Webscraping in Economics: Examples and Advice

Nathan Schiff
Shanghai University of Finance and Economics

Graduate Urban Economics, Week 14
May 22, 2017

Administration

Next class rescheduled to May 31st, same time

I will send out an email with location (hopefully same classroom); we will read Couture and Handbury paper

Following class is last class, would like to have grad students present (proposals or ongoing work)

Restaurant Variety in US Cities

Urban Economics literature suggests consumption amenities (consumer benefits: shops, restaurants, museums, concerts, etc...) attract people to live in cities (Glaeser, Kolko, Saiz 2001)

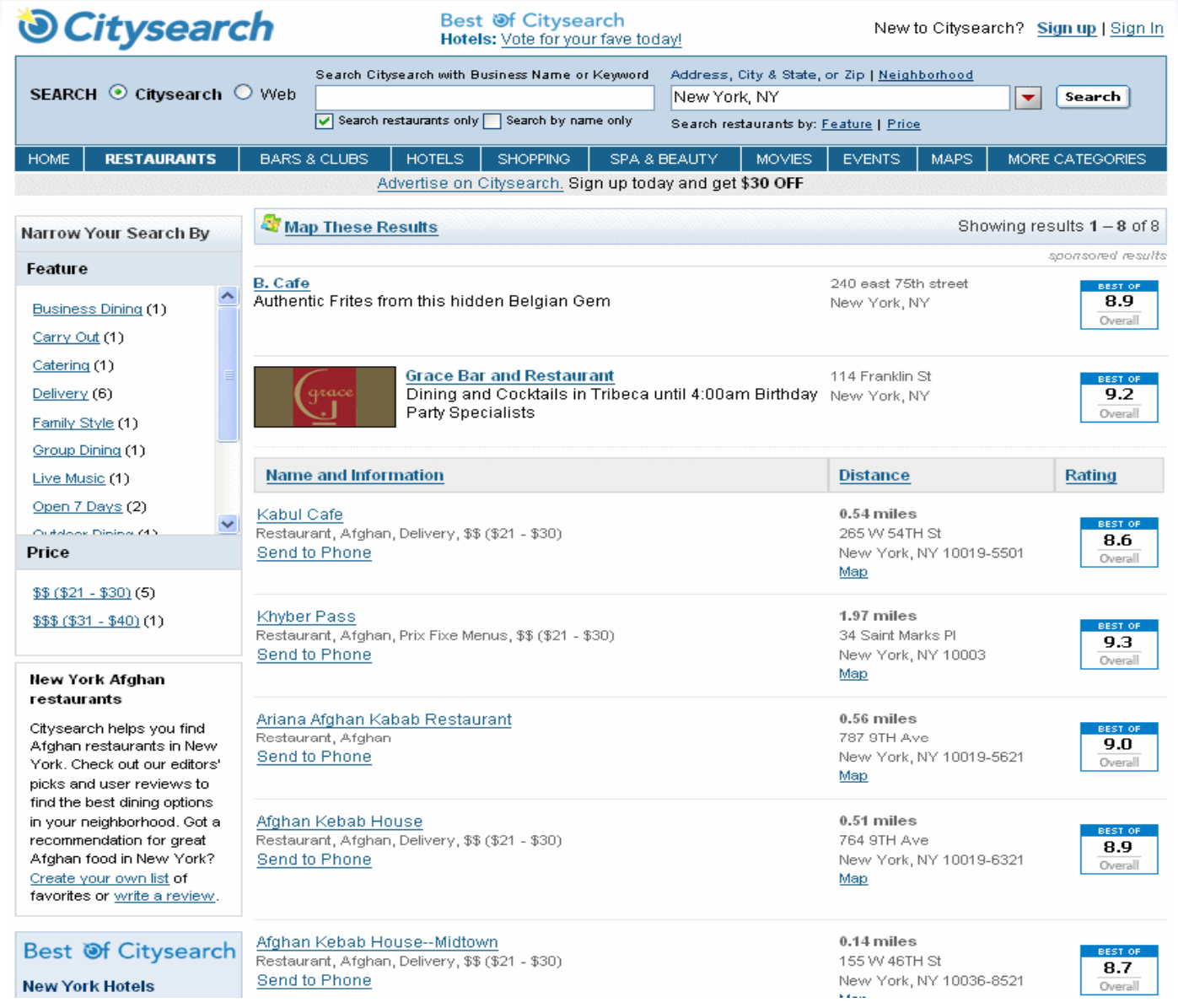
Some evidence that this force is becoming more important, helps to explain recent revival of US downtowns (Couture and Handbury 2016)

But how can we measure this?

I decided to try and measure consumption variety using city restaurant variety

No gov't data on this but lots of online restaurant directories/portals

Scraped citysearch.com to show relationship between city population, density, and product variety (Schiff 2015)


Citysearch
Best of Citysearch
Hotels: Vote for your fave today!
New to Citysearch? Sign up | Sign In
SEARCH Citysearch Web
Search Citysearch with Business Name or Keyword
Address, City & State, or Zip | Neighborhood
New York, NY
Search
Search restaurants only
Search by name only
Search restaurants by: Feature | Price
HOME RESTAURANTS BARS & CLUBS HOTELS SHOPPING SPA & BEAUTY MOVIES EVENTS MAPS MORE CATEGORIES
Advertise on Citysearch. Sign up today and get $30 OFF
Narrow Your Search By
Feature
Business Dining (1)
Carry Out (1)
Catering (1)
Delivery (6)
Family Style (1)
Group Dining (1)
Live Music (1)
Open 7 Days (2)
Price
$$ ($21 - $30) (5)
$$$ ($31 - $40) (1)
New York Afghan restaurants
Citysearch helps you find Afghan restaurants in New York. Check out our editors' picks and user reviews to find the best dining options in your neighborhood. Got a recommendation for great Afghan food in New York? Create your own list of favorites or write a review.
Best of Citysearch
New York Hotels
Map These Results
Showing results 1 – 8 of 8
sponsored results
B. Cafe
Authentic Frites from this hidden Belgian Gem
240 east 75th street
New York, NY
BEST OF 8.9 Overall
Grace Bar and Restaurant
Dining and Cocktails in Tribeca until 4:00am Birthday Party Specialists
114 Franklin St
New York, NY
BEST OF 9.2 Overall
Name and Information
Distance
Rating
Kabul Cafe
Restaurant, Afghan, Delivery, $$ ($21 - $30)
Send to Phone
0.54 miles
265 W 54TH St
New York, NY 10019-5501
Map
BEST OF 8.6 Overall
Khyber Pass
Restaurant, Afghan, Prix Fixe Menus, $$ ($21 - $30)
Send to Phone
1.97 miles
34 Saint Marks Pl
New York, NY 10003
Map
BEST OF 9.3 Overall
Ariana Afghan Kabab Restaurant
Restaurant, Afghan
Send to Phone
0.56 miles
787 9TH Ave
New York, NY 10019-5621
Map
BEST OF 9.0 Overall
Afghan Kebab House
Restaurant, Afghan, Delivery, $$ ($21 - $30)
Send to Phone
0.51 miles
764 9TH Ave
New York, NY 10019-6321
Map
BEST OF 8.9 Overall
Afghan Kebab House--Midtown
Restaurant, Afghan, Delivery, $$ ($21 - $30)
Send to Phone
0.14 miles
155 W 46TH St
New York, NY 10036-8521
BEST OF 8.7 Overall

Some More Noteworthy Examples

1. Davis and Dingell (2016): use Yelp to look racial segregation in consumption (do different races consume different things?)
2. Cavallo and Rigobon (2015): "Billion Prices Project" collects prices from online retailers to look at macro price changing issues; also Cavallo (2015) "Scraped Data and Sticky Prices"
3. Halket and Pignatti (2015): scrape Craigslist to better understand US rental market
4. Many papers on eBay, some on Alibaba
5. Edleman, B. "Using Internet Data for Economic Research", (JEP 2012): useful discussion of many issues

Two (Rough) Uses of Webscraping

Find data that does not exist elsewhere

1. Data is generated by contributions from large user base (Yelp, Craigslist, eBay)
2. Data is itself just measurement of activity on website (forums like Reddit) or network (Facebook, LinkedIn, WeiBo)

Re-arrange data in more convenient form (less common)

1. Data from many sources aggregated on one site (ex: Wikipedia)
2. Parsing techniques of webscraping can also be used when data provider gives you data in inefficient form (ex: 1000 spreadsheets)

Scraped Data and Text Analysis

Current project (with Jacob Cosman) uses restaurant menu data to better understand price responses to new competition

Restaurant menus are just text–we have to use statistical techniques to measure differences in restaurant menus and item changes

We use a CS technique (ngram processing) to measure similarity of distribution of clusters of letters in two menus

These techniques are time consuming but not as daunting as you might expect; techniques already exist and often implemented in Python or R packages

A Simple Example

Say you were working on a project about government and education networks

You want the university of every member of congress, for many congresses (115th, 114th, ...)

This information is available on Wikipedia because each member of congress has a page which usually lists the alma mater (crowd-sourced, many contributors)

In other words, the information exists but in inconvenient form; a good scraping exercise

https://en.wikipedia.org/wiki/108th_United_States_Congress

Tools and Tips

Tools

- Python: I like the Anaconda distribution of Python, includes many built-in packages useful for data science; I use the BeautifulSoup HTML parser (bs4)
- Browser: Chrome or Firefox both have good tools for inspecting the HTML code of any webpage
- Selenium: web-browser emulator, can be used to scrape dynamic webpages (data displayed with javascript) but with considerable more difficulty

Tips

- Space out website scrapes: put automatic pauses between requests to website server, helps website and decreases chance they block your IP. Sometimes better to download pages, then parse.
- Set up test code: especially important if you do scheduled scrapes–you need to know when the website changes their format

Project Considerations

1. What is the exact source of the data being displayed? Does it truly measure what you want? Ex: menupages.com has been problematic
2. If the data is user-contributed, who are the users? Is selection bias going to be a big problem? (ex: which houses are rented on craigslist, who lists on eBay, who posts on social media?)
3. Does the site customize the data based on characteristics of the browser (location of IP address, time of day, frequency of visits, etc...)? Can you deal with this (fake cookies)?
4. Do you need a panel? If the website changes or is pulled down, could you write a paper with just a few periods?
5. How much measurement error can you tolerate in your research design? Ex: if you assemble a panel and a few observations per period fail due to impartial scrapes, is this manageable?

Website Data Considerations

1. Would the site be willing to give you the data or partner with you? Do they have a "chief economist" or API?
2. Is the data formatted in a regular and systematic way? Ex: Yelp is great, Craiglist OK, Reddit is chaotic
3. Are the URLs set up in a systematic way–can you figure out the addresses of all pages you need to scrape?
4. Is the data displayed in HTML or generated by a script (e.g., Javascript)? Is some of the data in image form?

Would it be faster (and more reliable) to do this manually (or hire a squad of undergrads)?

Good reference on automation consideration: Shapiro and Gentzkow, "Code and Data for the Social Sciences: A Practitioner's Guide" (2014)