Web scraping considerations

Data Science in a Box

datasciencebox.org

Ethics

"Can you?" vs "Should you?"

Researchers just released profile data on 70,000 OkCupid users without permission

By Brian Resnick | @B_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

A group of researchers has released a data set on nearly 70,000 users of the online dating site OkCupid. The data dump breaks the cardinal rule of social science research ethics: It took identifiable personal data without permission.

The information — while publicly available to OkCupid users — was collected by Danish researchers who never contacted OkCupid or its clientele about using it.

The data, collected from November 2014 to March 2015, includes user names, ages, gender, religion, and personality traits, as well as answers to the personal questions the site asks to help match potential mates. The users hail from a few dozen countries around the world.

The data dump did not reveal anyone's real name. But it's entirely possible to use clues from a user's location, demographics, and OkCupid user name to determine their identity.

> If your OkC username is one you've used anywhere else, I now know your sexual preferences & kinks, your answers to thousands of questions.
>
> — Scott B. Weingart (@scott_bot) May 11, 2016

Source: Brian Resnick, Researchers just released profile data on 70,000 OkCupid users without permission, Vox.

"Can you?" vs "Should you?"

Emil OW Kirkegaard @KirkegaardEmil · May 8
The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) openpsych.net/forum/showthre...

26 38

Ethan Jewett @esjewett · May 11
@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

3 9

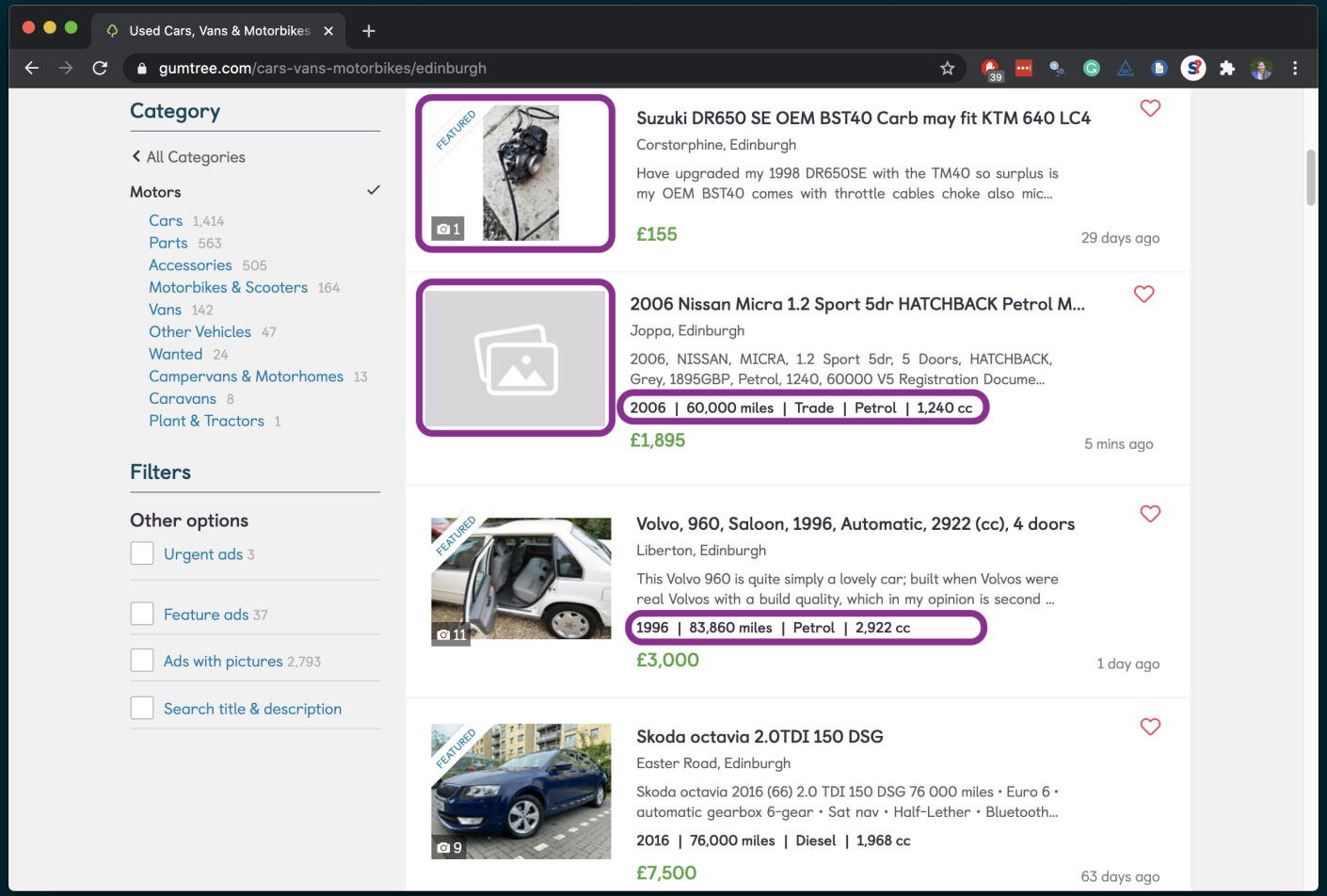
Emil OW Kirkegaard
@KirkegaardEmil

Follow

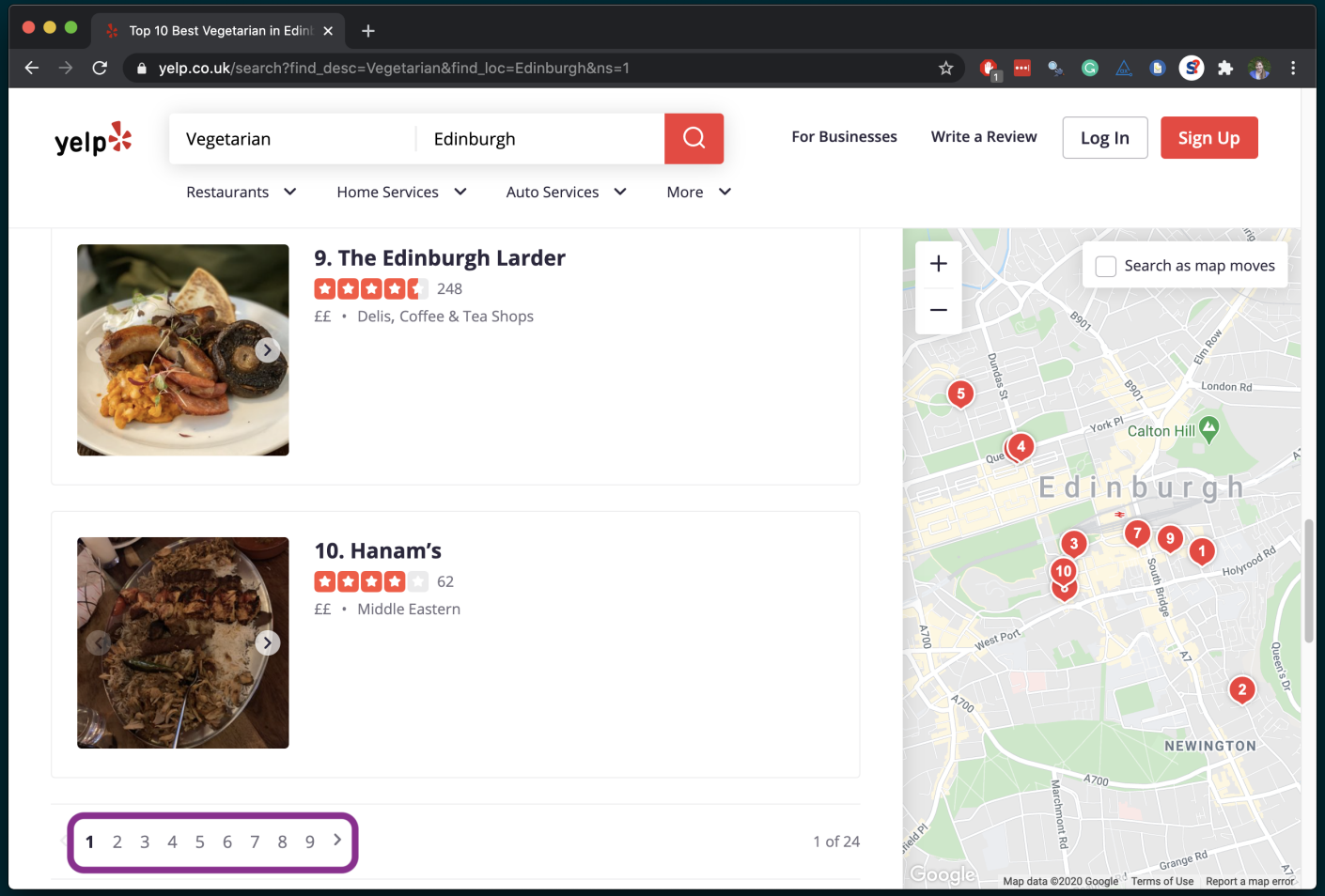
@esjewett No. Data is already public.

Challenges

Unreliable formatting at the source



Data broken into many pages



Workflow

Screen scraping vs. APIs

Two different scenarios for web scraping:

- Screen scraping: extract data from source code of website, with html parser (easy) or regular expression matching (less easy)
- Web APIs (application programming interface): website offers a set of structured http requests that return JSON or XML files

A new R workflow

- When working in an R Markdown document, your analysis is re-run each time you knit
- If web scraping in an R Markdown document, you'd be re-scraping the data each time you knit, which is undesirable (and not *nice*)!
- An alternative workflow:
  - Use an R script to save your code
  - Saving interim data scraped using the code in the script as CSV or RDS files
  - Use the saved data in your analysis in your R Markdown document