

Web scarping the restaurants of Copenhagen

Group 20

Exam in Social Data Science, Økonomisk Institut, Københavns Universitet

Sofie Dalum,

Naja Algreen Suhr, GRZ240

Rasmus Habel Svärd,

Michael Fogh Emcken,

Number of characters

## Abstract

People have different opinions when it comes to food. Flavours, ingredients, setting of the place your eating and price. Everything around a meal, places a role. This can be seen on the reviews from TripAdvisor, that rates restaurants all over the world, based on user’s experience.  
In this project, we

First we eat, then we do everything else. – M.F.K. Fisher

## Introduction to TripAdvisor and TripAdvisor Denmark

TripAdvisor Inc is a website company in the industry of travel services. The multiplatform company enables users to book hotels, vacation rentals, flights, find restaurants and activities, and brands itself as a travel guide. The platform also serves as a user review site of all on-site products, available in 28 languages and on 49 markets, according to the website.[[1]](#footnote-1) For this project we have limited the scope to the Danish subpage ‘tripadvisor.dk’ where the user interface and review language is Danish.

Users of TripAdvisor have the opportunity of finding, booking and reviewing products of the site, such as hotels or restaurants. The company revenue comes from commercial advertising, such as display based, click-based and subscription-based advertisement and from booking fees when users book through the website.[[2]](#footnote-2)

All registered users are able to review an element on TripAdvisor, but the reviews are not verified. Users can review as many or as few as they like, and the user can choose how much information alongside the ratings they want to disclose. Users are invited to review bookings they have made on TripAdvisor after the end of the booking, and some companies may ask their customers to review them online, sometimes motivated by participation in a contest or drawing of prices, if the customers submits a review. Because of this behaviour concerning TripAdvisor reviews cannot be trusted solemnly in general to be a true evaluation of a customer’s opinion.

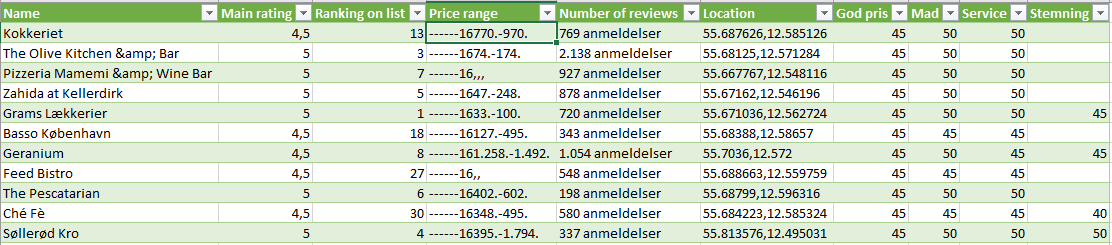
An example of this is the restaurant in London called ‘The Shed at Dulwich’ which at one point the highest rated restaurant on TripAdvisor in London. The problem was though that the restaurant didn’t exist and had never existed. The high rating came exclusively from fake reviews of family and friends, and after only 92 great fake review the restaurant entered first place.[[3]](#footnote-3) From this story it is to be learned that is it rather easy to manipulate TripAdvisor rating and that TripAdvisor as a platform doesn’t validate reviews or ratings. This is an important point to have in mind when moving forward using TripAdvisor data. We have chosen to use TripAdvisor data despite is disadvantage, because of its multiplatform nature and overall volume of Copenhagen based restaurants.

## Data generation and gathering

The data for our analysis has been retrieved from TripAdvisor.dk on dd-mm-yy. After choosing TripAdvisor as our area of interest we build a web scraping script in Python. The full scraping process has been logged with a script made by Snorre Ralund, provided in class. The script delays the requests of TripAdvisor by 0.5 seconds and logs the scraping in a text file for documentation and process control.

The scraping process of TripAdvisor consists of several parts. When searching for restaurants on TripAdvisor in Copenhagen the result is a number of overview pages containing the individual restaurants, as seen below.



The search result for restaurants in Copenhagen contains 78 overview pages. The script scrapes the overview pages for links on the individual restaurant pages and after removing duplicates, the script creates a csv file, ‘individual\_urls.csv’, with all individual restaurant links. The next layer of the scraper loops through ‘individual\_urls.csv’ loads all the restaurants individual data. The individual data is as follows: name, location, price range, price class, number of reviews, ranking, rating on value for money, food quality, service, atmosphere and overall rating. The individual data is here seen as an example.

REPLACE WITH FIANL DATA FRAME

The scraping script is attached as a Jupyter Notebook called “TripAdvisor\_scraper.ipynb”. This script creates **X** CSV files that contains the overview links (overview\_urls.csv) the individual restaurant links (individuals\_urls.csv) **and one with all the data for analysis (data.csv)**.

The HTML of TripAdvisor is structured in a way that makes the site somewhat easy to scrape their data. All links are absolute and stable links, and the data on the individual restaurant is stored the same way for all restaurants on site. Not all restaurants have the same data available on the individual pages, but the HTML structure is still preserved with values then being hidden. We have noted that data is missing on a limited number of restaurants, but after looking into it, we have concluded that it is because of the restaurants having a few or no reviews at all.

Before beginning the process of building the web scraper script we discovered that TripAdvisor has an API with restricted access. One can apply for a key to the content API, but the key is not granted for the use of content on data analysis, academic research and B2C application. Overall the API wouldn’t be effective for us in this project as the limitation on the API is 1000 calls per day, which means the data collecting process would have been stretched over several days. Furthermore, the time scale on the processing of the application on a key is unknown, but possible longer than this project. We therefore chose to build the web scraping script.

## Ethics of data collection and gathering:

If we had chosen to apply for the API key and had been successful in the scale of the project, we would have broken the terms and conditions when analysing the data. We could have deceived our purpose in our application and thus gotten access, but chose not to, as this would have been ethically wrong.

By scraping any web page under Danish jurisdiction is crucial to reflect on ethics and hacking, to avoid breaking the law. The scarping process we have scripted only collects data otherwise available when using the web page. All data collected from TripAdvisor is visible when clicking on and around the web page. The script is build using open source tools that are able to read HTML in a legal manner. Most, maybe even all, restaurants have website with the corresponding information on TripAdvisor, why we could have collected the data from the individual private web pages of the restaurants if needed, though we chose not to. Restaurants are public companies with an interest of sharing basic information, such as telephone number, addresses and other data, such as menu and pricing, why most restaurants have sample menus with pricing available online. Companies such as restaurants often benefits from being online as it creates more online reach and in the end more revenue. We find that scraping public restaurant data is ethically sound.

We have chosen to scrape only aggregated data, why we have only collected means and bundled data of restaurant reviews and therefore have no scarped data of individual physical or online persons, such as usernames, profile photo or other data. If we had chosen to scrape data of individual persons, we would have had to discuss the application use and the ethics thereof further. By limiting our self to only public restaurant data from TripAdvisor we have concluded that our scraping and analysis is ethically sound.

## Data processing

After extracting the data from TripAdvisor, the data needed processing and cleaning before being ready for analysis and modelling.

### Interactive Folium Map

The Python Folium package has been used for showing distribution of restaurants on different variables in Copenhagen in a simple map visualisation format. The Folium package has the benefit of being interactive, so it is possible to zoom in and out while using Jupyter Notebook. On map 1 each restaurant in the dataset has its own point based on its coordinates when zoomed in, whereas when zooming out the restaurants are gradually more and more clustered. The clusters and the individual points are coloured by rankings of the restaurants. E.g. a point of an individual restaurant will show the colour of its specific ranking grouped by two hundred. E.g. a cluster of restaurants will be coloured after the most dominant ranking group also in groups of two hundred. Restaurants with a ranking under two hundred will have the strongest green colour, whereas restaurants with a ranking between 200 and 400 will have a slightly lighter version of green. This pattern follows through from green over yellow and in the end red when rank is between two thousand and two thousand and two hundred.

## Analysis

## Conclusion

## Web pages

TripAdvisor:

Abouts us: “om os”, <https://tripadvisor.mediaroom.com/us-about-us>, 22.08.2019

Main page in English: [www.tripadvisor.com](http://www.tripadvisor.com), 22.08.2019-30.08.2019

Main page in Danish: [www.tripadvisor.dk](http://www.tripadvisor.dk), 22.08.2019-30.08.2019

Developer page about the TripAdvisor Api, <https://developer-tripadvisor.com/content-api/>, 22.08.2019

Vator:

Article about the business model of TripAdvisor <https://vator.tv/news/2018-04-13-how-does-tripadvisor-make-money>, 23.08.2019

Independent:

The Shed at Dulwich: <https://www.independent.co.uk/life-style/food-and-drink/the-shed-at-dulwich-was-london-s-top-rated-restaurant-just-one-problem-it-didn-t-exist-a8107791.html>, 27.08.2019

1. TripAdvisor, about us: ”om os” [↑](#footnote-ref-1)
2. Vator TV: “How does TripAdvisor make money?” [↑](#footnote-ref-2)
3. Independent: ”The Shed at Dulwich” [↑](#footnote-ref-3)