M3.
크롤링을 위한 파이썬 라이브러리

강사 프로필

김 호 성

| 학력사항

- 상명대학교 경영학 학사
- 상명대학교 빅데이터과학 학사

| 경력사항 (프로젝트 등)

- 데이터마케팅코리아 데이터엔지니어, 프론트엔드 개발자
- 텍스트 분석솔루션 FoxTA 2.0 개발 담당
- AI 마케팅 분석 솔루션(Ma:deri) 프론트엔드, 백엔드 개발
- 한화투자증권 텍스트분석 분류체계 개발 지원
- 솔루션 Gamification 및 DPI(Digital Power Index) 기획 참여
- 공공사업, 컨설팅 데이터 시각화 담당(QlikSense, R, Python)

| 강의 경력사항

- 예술경영지원센터 크롤링 강의 지원

커리큘럼

과정 목표

01 온라인(웹) 상의 다양한 데이터를 수집하기 위해 데이터의 형태와 웹의 기본구조를 이해한다.

02 크롤링을 위해 파이썬의 다양한 라이브러리의 기능과 역할을 학습한다.

03 실제 온라인에서 제공되는 다양한 형태의 데이터들을 원하는 방식으로 수집할 수 있도록 코드를 작성한다.

모듈 명	주요 내용
M1. 데이터의 형태	• 데이터를 정의하고, 강의에서 활용할 데이터에 대한 유형을 소개 • JSON, CSV, TSV, XLSX, 등 데이터의 형태를 소개하고 간단한 활용법 학습 • Database에 대한 소개 및 SQLite를 활용한 간단한 활용법 숙지
M2. 웹의 기본구조	• 웹의 개념과 간단한 역사 및 작동원리에 대한 내용을 파악 • 실제 다양한 웹페이지를 통해 웹을 구성하고 있는 요소(HTML, Javascript, JSON 등)들을 학습
M3. 크롤링을 위한 파이썬 라이브러리	• 크롤링을 위한 다양한 파이썬 라이브러리에 대한 소개 • 주피터노트북을 활용하여 각 라이브러리에 대한 실습 진행
M4. 웹페이지를 활용한 데이터 수집	• 실제 동작하는 웹페이지를 분석하고 수집하여, 원하는 형태의 데이터로 변환하는 실습 진행 • 대상 : 네이버금융, 네이버뉴스, DART, 증권사이트 등
M5. API를 활용한 데이터 수집	• 공공기관이나 민간기업에서 제공하고있는 OpenAPI를 활용하여 데이터를 수집하는 실습 진행 • 대상 : 공공데이터(부동산,금융), SNS API, 쇼핑몰 API 등
M6. 데이터 수집 시스템 만들기	• 위에서 작성된 파이썬 코드를 활용하여, 정기적인 수집 시스템을 만드는 실습 진행

*상기 커리큘럼의 내용 및 교육시간은 기관 및 교육과정의 사정에 따라 변동이 가능합니다.

M3 세부 목차

M3. 크롤링을 위한 파이썬 라이브러리

M4를 위한 API KEY발급

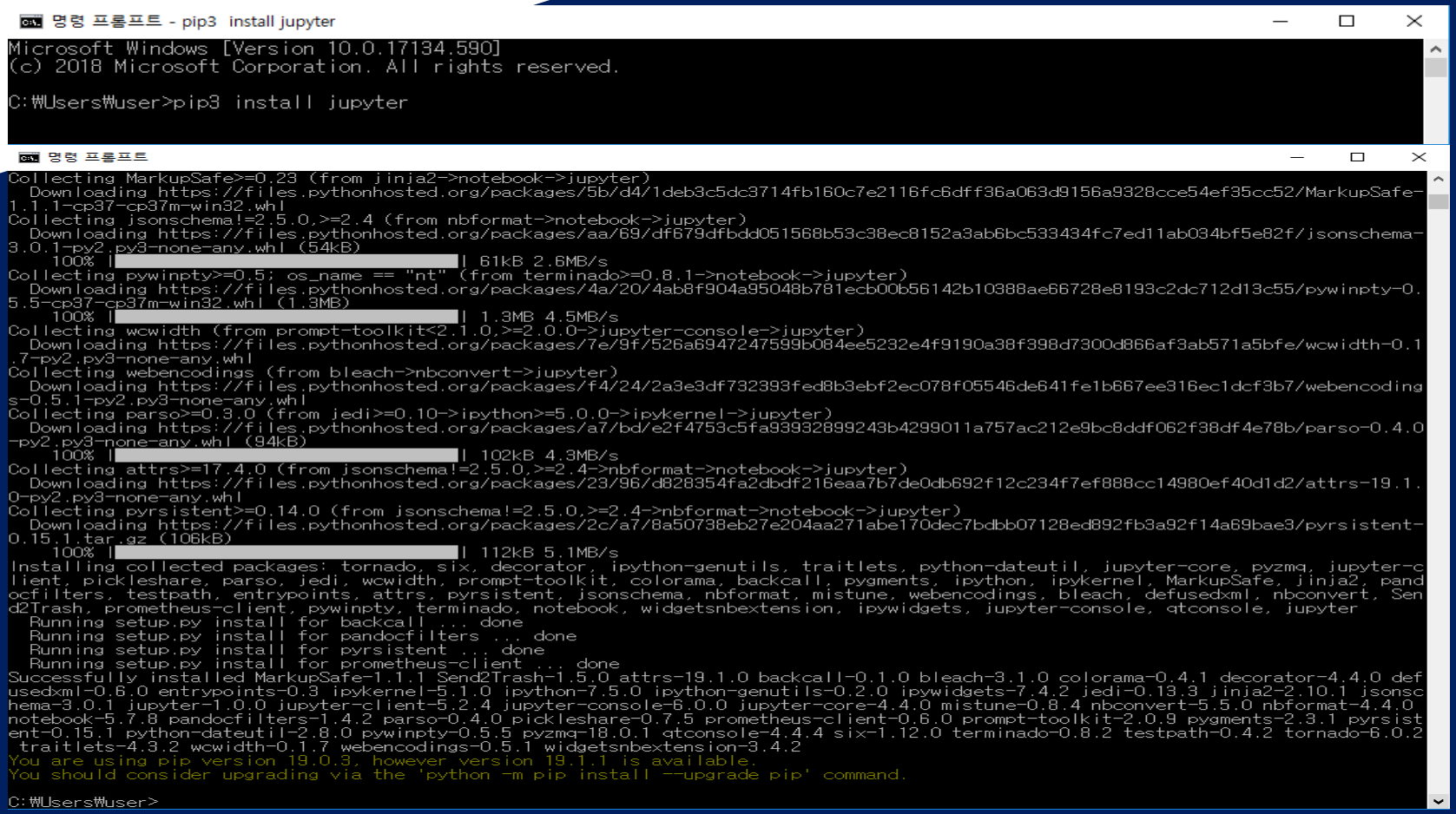
Jupyter 설치 및 크롤링 유의사항

urllib 기본 사용 및 예제

requests 및 BeautifulSoup 사용

Selenium을 통한 동적 크롤링

Jupyter 설치 및 크롤링 유의사항

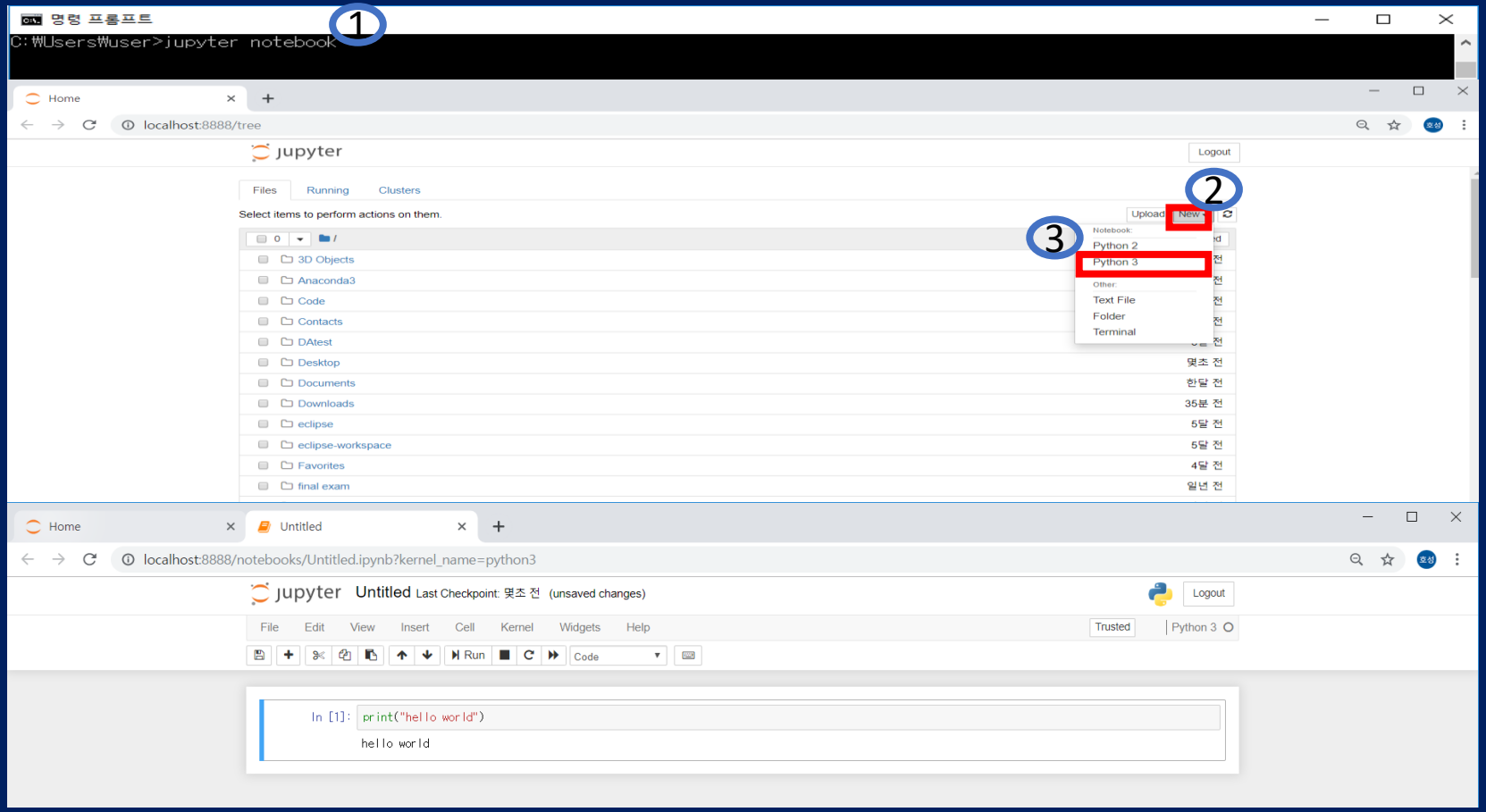


jupyter notebook설치

1. 실행창 〉 cmd (명령프롬프트) 실행

2. cmd창에 **pip3 install jupyter** 입력

3. 설치가 완료되면 다음 명령을 입력 받을 준비를 함



jupyter notebook설치

1.cmd창에 **jupyter notebook** 입력

2. jupyter notebook이 열리면 New버튼을 클릭

3.New버튼을 클릭한 뒤 나오는 메뉴에서 Python3선택

4. 명령어를 입력한 뒤 Ctrl+enter키를 눌러 제대로 작동 되는지 확인

구글 Colab

1.google 검색창에 colab을 검색
*웹 브라우저 크롬 권장

2. jupyter notebook과 같은 기능을 할 수 있는 것으로 구글계정만 있다면 따로 설치할 필요가 없음

3. 사용 목적에 따라서 jupyter notebook보다 좋을 수 있으나 강의에서는 jupyter를 기준으로 강의하므로 권장하지는 않음

실습

urllib 기본 사용 및 예제

데이터 예시

```
response = {
        "data":[
                {"rank":1, "keyword": " 초복", "val":17438},
                {"rank":2, "keyword": "중복", "val":5029},
                ...
                {"rank":10, "keyword": "철도의날", "val":1244}
                ]
        }
```

〈maderi.xx.xx로 요청시 〉

- html rendering
- 데이터가 비워진 영역이 있음

〈비동기 요청 〉

- html, json 형태로 데이터를 받아옴
- 데이터를 부르고 내부 javascript함수로 컴포넌트 에 맞춰 나타남

DataMKTKorea 데이터마케팅코리아

실습