|  |  |
| --- | --- |
| 관리번호 | 중앙대-컴공-15 |
| 작성자 | 염승윤 |
| 편집기간 | 2019.03.25 |
| 비밀등급 | 대외비 |

|  |  |
| --- | --- |
| 송준호 | 최지혁 |
|  |  |

|  |
| --- |
| **19년 상반기 AdBlogBlock 파싱 모듈 진행 상황 (보고)** |

|  |
| --- |
| **1** |

사용한 파이썬 라이브러리

① re : 정규식 라이브러리로써 파싱할 때 가장 속도가 빠르다.

② urllib.request.urlopen : Html을 String으로 변환해준다.

③ urllib.error.HTTPError : 웹 사이트를 열 때 발생하는 오류를 알려준다.(예시 404 Page not found)

|  |
| --- |
| **2** |

코드 내부 변수 및 함수 구조

변수

inner\_html (list of string) : 한동안 문제였던 <jframe> 태그에 숨어있던inner html의 URL을

저장하는 리스트이다.

pics (list of string) : inner html(블로그 내부)의 사진들의 URL을 저장하는 리스트이다.

num\_pic(int) : inner html(블로그 내부)의 사진들의 개수를 저장하는 정수이다.

target\_url(String) : 네이버 블로그의 상위 URL을 저장하는 문자열이다.

(https://blog.naver.com)

html, link, links : html을 파싱한 결과를 저장하는 문자열이다.

함수

Web\_parse(String url) : url 문자열을 기준으로 해당 사이트 Html코드 전체를 반환한다.

Re.findall(’정규식’,HTML) : HTML 코드 중 정규식 조건에 맞는 문자열을 반환한다.

|  |
| --- |
| **3** |

순서

① 파싱할 웹 페이지의 주소를 지정한다.

② 파싱한 결과로 <iframe [모든 문자열]> 안에 src=으로 시작하는 url을 읽어온다.

(inner html을 찾기 위함이다. 예시 /PostView.njn? 예시를 보면 앞에 naver.com이 없다. target\_url로 완전한 URL을 얻을 수 있다.)

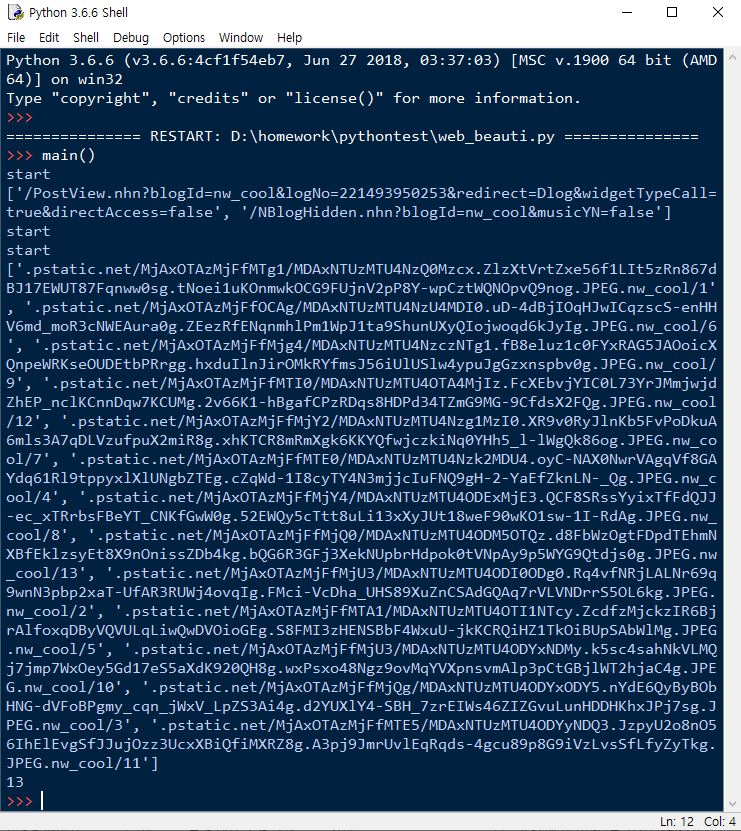
③ inner html을 파싱한다.

④ inner html의 이미지 파일을 파싱한다.(추후 네이버 스티커, 광고 문자열 등을 추가할 계획이다.)

⑤ 사진이 몇 개 있는지 센다. (pic\_url 리스트에 원소가 몇 개 있는지 보면 된다.)

|  |
| --- |
| **4** |

결과



<https://blog.naver.com/nw_cool/221493950253>

이 블로그를 파싱한 결과이다. 직접 세보면 13개의 사진을 업로드 했는데 프로그램의 결과도 동일하다.(JPG 파일들의 URL이 불완전 하다. 저 URL 앞에 target\_url만 붙혀주면 완성된다.)

완벽한 작전 수행을 위한 철저한 준비 태세 유지