

2022년 10월 2일 수정일 현재 디스코드에서 무료 핫딜 공지가 이루어지고 있다.

아래 디스코드 방 초대 링크를 클릭하면 서버에 입장할 수 있다.

<https://discord.gg/Frr7mxvV85>

얼마 전 지인을 통해 에펠펜코리아, 줄여서 펌코라고 불리는 커뮤니티 사이트의 핫딜 게시판에 괜찮은 할인 가격 상품이 많이 올라온다는 정보를 입수했다.

↑
9



사조 로하이 전자레인지 팝콘 80g 12입 x 2세트 [28]
쇼핑몰: **지마켓** / 가격: **16,420원** / 배송: **없음**
먹거리 / 20:58 / 남는라인감

↑
37



해남 세척 핫 팔고구마 실속형 5kg [38] 포텐
쇼핑몰: **옥션** / 가격: **9,900원** / 배송: **무료**
먹거리 / 20:10 / 코리프리

↑
67



유혜광 수제 통등심돈까스 10장 [140] 포텐
쇼핑몰: **인터파크** / 가격: **하나12,680원** / 배송: **0원**
먹거리 / 19:48 / 천사인4049

↓
-3



추희자두 5kg 중 55-65과 [27]
쇼핑몰: **하프클럽** / 가격: **18,320원** / 배송: **무료**
먹거리 / 19:41 / 라라랜드

↓
-16



TCL 4K UHD HDR 55인치 스마트티비 [34]
쇼핑몰: **쿠팡** / 가격: **478,000원** / 배송: **무배**
가전제품 / 19:41 / 아리카나리

↑
7



LG전자 트롬 워시타워 W20VAN 세탁23kg 건조20kg [22]
쇼핑몰: **하우스엘** / 가격: **2,329,000원** / 배송: **무료**
가전제품 / 19:22 / 꿀렁이수비대

↑
68



근본 트랙스타 가을할인 [122] 포텐
쇼핑몰: **트랙스타** / 가격: **30%~50%** / 배송: **무료**
의류 / 19:18 / 성글뉴캐슬랜

↑
33



배라 파인트 (포장) [31]
쇼핑몰: **해피오더** / 가격: **5,900원** / 배송: **무료**
먹거리 / 19:14 / 알면서도알지못해

↑
4



광동 우영차 500ml x 24pet (유통기한 22.10.21) [12]
쇼핑몰: **네이버쇼핑** / 가격: **7,900원** / 배송: **0원**
먹거리 / 19:13 / 히트만

↑
1



지금 수확!! 포슬포슬 2022년 햃 감자 중사이즈 5kg [13]
쇼핑몰: **쿠팡** / 가격: **와우8,900** / 배송: **0원**
먹거리 / 18:59 / 히트만

↑
26



2080 9모션 앵커리스 칫솔 12개 [25] 포텐
쇼핑몰: **지마켓** / 가격: **10,950원** / 배송: **무배**
생활용품 / 18:59 / 보쌈한일꺼억

↑
72



QCY H2 무선 블루투스 헤드셋 1+1 [232] 포텐
쇼핑몰: **티몬** / 가격: **32,500원** / 배송: **무료**
PC제품 / 18:46 / xvq

제목을 클릭하면 제목과 내용이 표시되는 팝업 메뉴가 나타나고, 1부터 10까지의 페이지 번호가 표시된다.

에펌코리아

Copyright © www.felker.com All rights reserved.

에펌코리아 핫딜 게시판 사진

가격인 무료인 상품만 구매하려고 하는 사람들이 분명 있을 것 같은데, 가격인 무료인 상품만 검색할 방법이 없었다.

그래서 직접 구현해 보기로 했다.

그리고 그 결과는...

[보안 시스템에 의한 자동 차단]

사람이 아닌 자동화 프로그램에 의한 비정상적인 반복적인 접속이 탐지되어 에펠포리아 보안 시스템에 의해 사용하시는 IP가 차단되었습니다.

일부 브라우저에서 새로고침 키를 실수로 계속 누르고 있으면 차단될 수 있습니다.

공용 IP인 경우에는 다른 사용자의 행동에 의해 잘못된 차단도 가능합니다.

잘못 차단되었다고 생각하는 경우 VPN이나 warning 우회용 프로그램 이용하고 있는 경우 껴보시길 바랍니다.

브라우저 아닌 프로그램/앱으로 접속하고 있는 경우 해당 프로그램/앱 문제일 것 입니다.

관련 문제되는 앱이 의심되면 이메일로 사용하던 앱 이름을 알려주시길 바랍니다.

24 시간 이후에 자동으로 차단이 풀립니다.

잘못 차단된 경우 help@fmkorea.com에 하단의 정보와 함께 관련 문의를 하시길 바랍니다.

시간:

차단 종류: 1 1

IP:

나라:

접속 종류:

ASNorg:

에펠포리아 2회차 차단 화면

정확히 세어보지는 않았지만 대략 50번 연속 get 방식으로 https 요청을 보냈더니 리턴 값의 status_code가 200으로 오지 않았다.

코드 문제인가 싶어 당황했는데, 해당 사이트에 IP 차단을 당해서 발생한 문제였다.

차단된 상태에서 에펠포리아에 접속하면 위 사진과 같은 화면이 나온다.

처음에는 reCAPTCHA 인증만 하면 차단이 풀렸지만, 다시 차단을 당했을 때는 24시간 동안 기다려야 했다.

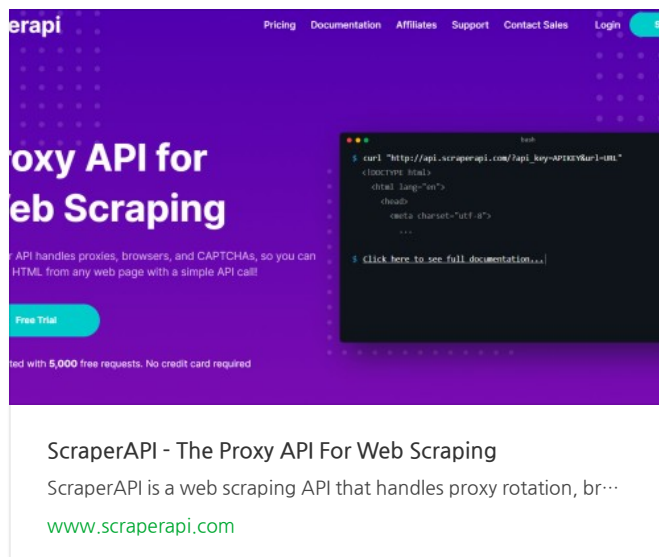
찾아보니 여러 가지 해결 방법이 있었다.

스크래핑 중간중간 `time.sleep()` 함수를 이용하여 스크래핑 시간을 지연시키는 방법, 토르를 이용하여 요청 IP를 바꾸는 방법, User-Agent를 바꾸는 방법 등등.

매 요청 시마다 User-Agent를 바꾸는 방법은 통하지 않아서(이 방법을 쓰다가 24시간 차단되었다), 매 요청 시마다 IP를 바꾸는 방식을 선택했다.

ScrapperAPI라는 곳에서 해당 서비스를 제공하고 있어 이용해 보았다.

<https://www.scraprapi.com>



회원가입은 GitHub 연동하면 금방인데, 이상하게 회원가입할 때 유독 사이트 속도가 너무 느려서 좀 걸렸다.

API 사용할 때 참고했던 [블로그 글](#)에 따르면 무료 플랜의 경우 매월 5,000건의 프록시 API 요청이 가능하고, 최대 5개의 동시 요청이 허용된다고 한다.

매달 5,000 크레딧이 갱신되는 것인지, 한 달 동안만 이용이 가능한 것인지, 5,000건 요청을 넘겼을 때 유료 플랜으로 전환해야 하는 것인지 명확히는 모르겠지만 일단 써보기로 했다.

```
$ python a.py
time taken to execute random_proxy(): 0.143
time taken to execute random_proxy(): 0.162
time taken to execute random_proxy(): 0.176
time taken to execute random_proxy(): 0.201
time taken to execute random_proxy(): 0.245
time taken to execute random_proxy(): 0.246
time taken to execute random_proxy(): 0.3
time taken to execute random_proxy(): 0.319
time taken to execute random_proxy(): 0.318

$ python main.py
어느 날짜 이후까지 검색할지 입력해주세요.
입력 예시: 2022.09.11
2022.09.23
time taken to execute random_proxy(): 3.479
time taken to execute random_proxy(): 1.493
time taken to execute random_proxy(): 2.109
time taken to execute random_proxy(): 2.412
time taken to execute random_proxy(): 6.648
time taken to execute random_proxy(): 1.277
time taken to execute random_proxy(): 5.136
time taken to execute random_proxy(): 3.615
time taken to execute random_proxy(): 2.551
검색이 완료되었습니다.
```

random_proxy()는 url과 path, parameter를 인자로 받아 requests의 get 함수로 http(s) 요청하는 함수다.
a.py는 대학교 공지 게시판에 IP 변환 없이 요청하는 소스 코드, main.py는 펌코 핫딜 게시판에 IP 변환해가며 요청하는 소스 코드이다.

함수 실행에 걸리는 시간을 터미널에 출력해 보니 대략 10배 정도 차이가 나는 것 같았다.

너무 오래 걸리면 사용하기 불편하니, 다음번에 코드를 수정하게 되면 좀 더 빠른 방식으로 바꾸어야 할 것 같다.

```
$ python main.py
어느 날짜 이후까지 검색할지 입력해주세요.
입력 예시: 2022.09.11
2022.09.23
동원 칠레 한정샐,가브리살 냉동 100g 1825 [146]
https://www.fmkorea.com/5051401657
[안드]던전공주 : 던전RPG [28]
https://www.fmkorea.com/5050142258
Republique (잠입 액션 게임) $4.99 -> 무료 [25]
https://www.fmkorea.com/5049219022
애플 TV+ 6개월 무료체험 (카카오페이) [281]
https://www.fmkorea.com/5046843968
검색이 완료되었습니다.
```

지금까지 완성된 프로그램의 출력 형태는 이런 느낌이다.

이걸 웹으로 이식하면 된다.

특정 기간을 스크롤바로 선택할 수 있게끔 하고, 내 무료 핫딜 상품 글 제목과 글 url을 웹상에 보기 좋게 출력하면 될 듯.

그런데 블로그에 정리하다 보니 생각이 좀 바뀌었다.

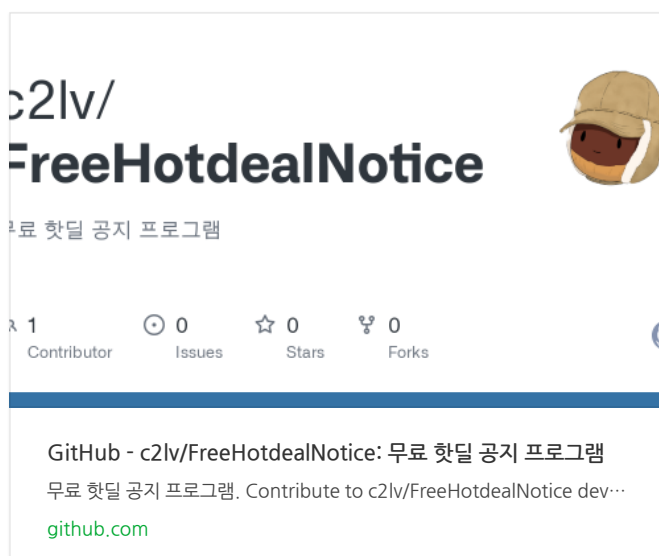
매번 검색하여 찾는 것보다 무료 상품 글이 올라왔을 때 알림을 주는 게 더 나은 서비스이지 않을까 싶다.

특정 시간마다 일정 시간 동안 올라온 글 중 무료 상품 글만 모아 카카오톡이나 디스코드 메시지로 알려주는 서비스로 만들어봐야겠다.

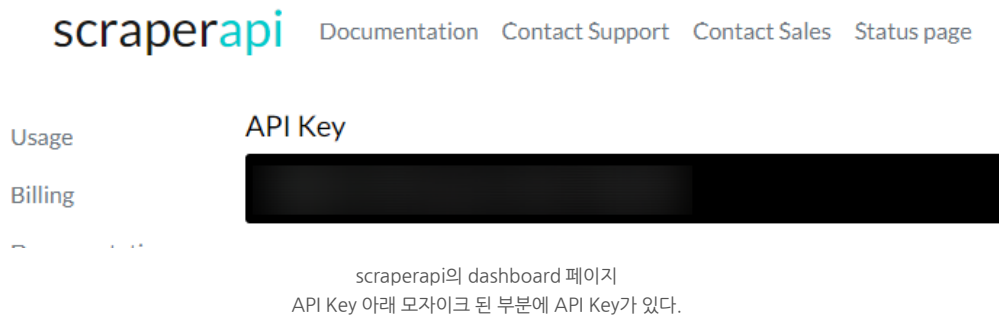
짧은 간격으로 자주 실행되게끔 만들면 코드가 실행될 때마다 많은 페이지를 요청할 필요가 없어 IP 차단을 염려할 필요도 없다.

글 작성 시점까지 작성한 소스 코드는 깃허브에 업로드해두었다.

<https://github.com/c2lv/FreeHotdealNotice>



const.py의 SCRAPER_API_KEY 값이 비어있는데, 여기에는 본인의 scraperAPI API key를 넣어주면 된다.
API key는 scraperapi 웹페이지에서 로그인하고 dashboard 가보면 있다.



robots.txt라는 것에 대해 알게 된 시점부터 스크래핑을 할 때 나름 신경 쓰려고 노력하는 편이다.

robots.txt는 웹사이트에 웹 크롤러같은 로봇들의 접근을 제어하기 위한 규약이다. 아직 권고안이라 꼭 지킬 의무는 없다.

크롤러들은 주로 검색엔진들의 인덱싱 목적으로 사용되는데, 웹사이트들 입장에서도 더 많은 검색 노출을 원하는게 일반적이므로 딱히 막을 이유는 없다. 다만 서버의 트래픽이 한정돼있거나 검색엔진의 노출을 원하지 않는 경우, 이 robots.txt에 “안내문” 형식으로 특정 경로에 대한 크롤링을 자제해 줄 것을 권고하는 것이다. 지킬 의무가 없다고 하나 지켜주는 게 상식이며, 마찬가지로 서버 주인 입장에서는 규칙을 지키지 않는 크롤링이 들어오는데도 계속해서 서비스를 제공할 의무 또한 없으므로 크롤러의 아이피를 차단하면 그만이다.

robots.txt는 웹사이트의 최상위 경로(=루트)에 있어야 한다. 즉, 사이트를 치고 슬래시 후 바로 robots.txt를 넣으면 볼 수 있다는 것이다

<https://namu.wiki/w/robots.txt>
최근 수정 시각: 2022-05-08 15:48:10

펄코의 robots.txt 내용은 다음과 같다.

```
User-agent: ia_archiver
Disallow: /

User-agent: PetalBot
Disallow: /

User-agent: ICCrawler
Disallow: /

User-agent: Linguee Bot
Disallow: /

User-agent: Tweetmemebot
Disallow: /

User-agent: dotbot
Disallow: /

User-agent: AhrefsBot
Disallow: /

User-Agent: MJ12bot
Disallow: /

User-agent: BoardReader
Disallow: /

User-agent: BLEXBot
Disallow: /

User-agent: CCBot
Disallow: /

User-Agent: The Knowledge AI
Disallow: /

User-agent: *
Disallow: /*act=IS&
Disallow: /*search_keyword=
Disallow: /*act=dispMemberBookmark
Disallow: /*_filter=
Disallow: /*m=
Disallow: /*_loader
```

<https://www.fmkorea.com/robots.txt>

나의 요청 URL은 아래 6가지 Disallow 형식에 해당하지 않아서, 해당 URL로는 스크래핑을 해도 괜찮다고 판단했다.

그래도 너무 빠른 스크래핑은 서버에 부하를 주고 관리자와 다른 유저들에게 피해를 줄 수 있는 행동이니 IP 차단이 풀리면 조심해야겠다.

+ 2022.10.01 추가

[Python] 펌코 무료 핫딜 상품 스크래핑 프로그램 (2)

<https://blog.naver.com/hyeonjun7/222889268953>

이 글과 이어진다.