웹크롤링 개발환경 (파이썬 몇버전을 사용해서 하겠다 / CSV 파일은 어떤식으로 처리해서 추출된 데이터를 어떻게 보관하고 할 것이다)

1. Python version 3 (Python 3.7.4)

Python은 고급 프로그래밍 언어로, 플랫폼 독립적이며 인터프리터식, 객체지향적, 동적 타이핑 대화형 언어이다.

< Python version 3의 특징>

* Python 2와 하위호환성이 없음

Python 2의 경우 2020년 1월 1일부로 지원을 종료할 에정

* 교차 플랫폼 언어

Python은 윈도우나, 리눅스, 유닉스, 매킨토시 등 다른 플랫폼에서도 동일하게 구동할 수 있는 이식성 언어

* C, C++, Java 같은 언어와 쉽게 통합

<사용할 예정 Library>

* 1. Beautiful Soup : HTML 이나 XML에서 데이터를 추출하는 Python Library. 웹문서의 구조를 찾아내는 파서를 이용해 찾고자 하는 데이터의 위치를 찾아 내어 값을 추출
  2. Selenium : Web앱을 테스트하는데 이용하는 프레임워크. ‘webdriver’ 라는 API를 통해 운영체제에 설치된 Chrome등의 브라우저를 제어 가능. 브라우저를 직접 동작시키기 때문에 JavaScript를 이용해 비동기적으로 혹은 뒤늦게 불러와지는 컨텐츠들을 가져올 수 있음. 또한 JS로 렌더링이 완료된 후의 DOM 결과물에 접근이 가능
  3. Tweepy : Twitter에게 API를 요청하고 정보를 얻는 모듈 (단, 7일 내의 자료만 얻을 수 있다는 단점이 있음)
  4. Pandas : 엑셀, 텍스트 파일, SQL 데이터 베이스 등의 데이터를 분석하기 위한 라이. 브러리. 행과 열로 이루어진 데이터 객체를 만들어 다룰 수 있어 보다 안정적으로 대용량 데이터들을 처리 가능한 도구
  5. Numpy : 행렬이나 일반적으로 대규모 다차원 배열을 쉽게 처리 할 수 있도록 지원하는 라이브러리. 데이터 구조 외에도 수치 계산을 위해 효율적으로 구현된 기능 제공