Web Crawler: yelp

Nick Halliwell, Aina López, Yaroslav Marchuk

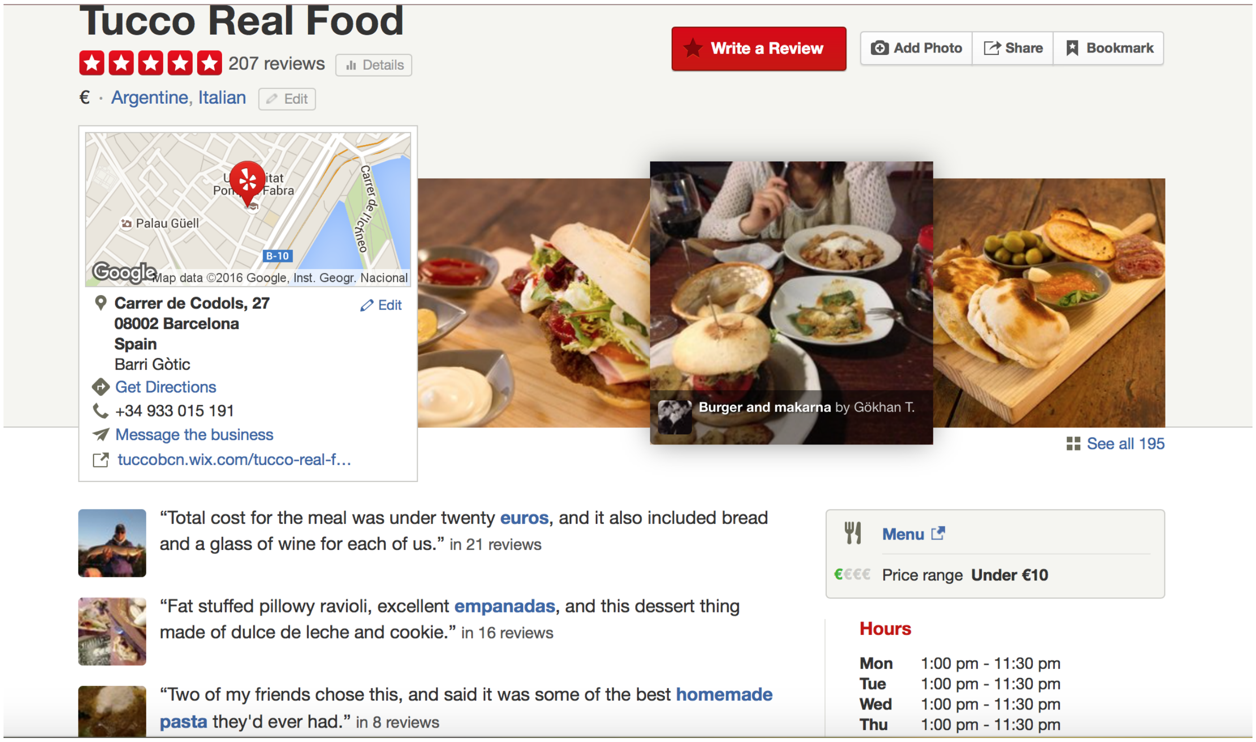
15 May 2016

We decided to scrape data from yelp.com. Yelp is a website where people can write reviews and recommendations and rate the restaurants of a city.

For this project, we scraped the reviews of all the restaurants in Barcelona, which included:

- Text
- Score
- Information of the user:
  - Number of friends
  - Number of reviews
  - Nacionality
- Details of the restaurant:
  - Type of food
  - Price category

And we stored each review in a *.json* file.



Tucco Real Food

207 reviews

€ · Argentine, Italian

Carrer de Codols, 27
08002 Barcelona
Spain
Barri Gòtic

+34 933 015 191

tuccobcn.wix.com/tucco-real-f...

Burger and makarna by Gökhan T.

"Total cost for the meal was under twenty euros, and it also included bread and a glass of wine for each of us." in 21 reviews

"Fat stuffed pillowy ravioli, excellent empanadas, and this dessert thing made of dulce de leche and cookie." in 16 reviews

"Two of my friends chose this, and said it was some of the best homemade pasta they'd ever had." in 8 reviews

Price range Under €10

Hours

Mon	1:00 pm - 11:30 pm
Tue	1:00 pm - 11:30 pm
Wed	1:00 pm - 11:30 pm
Thu	1:00 pm - 11:30 pm

Figure 1. Yelp ScreenShot

1. Instructions

2. How it works?

Save all the URLs of the restaurants you want to scrape in a file. For each URL, do:

1. Get the html code.
2. Each review has a link attributed. Get all the urls of the reviews of that restaurant.
3. Store the characteristics of the restaurant in variables: the type and the price category of the restaurant.
4. For each review (or link found in step 2): get the html code of that page and store the data: text, score, user name, nacionality, number of friends and number of reviews. After each iteration, stop some seconds.

Write a *.json* file for each of the reviews. If some of the previous steps fails, write the URL of the restaurant in a file called *rejected.txt*.

Finally, after each iteration, stop a random amount of seconds.

3. Dealing with errors

How is our code robust?, How do we deal with unexpected failures? What Logs have you mantained?

4. What features have you implemented specifially to tadhere to the robots.txt file of your target website?