 이광만)는 25일 야놀자가 여기어때컴퍼니를 상대로 낸 데이터베이스 제작자 권리침해 소송에서 원고 일부 승소 판결했다. 판결이 확정되면 여기어때는 야놀자에게 10억원을 배상해야 한다.

네이버-다윈중개 '크롤링' 소송戰…플랫폼 업계 데이터 IP 논쟁 가열

구민기 기자

입력 2022.05.10 12:01 수정 2022.05.10 13:19

가가

네이버가 부동산 중개 플랫폼 다윈중개에 대해 ‘크롤링 금지’ 민사 소송을 제기했다. 다윈중개가 네이버부동산을 대상으로 하는 웹 크롤링(온라인 정보 수집 행위) 행위를 금지해달라는 내용이다. 가처분 금지 소송에서 ‘화해권고’가 난 사안이지만, 향후 다윈중개 뿐 아니라 경쟁자 크롤링을 원천차단하기 위한 법적 해석을 남기려 민사소송을 제기한 것으로 분석된다.

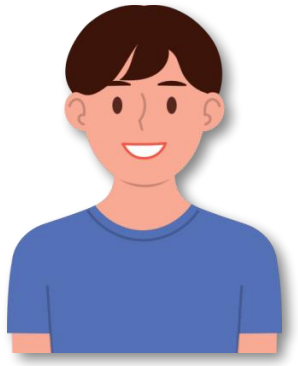
오늘의 주요





- 크롤러는 사람보다 빠르게 웹 페이지에 접근함
- 상대 웹사이트에 부하 발생 가능
- 서버는 전송량에 제한이 있고, 전송량에 따라 요금이 발생
- 우리 프로그램 때문에 상대방/기업이 경제적 피해를 입을 수 있음

크롤러를 사용할 때는 주의가 필요함

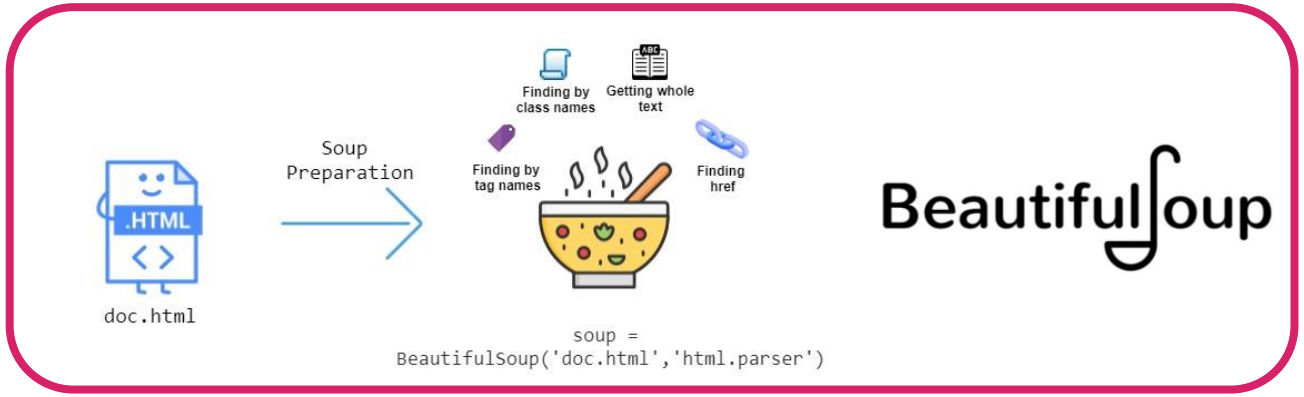


사용자

개발 언어
(프로그래밍 언어)

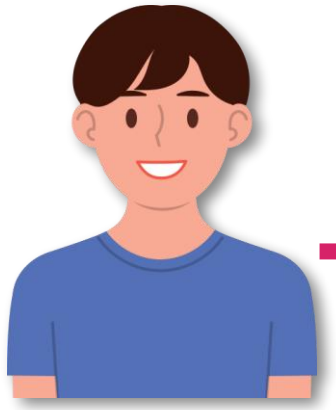
1 HTTPS Request 요청

Web Driver



2 Beautiful Soup를 이용한
웹페이지 분석 및 텍스트 추출

웹 페이지
(네이버 뉴스)



사용자

개발 언어
(프로그래밍 언어)

3 네이버 뉴스를 텍스트 파일로 만들기

naver_movie.txt

```
1 압도적인 영상미, 눈을 뗄 수 없는 연출. 본인만의 스타일 확실한 SF 영화. 무엇보다 음악이 신의 한 수
2 주연에 모래벌레도 출연시켜야..
3 배우들이 화려하게 뭘 하는 건 아닌데, 그냥 영화 자체가 화려함ㅋㅋㅋ내용은 그냥 서막 수준이라 지루할 수도 있는데, 비주얼적으
4 보는게 아니라 체험하는 영화. 두시간 반 모래바람 속에 있다 나온 기분..
5 이게영화지이게영화지
6 개인적이지만 이런 영화 좋아한다면 취향 다 때려박은 영화입니다.한번 더 볼 예정입니다.
7 기대했던 만큼 재미있었다. 영화관에서 안 봤으면 평생 후회했을 듯. 시간 가는 줄 모르고 봤음. 근데… 설마 여기서 끝나 에이 말
8 서사에 불과할뿐 앞으로 대장정의 시작 명쾌한 연기흡입력..
9 2편이 기다려지는 또하나의 SF걸작
10 드니 빌뇌브 감독에, 영상미는 황홀하고 음악은 웅장하고 캐스팅도 대박인데 이상하게 영화가 밋밋함
11 압도적인 영상미, 눈을 뗄 수 없는 연출. 본인만의 스타일 확실한 SF 영화. 무엇보다 음악이 신의 한 수
12 주연에 모래벌레도 출연시켜야..
13 배우들이 화려하게 뭘 하는 건 아닌데, 그냥 영화 자체가 화려함ㅋㅋㅋ내용은 그냥 서막 수준이라 지루할 수도 있는데, 비주얼적으
14 보는게 아니라 체험하는 영화. 두시간 반 모래바람 속에 있다 나온 기분..
15 이게영화지이게영화지
16 개인적이지만 이런 영화 좋아한다면 취향 다 때려박은 영화입니다.한번 더 볼 예정입니다.
17 기대했던 만큼 재미있었다. 영화관에서 안 봤으면 평생 후회했을 듯. 시간 가는 줄 모르고 봤음. 근데… 설마 여기서 끝나 에이 말
18 서사에 불과할뿐 앞으로 대장정의 시작 명쾌한 연기흡입력..
19 2편이 기다려지는 또하나의 SF걸작
20 드니 빌뇌브 감독에, 영상미는 황홀하고 음악은 웅장하고 캐스팅도 대박인데 이상하게 영화가 밋밋함
21 압도적인 영상미, 눈을 뗄 수 없는 연출. 본인만의 스타일 확실한 SF 영화. 무엇보다 음악이 신의 한 수
22 주연에 모래벌레도 출연시켜야..
23 배우들이 화려하게 뭘 하는 건 아닌데, 그냥 영화 자체가 화려함ㅋㅋㅋ내용은 그냥 서막 수준이라 지루할 수도 있는데, 비주얼적으
24 보는게 아니라 체험하는 영화. 두시간 반 모래바람 속에 있다 나온 기분..
25 이게영화지이게영화지
26 개인적이지만 이런 영화 좋아한다면 취향 다 때려박은 영화입니다.한번 더 볼 예정입니다.
27 기대했던 만큼 재미있었다. 영화관에서 안 봤으면 평생 후회했을 듯. 시간 가는 줄 모르고 봤음. 근데… 설마 여기서 끝나 에이 말
28 서사에 불과할뿐 앞으로 대장정의 시작 명쾌한 연기흡입력..
29 2편이 기다려지는 또하나의 SF걸작
30 드니 빌뇌브 감독에, 영상미는 황홀하고 음악은 웅장하고 캐스팅도 대박인데 이상하게 영화가 밋밋함
31 압도적인 영상미, 눈을 뗄 수 없는 연출. 본인만의 스타일 확실한 SF 영화. 무엇보다 음악이 신의 한 수
32 주연에 모래벌레도 출연시켜야..
33 배우들이 화려하게 뭘 하는 건 아닌데, 그냥 영화 자체가 화려함ㅋㅋㅋ내용은 그냥 서막 수준이라 지루할 수도 있는데, 비주얼적으
34 보는게 아니라 체험하는 영화. 두시간 반 모래바람 속에 있다 나온 기분..
35 이게영화지이게영화지
36 개인적이지만 이런 영화 좋아한다면 취향 다 때려박은 영화입니다.한번 더 볼 예정입니다.
37 기대했던 만큼 재미있었다. 영화관에서 안 봤으면 평생 후회했을 듯. 시간 가는 줄 모르고 봤음. 근데… 설마 여기서 끝나 에이 말
38 서사에 불과할뿐 앞으로 대장정의 시작 명쾌한 연기흡입력..
39 2편이 기다려지는 또하나의 SF걸작
```

시각화

영화 리뷰 텍스트 파일

개발 언어
(프로그래밍 언어)

KoNLPy

4 KoNLPy 라이브러리를 통한 한국어 분석

워드 클라우드

koNLPy 형태소 분석

STEP 5. 한글 분석 모듈 'koNLPy'을 통한 형태소 분석

STEP 5-1. koNLP 모듈 설치

[11]
```
# ======= 형태소 분석을 위해 한글 분석 모듈 konlpy를 설치한다. ==========
!python -m pip install konlpy
import konlpy
print('KoNLPy version...:', konlpy.__version__)
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting konlpy
  Downloading konlpy-0.6.0-py2.py3-none-any.whl (19.4 MB)
     |████████████████████████████████| 19.4 MB 7.6 MB/s
Requirement already satisfied: numpy>=1.6 in /usr/local/lib/python3.7/dist-packages (from konlpy) (1.21.6)
Requirement already satisfied: lxml>=4.1.0 in /usr/local/lib/python3.7/dist-packages (from konlpy) (4.9.1)
Collecting JPype1>=0.7.0
  Downloading JPype1-1.4.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.whl (453 kB)
     |████████████████████████████████| 453 kB 49.5 MB/s
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from JPype1>=0.7.0->konlpy) (4.1.1)
Installing collected packages: JPype1, konlpy
Successfully installed JPype1-1.4.0 konlpy-0.6.0
KoNLPy version...: 0.6.0
```

koNLPy 형태소 분석

STEP 5-2. 형태소 분석 진행

[12] 0초
```
# 한국어 텍스트 분석에 필요한 모듈(konlpy)의 Open Korean Text 형태소 분석기를 불러온다.
from konlpy.tag import Okt  # 토큰화 하는 작업. 속도가 매우 느림.
from collections import Counter # 빈도수 딕셔너리
from wordcloud import WordCloud # 워드 클라우드
import matplotlib.pyplot as plt # 차트 지원
import matplotlib as mpl
```

[13] 0초
```
# 앞서 저장한 리뷰 텍스트 파일 열기
with open ('/content/drive/MyDrive/1.BokjaCrawler/naver_movie.txt', 'r', encoding='utf-8') as f:
  doc = f.read()
```

[14] 1초
```
# Open Korean Text 형태소 분석기 객체를 생성한다.
okt = Okt()
```

[15] 6초
```
# 문장에서 명사만 추출
nouns = okt.nouns(doc)
```

[16] 0초
```
# 단어의 길이가 1개인 것은 제외하고, 리스트에 담는다.
words = [n for n in nouns if len(n) > 1]
print (len(nouns))
```

```
870
```

워드 클라우드 생성

STEP 6. 워드 클라우드 생성

[21] 1초

```
from wordcloud import WordCloud

wordcloud = WordCloud(
  font_path=font_path,
  width=1000, height = 1000,
  max_words=100,
#  scale=2.0,
#  relative_scaling=0.2,
  max_font_size=250,
  background_color='white',
).generate_from_frequencies(dict(data))
```

워드 클라우드 생성

```
1 from matplotlib import pyplot
2
3 pyplot.figure(figsize=(16, 10))  # width, height in inches
4 pyplot.imshow(wordcloud)
5 pyplot.axis('off')
6 pyplot.show()
```