Practice1 _ 2021105662 _ 홍예림

```python
#!/user/bin/env python3
from bs4 import BeautifulSoup
from urllib.request import *
from urllib.parse import *
from os import makedirs
import os.path, time, re

proc_files = {}

def enum_links(html, base):
    soup = BeautifulSoup(html, "html.parser")
    links = soup.select("link[rel='stylesheet']")
    links += soup.select("a[href]")
    result = []

    for a in links:
        href = a.attrs['href']
        url = urljoin(base, href)
        result.append(url)
    return result

def download_file(url):
    o = urlparse(url)
    savepath = "./" + o.netloc + o.path
    if re.search(r"/$", savepath):
        savepath += "index.html"
    savedir = os.path.dirname(savepath)

    if os.path.exists(savepath): return savepath

    if not os.path.exists(savedir):
        print("mkdir=", savedir)
        makedirs(savedir)

    try:
        print("download=", url)
        urlretrieve(url, savepath)
        time.sleep(1)
        return savepath
    except:
        print("다운 실패: ", url)
        return None
```

```python
def analyze_html(url, root_url):
    savepath = download_file(url)
    if savepath is None : return
    if savepath is proc_files : return
    proc_files[savepath] = True
    print("analyze_html=", url)

    html = open(savepath, "r", encoding="utf-8").read()
    links = enum_links(html, url)
    for link_url in links:
        if link_url.find(root_url) != 0:
            if not re.search(r".css$", link_url): continue

        if re.search(r".(html|htm)$", link_url):
            analyze_html(link_url, root_url)
            continue

        download_file(link_url)

if __name__ == "__main__":
    url = "https://docs.python.org/3.5/library/"
    analyze_html(url, url)
```

실행 코드

Enum_links 함수를 이용해 href 속성을 추출한 뒤, link를 절대경로로 변환한다.

Download_file 함수를 이용해 폴더를 생성해 link를 저장한다.

Analyze_html함수를 이용하여 html파일을 분석한다. Html 파일 안에 있는 link들을 추출하여 저장한다.

```
practice1.py
mkdir= ./docs.python.org/3.5/library
download= https://docs.python.org/3.5/library/
analyze_html= https://docs.python.org/3.5/library/
mkdir= ./docs.python.org/3.5/_static
download= https://docs.python.org/3.5/_static/pydoctheme.css
download= https://docs.python.org/3.5/_static/pygments.css
download= https://docs.python.org/3.5/library/intro.html
analyze_html= https://docs.python.org/3.5/library/intro.html
download= https://docs.python.org/3.5/library/functions.html
analyze_html= https://docs.python.org/3.5/library/functions.html
download= https://docs.python.org/3.5/library/constants.html
analyze_html= https://docs.python.org/3.5/library/constants.html
download= https://docs.python.org/3.5/library/stdtypes.html
analyze_html= https://docs.python.org/3.5/library/stdtypes.html
download= https://docs.python.org/3.5/library/exceptions.html
analyze_html= https://docs.python.org/3.5/library/exceptions.html
download= https://docs.python.org/3.5/library/text.html
analyze_html= https://docs.python.org/3.5/library/text.html
download= https://docs.python.org/3.5/library/string.html
analyze_html= https://docs.python.org/3.5/library/string.html
download= https://docs.python.org/3.5/library/re.html
analyze_html= https://docs.python.org/3.5/library/re.html
download= https://docs.python.org/3.5/library/difflib.html
analyze_html= https://docs.python.org/3.5/library/difflib.html
download= https://docs.python.org/3.5/library/textwrap.html
analyze_html= https://docs.python.org/3.5/library/textwrap.html
download= https://docs.python.org/3.5/library/unicodedata.html
analyze_html= https://docs.python.org/3.5/library/unicodedata.html
download= https://docs.python.org/3.5/library/stringprep.html
analyze_html= https://docs.python.org/3.5/library/stringprep.html
download= https://docs.python.org/3.5/library/readline.html
analyze_html= https://docs.python.org/3.5/library/readline.html
download= https://docs.python.org/3.5/library/rlcompleter.html
```

실행 결과

html파일을 분석하여 html안에 있는 link들을 추출한다. 해당 link들이 파일인 경우 다운을 받는다. 다운 받은 파일이 html 파일이라면 해당 html파일도 분석하여 link들을 추출한다.