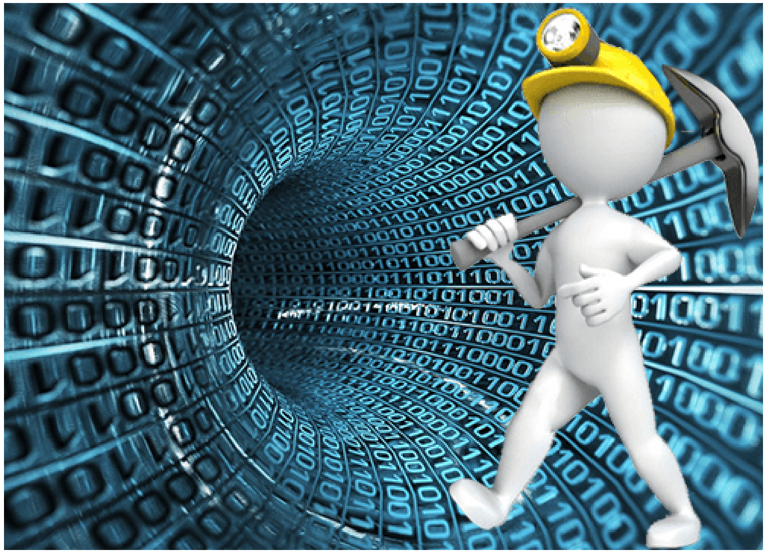
Practical Applications of Web Scraping Discussion

Discussant - Mike Gerling

Contributors: Katherine Vande Pol and Chad Garber



Research and Development

Four Diverse Web Scraping Applications

- **National Agricultural Statistics Service**
  - Industrial Hemp Growers
    - New commodity to be surveyed
- **RTI International**
  - Quickly determine substance use/abuse and rapidly deploy the appropriate type and level of treatment
    - Reddit (social media)
- **U.S. Census Bureau**
  - State and local tax revenue
    - Leveraging other government agencies’ websites
- **National Center for Education Statistics**
  - Mask mandates in schools (Covid-19) and the effect on students’ learning

Endless Number of Applications

- **Stock market (hedge funds)**
  - Examining tweets to determine market sentiment
- **Start-up companies**
  - Find a niche market
- **Impact and extent of national disasters**
  - Food supply
  - Overall economy

Takeaways & Common Themes Across Presentations

- **Define what data are to be collected**
  - Development of sophisticated web-scraping tools and analysis tools to obtain the right information

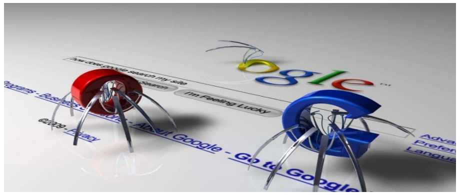
- **Data formatting**
  - Tools were developed to convert collected information into the proper format for the database layout

- **Human intervention**
  - Still required for complex scenarios/records

Takeaways & Common Themes Across Presentations

- **Utilization of outside expertise (contractors) to assist with these web-scraping projects**
- **Reduce respondent burden & costs**
  - Obtain the information another way
  - Replace/supplement a survey
- **Timeliness**
  - Obtain information faster than preparing and conducting a field study

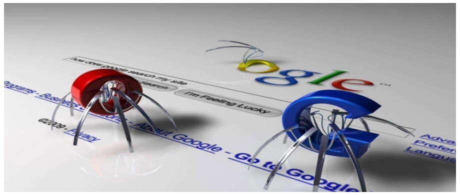
Discussion Points



- **Does coverage of the web-scraped population differ from the population of interest?**
  - Coverage & Representativeness
    - Demographic differences?
    - Social media - are the most vocal swaying the sentiment?

- **How to measure the quality of the data compared to surveys or other traditional data sources?**
  - How current is the web-scraped data?
    - Source and the date of the source
  - Quality will likely depend on the subject matter

Additional Discussion Points



- **Completeness of the data?**
  - Address, phone, email, etc.
  - Does the data available answer the questions of interest?

- **When should not web scraping be utilized?**
  - Relevance (out of scope)
  - Accuracy (errors in posting/scraping)
  - Recency (age of data)
  - Completeness/bias (missing a portion of the population)
  - Interpretability (applicable/practical)

Additional Discussion Points

- **In-house or contract out?**
  - One time or repetitive process?
  - Staffing
    - Availability of in-house expertise?
      - Subject matter
      - Web crawling/scraping (programming)
  - Scale of the project
    - 1,000 records or 100,000 records, etc.
  - System resources (hardware requirements)
    - Virtual machines and IP addresses

- **How does one update a web-scraped list over time?**

Future of Web Crawling/Scraping

- Software tools will become more accessible for non-programmers

- Improved automation of data organization and data cleaning

- Super-intelligent scraping software (artificial intelligence) to handle complex scrapes

Future of Web Crawling/Scraping

- **Legalities**
  - Privacy concerns of personal information
  - Copyrighted and trade secret information
  - Increasing legislation and court rulings

- **Advancements in websites to detect and block scraping attempts**
  - Data for purchase will increase
    - Internally scraped by the website's owner for resale

Remember There is a Wealth of Data Out There

- 5.47 billion active internet users
  - 4.32 billion people use their mobile devices to go online
- 1.98 billion websites online
  - 198.4 million active websites on the web
- 7 million blog posts get published per day
- 4.2 billion active social media users
- 500 hours of video are uploaded to YouTube every minute
- 26 smart objects are located near every human on earth

First Site Guide - 2022

Articles & Publications

- 3 Predictions about Future of Web Scraping. (2022). ProWebScraper. https://prowebscraper.com/blog/future-of-web-scraping/.

- Blazquez, D., J. Domenech, J. A. Gil, and A. Pont. (2019). Monitoring e-Commerce Adoption from Online Data. Knowledge Information Systems, 60, 227–245. DOI: https://doi.org/10.1007/s10115-018-1233-7.

- Chand, M. (2019). How Much Data On The Internet. C#Corner. https://www.c-sharpcorner.com/article/how-much-data-is-on-the-internet/.

- Boettcher, I. (2016). Quality Control of Web-Scraped and Transaction Data (Scanner Data). European Conference on Quality in Official Statistics (Q2016). Madrid, Spain. https://www.ine.es/q2016/docs/q2016Final00139.pdf.

Articles & Publications

- Minaev, A. (2022). Internet Statistics 2022: Facts You Need To Know. First Site Guide. https://firstsiteguide.com/internet-stats/.

- Pai, S. (2021). Scraping vs API: What's the Best Way to Extract Data. Datahut. https://www.blog.datahut.co/post/web-scraping-vs-api#:~:text=While%20web%20scraping%20gives%20you,is%20available%20on%20a%20website.

- What Does the Future of Data Scraping Hold?. (2021). Datahen. https://www.datahen.com/blog/what-does-the-future-of-data-scraping-hold/.

- Young, L., M. Hyman, and B. Rater. (2018). Exploring a Big Data Approach to Building a List Frame for Urban Agriculture: A Pilot Study in the City of Baltimore. Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 323–340. DOI: http://dx.doi.org/10.2478/JOS-2018-0015.

Mike Gerling	michael.gerling@usda.gov	202-692-0277
Katherine Vande Pol	katherine.vandepol@usda.gov	217-493-2999
Samuel Garber	samuel.garber@usda.gov	202-692-0284

Discussion Points

- How to measure the quality of the data compared to surveys or other traditional data sources?
- When should not web scraping be utilized?
- How does one update a web-scraped list over time?
- How to determine if the coverage of the web-scraped population differs/aligns with the population of interest?