Project 4. Building a Web Crawler in Python

김지환 교수

Office: AS713

Tel: 705-8924

Email: kimjihwan@sogang.ac.kr

프로젝트 목표

- **Python3와 Python 라이브러리(BeautifulSoup4, Requests)를 사용하여 해당 웹 사이트의 모든 하이퍼링크와 하이퍼링크를 재귀적으로 방문하여 방문되는 모든 페이지의 text를 수집한다.**
- **Azure 서비스를 활용하여 과제를 수행하여 클라우드 서비스를 사용하는 법을 체득한다.**

프로젝트 요구사항

- 반드시 **Python3**를 이용하여 프로그래밍 한다.
- **Requests**와 **BeuatifulSoup4**를 이용하여 프로그래밍한다.
- **Crawling** 할 웹사이트는 **http://cspro.sogang.ac.kr/~gr120170213**
- 웹 사이트의 모든 하이퍼링크와 하이퍼링크를 재귀적으로 방문하여 방문하는 모든 페이지의 **text**를 수집한다.
- 중복된 **Crawling**이나 **Page**에 존재하지 않는 **Link**의 **Crawling**은 허용하지 않는다.

프로젝트 설명

- http://cspro.sogang.ac.kr/~gr120170213 를 Crawling 한다.
  - http://cspro.sogang.ac.kr/~gr120170213 는 닫힌계로 이루어진 웹사이트
    - 하이퍼링크로 연결된 page의 갯수는 유한하다.
  - Requests를 사용하여 HTML 파일을 얻는다.
  - BeautifulSoup4 library로 Parse tree를 생성한다.
    - http://cspro.sogang.ac.kr/~gr120170213 의 웹사이트의 첫 페이지를 Root page라 하고, 하이퍼 링크로 연결된 페이지들을 Descendant Page로 연결한다.
- 프로젝트 요구사항을 모두 만족하는 프로그램을 만든다.

프로젝트 설명

- 작업은 Microsoft Azure에서 가상 컴퓨터를 만들어 진행한다.
  - Domain name: sp+학번+.{리전}.cloudapp.azure.com
    - Ex> sp20181234.southeastasia.cloudapp.azure.com
  - Hardware spec: DS1_V2 standard
  - Software spec: Ubuntu 16 or 17 64bit
  - Username과 password or key 자유 – key 추천

제출 형식

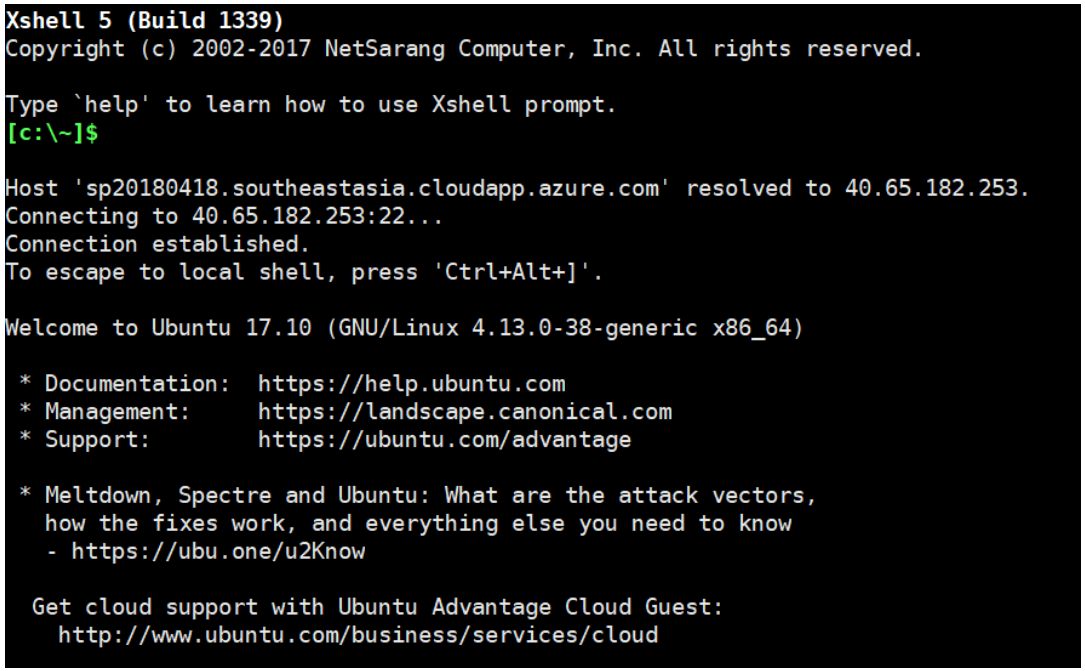
- **python3 : 반드시 python3를 사용하여 구현합니다.. 다른 프로그램을 사용하는 경우 0점 처리합니다.**
- **제출물(아래의 파일 중 1개라도 없는 경우에는 0점 처리합니다.)**
  - **URL.txt**
    - **방문한 Page의 URL을 방문한 순서대로 줄단위로 출력합니다.**

```
http://cspro.sogang.ac.kr/~gr120170213/index.html
http://cspro.sogang.ac.kr/~gr120170213/comments.html
http://cspro.sogang.ac.kr/~gr120170213/exception.html
(공백없이)방문한 URL\n      (마지막 줄엔 \n 생략)
```

  - **방문한 Page의 결과 텍스트 파일**
    - **Output_0001.txt, Output_0002.txt**
  - **Python 코드**
    - **학번.py**

제출 형식

- 제출물(아래의 파일 중 1개라도 없는 경우에는 0점 처리합니다.)
  - 학번.jpg(or png, gif 등, windows 환경에서 열 수 있는 임의의 그림 파일 형식)
    - IaaS 서비스를 이용하여 실행시킨 캡쳐 화면



채점 방식

- 제출물 중 1개라도 없는 경우에는 0점 처리한다.
  - Readme, document 제출 불필요
- 모두 제출한 URL.txt 파일 중 random하게 3개의 파일로 내용이 모두 완벽하게 Crawling 되었는지 판단한다.
  - Page당 5점 감점
- Azure의 크레딧을 전부 사용하거나, 기타 사용자의 실수로 Azure를 사용할 수 없는 경우 Amazon의 EC2를 사용하여 진행한다.
  - 이 경우, 1일 Late와 동일한 10%감점으로 처리한다
  - 학생의 실수가 아닌 예기치 못한 경우 조교에게 별도 연락한다.
- 무한Loop 발생 시 0점 처리한다.

제출 방법

- **sp학번_proj4라는 이름의 디렉터리를 만들고, 이 디렉터리에 제출파일, Document, readme 파일을 넣어서 디렉터리를 tar로 압축하여 한 파일로 만들어 메일로 보내시기 바랍니다.**
- **제출 파일은 sp<학번>_proj4.tar 입니다.**
- **제출 기한 : 2018년 5월 14일**
- **제출 주소 : ied2017c3@gmail.com**

  메일제목 형식 : [SPproj #4] 학번 이름

주의 사항은 이전 프로젝트와 같습니다.