Veri Kazıma Etik Midir?

Pınar Dağ

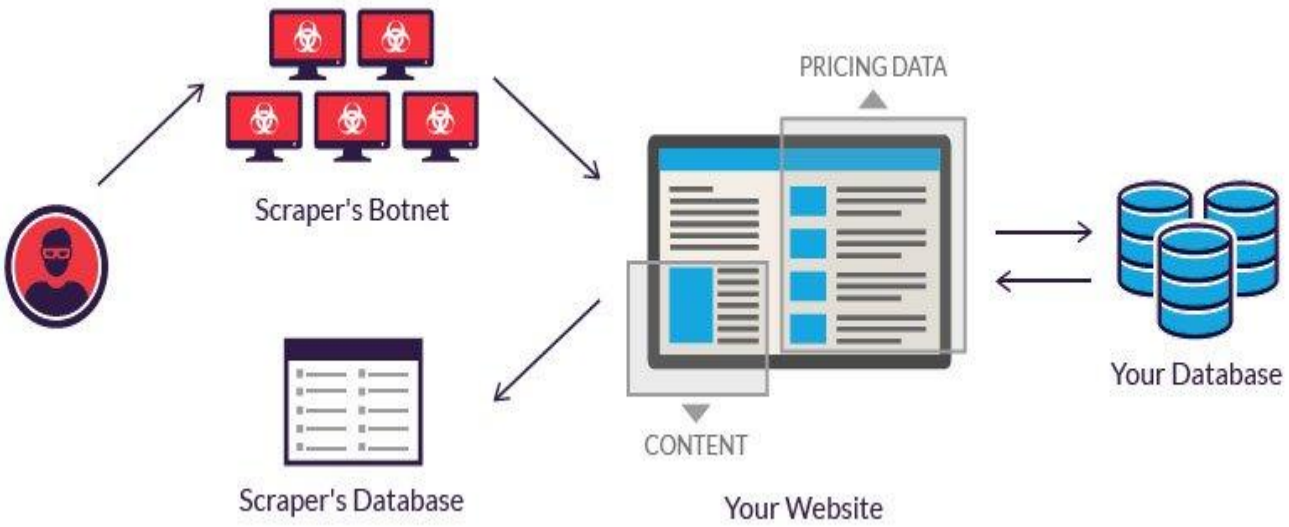
Neden veri kazıyoruz?

-Kazımak demek veri ihtiyacımız var demek

-Veriye erişimde sorun yaşıyoruz demek

-Veriyi doğrulamaya ihtiyacınız var demek ve doğru bilgiyi kazımaya ihtiyacınız var demek.

Sitelerden veri kazıma maliyetli



- Yazılımcının ciddi zamanını alır kazıma
- Düzenli olarak maliyet yaratır
- Ve kazıma düzenli bakıma ihtiyacı vardır

Web kazımanın iyi tarafları

Analiz ve görselleştirme için kullanılabilir
Araştırmaların gelişmesine yardımcı olabilir
Piyasa analizinde ve fiyat karşılaştırmasında kullanılabilir.
Gazeteciliğe yardımcı olur
Hikaye fikirleri geliştirmeyi sağlar
Toplu veriler hakkında kaba bir fikir edinmemize yardımcı olur
Temel veri sistemlerini araştırmak / anlamak için fayda sağlar

Web kazımanın kötü tarafları

İntihal teşvik edebilir

Spam için kullanılır

Kimlik hırsızlığı için kullanılabilir

Stop Web Scraping

STOP CONTENT THEFT AND DOWNTIME FROM WEB SCRAPING BOTS

Tired of watching thieves at work and feeling powerless to stop them? Web scraping bots steal whatever content they've been programmed to fetch – articles, prices, promotions, and API data tha should only be available to legitimate customers or authorized partners.

Distil Networks puts an immediate stop to content theft, competitive data mining, negative SEO attacks, as well as **application denial of service** due to aggressive web scraping.

How to use Terms & Conditions for web scraping protection

Is your website the target of web scrapers? You're not alone! Billion-dollar businesses are built on the back of people like you, by companies that scrape and use your data for all sorts of purposes.

The good news is that you can protect even non-copyrighted content, like prices or customer reviews, via your **website terms and conditions**.

Web Kazımanın Ekonomisi

- Web kazıma yapan şirketlerin% 38'i içerik elde etmek için bunu yapıyor
- Gayrimenkul siteleri bir numaralı web kazıma kurbanı
- Web kazıma hizmetleri saatte 3,33 dolara mal oluyor
- Ortalama web kazıma projesi yaklaşık olarak 135 dolar.
- Ortalama web kazıyıcı yıllık 58.000 $ kazanıyor

Web'de kazımaya yönelik hangi sektörler en fazla korumaya ihtiyaç duyuyor?

- Dijital yayıncılar ve dizinler
- Seyahat
- Emlak
- E-ticaret

- Yasal bir uygulama geliştirin
- Hizmet reddi (DoS) saldırılarını engelle

ETİK TARAFI

Bu verileri alabilir miyim?
Bu verileri yeniden yayınlayabilir miyim?
Web sitesinin sunucularını aşırı yüklüyor muyum?
Bu verileri ne için kullanabilirim?

Flickr user: Greg Emel

Question Asked a year ago

Oliver C. Stringham
7.9 · Rutgers, The State University of New Jersey

Is it legal to use web scraped data for research?

Many websites say in their terms that use of anything (text, pictures, etc) on their site is prohibited because it is their intellectual property. Does anyone know if it is actually illegal or legal to web scrape data from websites to use in research? Do I need to get permission from each individual website I want to scrape? Does the data need to be "anonymous" when published (i.e. someone can't determine which website it came from)?

Research Ethics | Web Scraping

Share

Web kazıma için etik nedir?

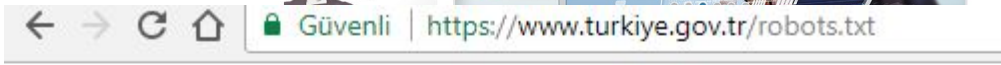
Telif hakkı gibi, kurallar, protokoller ve en iyi uygulamaları sağlama

1. Kimliğinizi belirtiyor musunuz?
2. İndirme aşamalarınızı belgeliyor musunuz?
3. Analiz aşamalarınızı görünür kılıyor musunuz?
4. Araştırmanız tekrarlanabilir mi?
5. Yöntemleriniz ve süreçleriniz gizli mi yoksa özel mi?
6. Kaynaklarınızı nasıl koruyor ve belgeliyorsunuz?
7. Kaynaklarınıza kredi veriyorsunuz?
8. Articles:
    a. *To Scrape or Not to Scrape: Technical and Ethical Challenges of Collecting Data off the Web*
    b. *On the ethics of web scraping from a data journalism perspectiv*e

Etik Veri Kazıyıcı Nasıl Olunur?

- robots.txt dosyasını okuyunuz (varsa eğer)
- Sitenin Terms of Use kısmını okuyunuz (Mevcutsa)
  - Medya alanında çalışan avukata danışınız
- Verileri nasıl koruyacağınızı düşünün
- Sitenin yapısına dikkat edin
- Kuruluşla görüştünüz mü
- siteyi yöneten?Görüşünüz

https://en.cbar.az/robots.txt




```
User-agent: *
Allow: /
Sitemap: https://www.turkiye.gov.tr/sitemap.xml
```


```
User-agent: *
Disallow: /wp-admin/
Allow: /wp-admin/admin-ajax.php
```

https://facebook.com/robots.txt


```
# Notice: Crawling Facebook is prohibited unless you have express written
# permission. See: http://www.facebook.com/apps/site_scraping_tos_terms.php

User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
Disallow: /live/
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photo.php
Disallow: /photos.php
Disallow: /sharer/

User-agent: baiduspider
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
Disallow: /live/
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photo.php
Disallow: /photos.php
Disallow: /sharer/

User-agent: Bingbot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
```

“Kazıma, web siteleri tarafından belirlenen tüm kurallara riayet edildiği sürece ve kazıma verileri iyi niyetlerle kullanıldıkça, veri kazıma etiktir.”

source

Kaynak/Kitap/Proje

http://www.verigazeteciligi.com/veri-gazeteciligi-ve-web-kazimanin-etigi-uzerine/
http://www.garethjames.net/a-guide-to-web-scrapping-tools/
https://onlinejournalismblog.com/2013/09/18/ethics-in-data-journalism-mass-data-gathering-scraping-foi-and-deception/
http://datajournalismhandbook.org/1.0/en/getting_data_3.html
https://docs.google.com/presentation/d/1IC18k19qejGHLCuEtV_2yY0hbbbS48kVoSuYKZZHezQ/edit#slide=id.g10d63abbc3_1_14
https://www.promptcloud.com/blog/is-data-scraping-ethical
https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01
https://www.avvg.org.tr/konular/40-ders-5-pdfden-ve-webden-veri-kazima-.html
http://robertorocha.info/on-the-ethics-of-web-scraping/
http://www.udu.co/blog/what-is-web-scraping-is-it-ethical
https://www.targetinternet.com/what-is-data-scraping-and-how-can-you-use-it/
http://www.verigazeteciligi.com/veri-gazeteciligi-ve-web-kazimanin-etigi-uzerine/

Source

Kaynak/Kitap/Proje

- Catalog of Web Scraping tools
  https://docs.google.com/spreadsheets/d/1A_9wBEmc8VP6HUm2R-7jdPU9FWY8dSfNsfljQ6cz638/edit#gid=0

Tools used or demonstrated in this Workshop

- Webscraper.io - http://webscraper.io/ (script and crawl a website)
- OpenRefine - http://openrefine.org (scripting, cleaning, parsing, utility data tool)
  - Regular Expressions
  - GREL
  - Jsoup
- TAGS - https://tags.hawksey.info/ (Twitter Stream collector)
- Splunk - Academic Program (Twitter Stream collector / Search and Discover)

Ethics

- To Scrape or Not to Scrape: Technical and Ethical Challenges of Collecting Data off the Web
- http://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/