Project 4. Building a Web Crawler in Python

박운상 교수

Office: K 336

Tel: 705-8936

Email: unsangpark@sogang.ac.kr

프로젝트 목표

- **Python과 Python 라이브러리(BeautifulSoup4, Requests)를 사용하여 웹 사이트의 모든 하이퍼링크와 하이퍼링크를 재귀적으로 방문하여 방문되는 모든 페이지의 text를 수집한다.**

프로젝트 요구사항

- **Python을 이용하여 프로그래밍 한다.**
- **Requests와 BeuatifulSoup4를 이용하여 프로그래밍한다.**
- **Crawling 할 웹사이트는 cspro.sogang.ac.kr/~gr120160213**
- **웹 사이트의 모든 하이퍼링크와 하이퍼링크를 재귀적으로 방문하여 방문하는 모든 페이지의 text를 수집한다.**
- **중복된 Crawling이나 Page에 존재하지 않는 Link의 Crawling은 허용하지 않는다.**

프로젝트 설명

- **cspro.sogang.ac.kr/~gr120160213 의 웹사이트를 Crawling 한다.**
  - **cspro.sogang.ac.kr/~gr120160213 는 닫힌계로 이루어진 웹사이트**
    - **하이퍼링크로 연결된 page의 갯수는 유한하다.**
  - **Requests를 사용하여 HTML 파일을 얻는다.**
  - **BeautifulSoup4 library로 Parse tree를 생성한다.**
    - **cspro.sogang.ac.kr/~gr120160213 의 웹사이트의 첫 페이지를 Root page라 하고, 하이퍼 링크로 연결된 페이지들을 Descendant Page로 연결한다.**
- **프로젝트 요구사항을 모두 만족하는 프로그램을 만든다.**

제출 형식

- **python2 : 반드시 python2를 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.**

```
cspro.sogang.ac.kr/~gr120160213
cspro.sogang.ac.kr/~gr120160213 /0001.htm
cspro.sogang.ac.kr/~gr120160213 /0002.htm
(공백없이)방문한 URL\n      (마지막 줄엔 \n 생략)
```

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

채점 방식

- 제출물 중 1개라도 없는 경우에는 0점 처리한다.
- 모두 제출한 URL.txt 파일 중 random하게 3개의 파일로 내용이 모두 완벽하게 Crawling 되었는지 판단한다.
  - Page당 5점 감점
- 무한Loop 발생 시 0점 처리한다.
- Time Complexity를 계산하여, 평균시간보다 2배 낮을 경우는 감점 요인이 되며 2배 높을 경우는 가산점 요인이 된다.

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 : 2016년 5월 10일**
- **제출 주소 : sp2016proj@gmail.com**

  메일제목 형식 : [SP숙제 #4] 학번 이름

주의 사항은 이전 프로젝트와 같습니다.