

유튜브 데이터 수집

[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

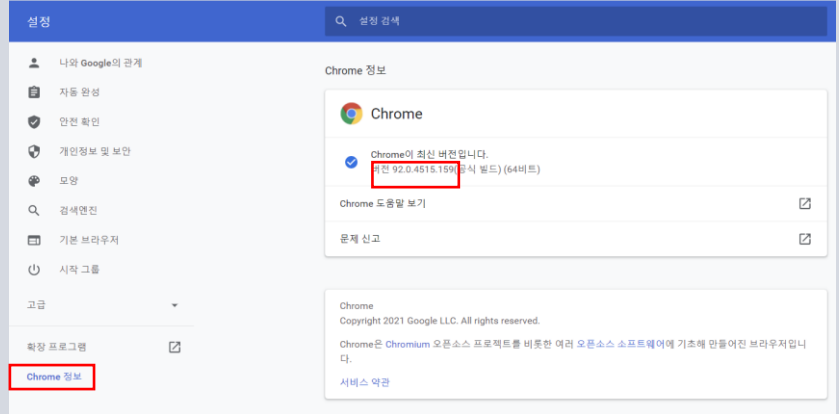
09/03 (금)



STEP 1 | selenium 라이브러리 사용을 위한 크롬 드라이버 설치

NH

크롬 창을 켜고 “우측 상단 ⋮ 버튼 - 설정”을 클릭합니다.
설정 페이지의 좌측 하단 “Chrome 정보”를 클릭하여 크롬 버전을 확인합니다.



크롬 드라이버 공식 홈페이지 (<https://chromedriver.chromium.org/downloads>)에서 사용하는 크롬 버전에 맞는 파일을 다운로드합니다.

SNUDM | 참고

크롬 드라이버는 크롬 브라우저를 제어하기 위해 사용되는 도구이다.

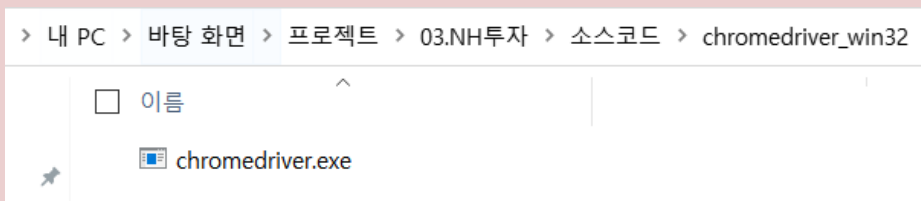
이후 웹 자동화 Python 라이브러리인 selenium을 사용할 때 이 크롬 드라이버를 활용하게 된다.

SNUDM | 주의

크롬 버전이 업데이트 되면 업데이트 된 맞추어 크롬 드라이버 파일을 새로 다운로드 해야 한다.

아래와 같은 파일이 생성되었으면 크롬 드라이버 설치가 성공적으로 완료된 것입니다.

실행 결과 예시



[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

09/03 (금)

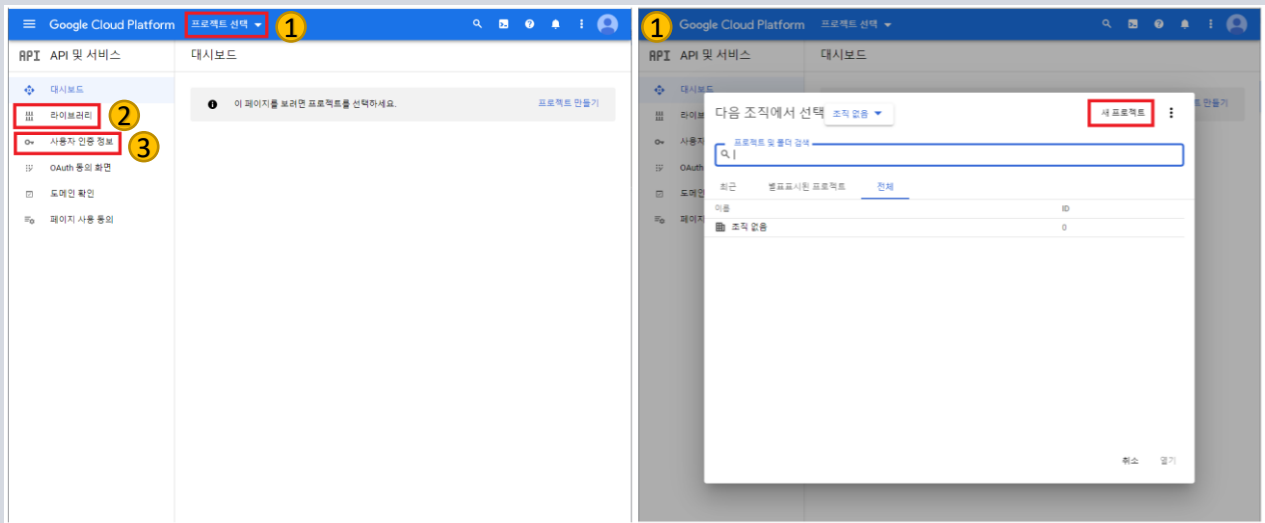


STEP 2 | Youtube Data API 사용을 위한 Google API Key 생성

NH

(1) 프로젝트 생성

- Google 계정 로그인 후 Google API console (<https://console.developers.google.com>)에 접속합니다.
- 상단 “프로젝트 선택 - 새 프로젝트” 버튼을 클릭하여 프로젝트를 생성합니다.



- 내용을 자유롭게 작성한 뒤 “만들기” 버튼을 눌러 프로젝트 생성을 완료합니다.

프로젝트 이름 *

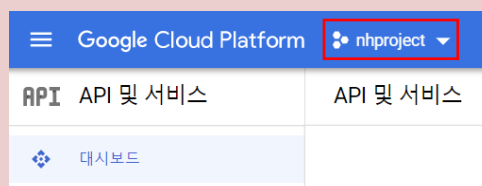
프로젝트 ID: nhproject-3

위치 *

상위 조직 또는 폴더

아래와 같이 위에서 설정한 프로젝트명이 표시되면
성공적으로 프로젝트가 생성된 것입니다.

실행 결과 예시



[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

09/03 (금)

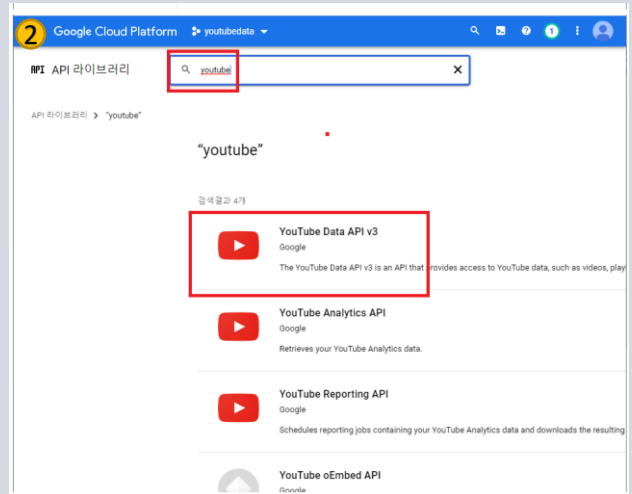
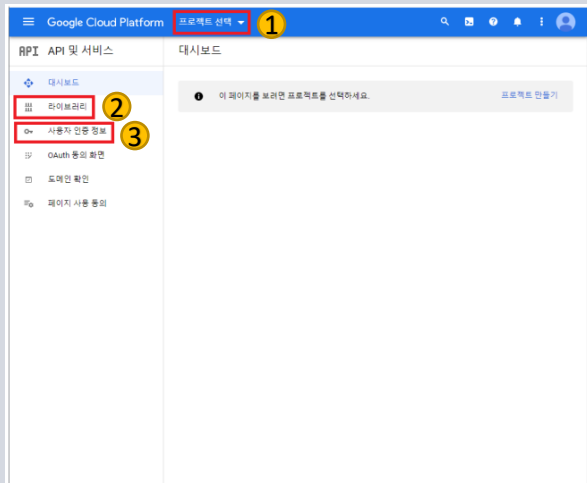


STEP 2 | Youtube Data API 사용을 위한 Google API Key 생성

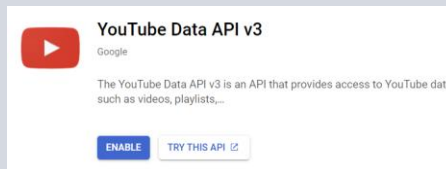
NH

(2) Youtube Data API V3 라이브러리 활성화

- Google API console (<https://console.developers.google.com>)에 접속합니다.
- 좌측 “라이브러리” 버튼을 클릭합니다.
- 검색창에 “youtube”를 입력하고, 검색된 ‘YouTube Data API v3’를 클릭합니다.



- “사용(ENABLE)” 버튼을 클릭하여 YouTube Data API v3를 활성화합니다.



[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

09/03 (금)

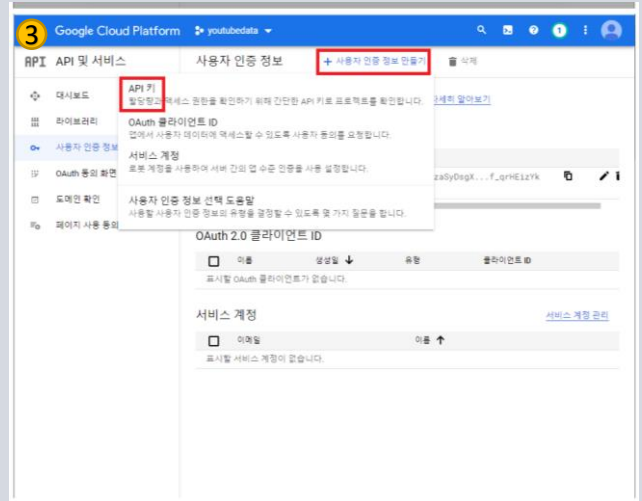
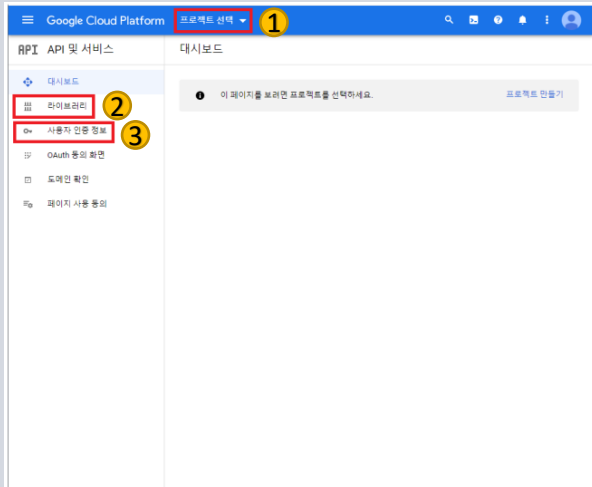


STEP 2 | Youtube Data API 사용을 위한 Google API Key 생성

NH

(3) 사용자 인증 정보 API key 발급

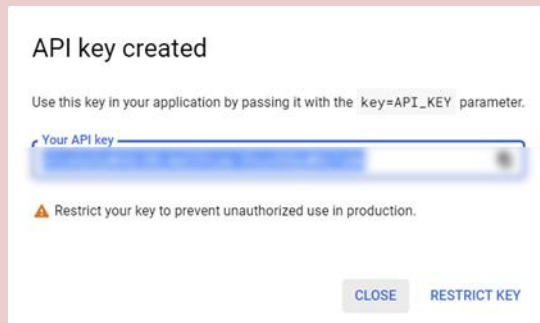
- Google API console (<https://console.developers.google.com>)에 접속합니다.
- 좌측 “사용자 인증 정보” 버튼을 클릭합니다.
- “+ 사용자 인증 정보 만들기 - API 키” 버튼을 클릭합니다.



아래와 같은 화면이 뜨면 성공적으로 API Key가 발급된 것입니다.

Key 값을 따로 메모해두시기 바랍니다.

실행 결과 예시



SNUDM | 참고

발급된 Key를 이용하여 유튜브 데이터 수집을 위한 Google API를 이용할 수 있다.

API Key 값이 외부에 노출되지 않도록 주의하여야 한다.

[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

09/03 (금)

STEP 3 | 라이브러리 설치

NH

Anaconda Prompt 상에서 가상환경 nh에 접속한 후 Google API 라이브러리를 설치합니다.

```
conda activate nh

pip install selenium
pip install google-api-python-client
```

Anaconda Prompt

STEP 4 | 특정 채널에 게시된 동영상 ID 목록 수집

NH

Jupyter Notebook 상에서 **Data_Youtube_1_video_ids.ipynb** 파일을 클릭합니다.
상단 메뉴에서 Kernel - Change kernel - nh를 클릭합니다.

아래 예시를 참고하여 앞에서 설치한 크롬드라이버 파일 경로를 입력합니다.

Python Code

```
CHROME_DRIVER_FILEPATH = r'C:\Users\Jihye Park\OneDrive - SNU\바탕 화면\프로젝트\03.NH투자\소스코드\chromedriver_win32\chromedriver'
```

SNUDM | 참고

위 코드에서 **r**은 raw string을 의미하는 Python 기호로, 윈도우 파일경로 안에 있는 역슬래시(\)가 문자열로 올바르게 인식되도록 처리한다.

Run 버튼을 클릭하여 코드를 실행합니다.
(채널명, 동영상 목록 링크) 형식으로 작성된 아래 두 정보를 대상으로 채널 별 동영상 ID 목록을 수집합니다.

Python Code

```
channels = [("Nate O'Brien", 'https://www.youtube.com/c/NateOBrien/videos'), \
            ("Graham Stephan", 'https://www.youtube.com/c/GrahamStephan/videos')]
```

Video_ids.csv에 저장된 내용을

실행 결과 예시

Jupyter Notebook 상으로 불러와 확인할 수 있습니다.

```
import pandas as pd
pd.read_csv('Video_ids.csv')
```

	channel_name	video_id	video_title
0	Nate O'Brien	CHimCbFWj7k	I Meditated Every Day... Here's What I Learned
1	Nate O'Brien	-LZyX9uHoJg	My 13 Sources Of Income (Explained)
2	Nate O'Brien	YZxeQ7xiOyU	How I Will Profit From Inflation

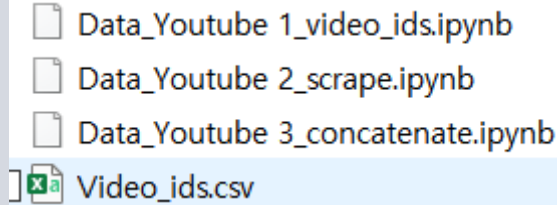
[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

09/03 (금)

STEP 5 | 유튜브 동영상 데이터 수집

NH

아래 이미지와 같이 Video_ids.csv 파일과 Data_Youtube *.ipynb 파일들을 같은 공간에 위치시킵니다.



Jupyter Notebook 상에서 **Data_Youtube 2_scrape.ipynb** 파일을 클릭합니다.
상단 메뉴에서 Kernel - Change kernel - nh를 클릭합니다.

아래 예시를 참고하여 앞에서 발급받은 API Key 값을 입력합니다.

video_ids_filepath 값으로 Video_ids.csv를 입력합니다.

Run 버튼을 클릭하여 코드를 실행합니다.

Video_ids.csv 파일에 기입된 동영상 ID 정보를 대상으로 동영상의 제목, 조회수 등의 정보를 수집합니다.

```
api_key='
video_ids_filepath = 'Video_ids.csv'
```

Python Code

SNUDM 참고

하나의 API Key 값으로 Google Youtube API를 사용할 수 있는 일일 할당량(quota)이 정해져 있다.

일일 할당량을 초과할 경우 아래와 같은 에러 메시지가 출력된다.

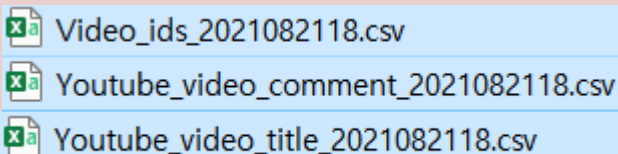
```
<HttpError 403 when requesting https://youtube.googleapis.com/youtube/v3/videos?part=snippet%2C+contentDetails%2C+statistics&id=C
HimCbFWJ7k&maxResults=100&key=AlzaSyDyWASLORL4mZlPcow-Ehus90QuWRzTs0w&alt=json returned "The request cannot be completed because
you have exceeded your <a href="/youtube/v3/getting-started#quota">quota</a>.". Details: "[{'message': 'The request cannot be com
pleted because you have exceeded your <a href="/youtube/v3/getting-started#quota">quota</a>.', 'domain': 'youtube.quota', 'reason':
'quotaExceeded'}]">
```

일일 할당량은 우리나라 시간으로 오후 5시에 초기화된다. 따라서 일일 할당량을 초과하지 않는 데까지 데이터를 수집하여 저장하고, 미처 수집하지 못한 동영상 ID 목록은 따로 저장하여 다음 날 오후 5시 이후에 수집해야 한다.

실행이 완료되면 아래와 같이 3개의 파일이 생성됩니다.

이때 파일명 맨끝에 있는 yyyymmddhh 형식의 문자열은

코드 실행이 완료된 때의 날짜 및 시간을 나타냅니다.



→ 처리되지 못한 동영상 ID 목록

→ 처리완료된 동영상 ID의 댓글 데이터

→ 처리완료된 동영상 ID의 제목 데이터

실행 결과 예시

아직 처리되지 못한 동영상 ID가 담긴 파일명(e.g., Video_ids_yyyymmddhh.csv)으로 바꾸어 입력합니다.
이 파일 내용을 대상으로 동영상의 제목, 조회수 등의 정보를 추가 수집합니다.

```
api_key='
video_ids_filepath = 'Video_ids_2021082118.csv'
```

Python Code

[SNUDM × NH투자증권] 소셜미디어 텍스트 분석을 통한 투자 관련 핫토픽 탐지 유튜브 데이터 수집

09/03 (금)

STEP 5 | 유튜브 동영상 데이터 수집

NH

Jupyter Notebook 상에서 **Data_Youtube 3_concatenate.ipynb** 파일을 클릭합니다.
상단 메뉴에서 Kernel - Change kernel - nh를 클릭합니다.



아래 예시를 참고하여 수집한 데이터가 담긴 파일들을 입력합니다.
Run 버튼을 클릭하여 코드를 실행합니다.
여러 개의 파일에 담겨있는 데이터들을 하나로 모으는 작업을 수행합니다.

Python Code

```
title_filepaths = ['Youtube_video_title_2021082118.csv', 'Youtube_video_title_2021082119.csv', 'Yo
utube_video_title_2021082216.csv']
comment_filepaths = ['Youtube_video_comment_2021082118.csv', 'Youtube_video_comment_2021082119.csv', 'Youtube_video_comment_2021082216.csv']
```

실행이 완료되면 아래와 같이 2개의 파일이 생성됩니다.

실행 결과 예시

 Youtube_comment_FULL.csv
 Youtube_title_FULL.csv

→ 동영상 댓글 데이터

→ 동영상 제목 데이터

앞으로 작업할 폴더/파일 구조는 아래와 같습니다.

```
├─ chromedriver_win32 # 크롬드라이버
├─ chromedriver
├─ data
│   ├── 0819_Save_FULL.csv
│   ├── Video_ids.csv
│   ├── Youtube_title_FULL.csv
│   ├── Youtube_comment_FULL.csv
│   ├── Data_twitter.ipynb # 트윗 수집
│   ├── Data_Youtube 1_video_ids.ipynb # 유튜브 데이터 수집 step 1
│   ├── Data_Youtube 2_scrape.ipynb # 유튜브 데이터 수집 step 2
│   └── Data_Youtube 3_concatenate.ipynb # 유튜브 데이터 수집 step 3
├─ Frequency_Twitter.ipynb # 트윗 대상 빈도수 기반 핫토픽 탐지
├─ Frequency_Youtube.ipynb # 유튜브 동영상 대상 빈도수 기반 핫토픽 탐지
└─ ...
```