2nd Lab Project - Web Scraping

Web Data Mining - SS 2022

This is a hands-on lab project on Web scraping with Selenium[1]. You will practice how to extract raw textual data from websites, and how to preprocess and analyze them to unveil new knowledge. Each group (same as 1st lab project) will experiment on a task that contains TripAdvisor data, Python and Selenium. **The group will have to prepare a consistent, detailed hands-on tutorial as a Medium[2] article**, focused on the general topic of "Analysis of TripAdvisor Data with the help of Selenium and Python"; no further report will be required. An indicative example of a Medium tutorial article can be found here or here.

The tutorial will have the following sections:

A. Installing and setting up Selenium
B. Presentation of the TripAdvisor dataset requirements (what you need to extract)
C. TripAdvisor dataset web scraping process (how to extract and store it)
D. Data preprocessing for textual TripAdvisor data (how to clean your data)
E. Your final DataFrame format (how does your processed data look like, i.e., which are your columns and their contents)
F. Answers to the questions of Section "Data Analysis" in Python (how to analyze your data)
G. Insightful conclusions (what you've learnt)

Don't forget to:

- Document the process and **your conclusions** in your Medium-like article
- Share your code with Jupyter Notebook
- Upload/send your solutions until **02/05/2022**

Topics

Tripadvisor, Inc. is an American online travel company that offers online hotel reservations, bookings, and reviews for transportation, lodging, travel experiences, and restaurants. For this project -for educational purposes only-, **you are requested to scrape TripAdvisor data for local businesses and landmarks in Thessaloniki**. Specifically:

Table 1: Team allocations

Team	Topic
[103 - 88 -101 - 100]	Hotels

[1] https://pypi.org/project/selenium/
[2] https://medium.com

[108 -119 - 113 - 106]	Restaurants
[111 - 97 - 95 - 117]	Things to Do
[74 - 93 - 98]	Cafes & Bars
[123 - 116 - 91 - 107]	Restaurants

Attention

- For scraping cafes and bars only, on filtering options, select “Establishment Type” Bars & Pubs **AND** Coffee & Tea
- For scraping restaurants only, on filtering options, select “Establishment Type” Restaurants, and then start scraping at section “Top Restaurant in Thessaloniki” (scroll down)

Data Collection

Utilizing Selenium for web scraping, for each business that appears in the starting page given above, you should follow the process below:

1. Enter the business profile page
2. Go to the reviews section
3. Scrape data for all reviews (including pagination) as indicated below.
4. Store data in a convenient format for later analysis (either in a database, e.g., MongoDB, or in a simple file, e.g., CSV)

For each review, you should scrape the following data (Figure 1):

1. Reviewer username
2. Business reviewed
3. Review date (above the review text)
4. Visit data (below the review text)
5. Review title
6. Review text
7. Review rating

Selenium gives you the ability to interact with the webpage. Specifically, to extract more data about the reviewer, **for each review** you should click on the reviewer’s profile picture (automatically with Selenium) to extract the following information (if available) from the pop-up that appears (Figure 2):

8. Reviewer’s age
9. Reviewer’s gender
10. Reviewer’s location
11. Reviewer’s review distribution

Attention

- Long reviews do not appear fully in the reviews page. Hence, you should click on “Show more” (automatically with Selenium) to make the full text appear (Figure 3).
- For teams working on “Things to Do” or “Hotels”, clicking on the reviewer’s profile picture in the business reviews’ page will only redirect you to the reviewer’s profile (unwanted behavior). Instead, you want to click on the review’s title and open up a new tab with the review’s content (wanted behavior). Now you can click on the user’s profile picture to get demographic information (if available).
- Some reviewers do not have full profiles, namely age, gender, location, or all of the above, which may be missing from the pop-up. In this case, store the review data (without information about the reviewer) and proceed to the next review. Expect only 10% of the reviews to contain demographic information.

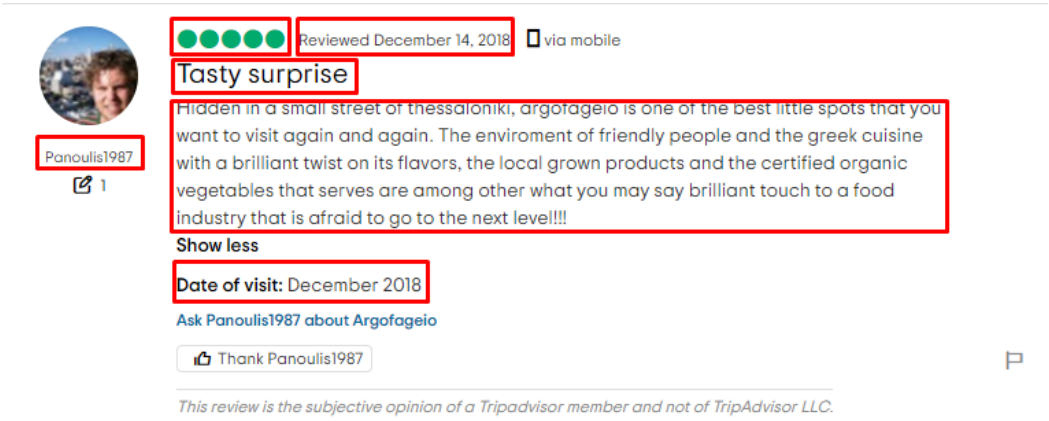

Figure 1: An indicative example of a review for a business on TripAdvisor.

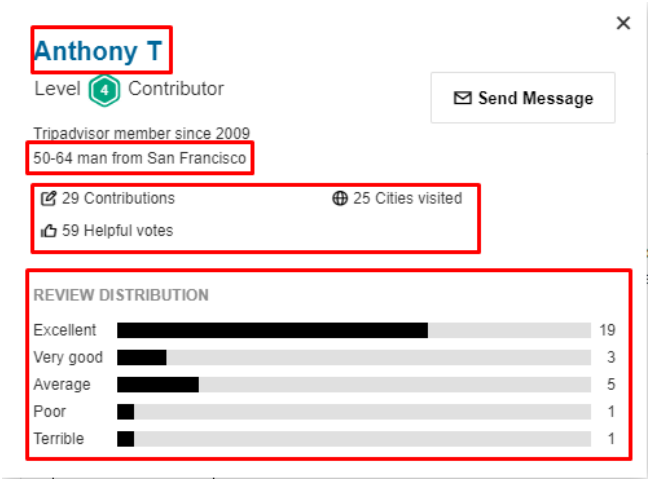

Figure 2: Once you click on the reviewer’s profile picture, a pop-up similar to this will appear.

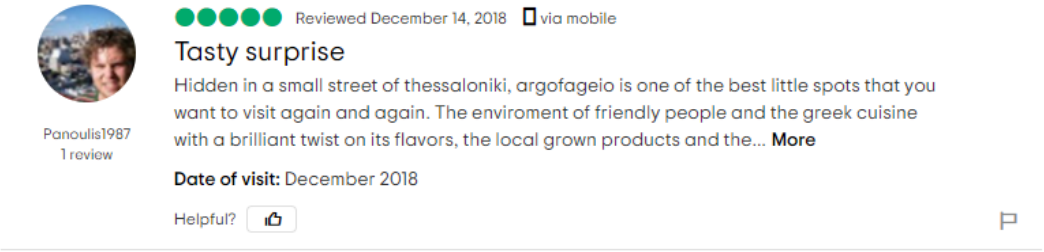

Figure 3: An indicative example of a review that needs to be expanded with “Show more”

Data Analysis

Once you collect all data, you are ready for the data analysis part. For the whole category you are analyzing, e.g., restaurants, you will create basic visualizations, and you will perform basic sentiment analysis and user profiling on aggregated user reviews. You should not visualize results referring to individual reviews or users.

Attention

Apply all necessary preprocessing steps, e.g., stop-word removal, lowercase conversion, tokenization, stemming, etc., according to your intuition and experience. Refine the preprocessing steps as needed while you proceed with your analysis. Do not forget to document this process in the Medium-like article deliverable.

Basic Visualizations

In this step, you will create some interesting visualizations for getting a better understanding of your data.

Table 2: Requested basic visualizations

#	Question
1	Visualize the number of monthly reviews over time. Which was the month with the most reviews? Is there any seasonality in the volume of reviews?
2	Visualize the most common words, bi-grams, and tri-grams across all reviews through a bar chart or word cloud. Also, visualize the most common words, bi-grams, and tri-grams in 5-star versus 1-star reviews.
3	Which are the 10 fastest growing and the 10 fastest shrinking words (based on usage frequency) in TripAdvisor reviews over time?
4	Explore and visualize emerging topics from all the user reviews across time, similar to here. Which topics do you identify? Do they make sense to you?

5	Visualize in a map the locations of the reviewers based on the collected data with color coding based on the location frequency. Prefer countries over cities to identify more than one occurrence.

Sentiment Analysis

In this step, you will calculate and visualize the expressed sentiment in user reviews. For the purpose of this assignment a text-based approach utilizing SenticNet[3] or TextBlob[4] is sufficient, but you are free to experiment with Machine Learning (ML) approaches if you are interested.

Attention
A bonus will be given for ML-based sentiment analysis, but it is by no means necessary, and it will not affect your grade negatively if you choose to omit it.

Table 3: Requested visualizations for sentiment analysis

#	Question
1	Calculate the sentiment or polarity score (negative versus positive) for each review and then print the top-5 positive and the top-5 negative reviews.
2	Find the most common words in positive reviews and the most common words in negative reviews and visualize them through a bar chart or tag cloud.
3	Find the sentiment or polarity score (negative versus positive) for each business/landmark and then print the top-10 places with positive and negative scores, respectively. Do they overlap with the businesses/landmarks with the highest and lowest scores in your category?

User Profiling

In this step, you will build an ML model for user profiling for an attribute of your choice (reviewer gender or reviewer age or both). To achieve this, you will only use the subset of the reviews with information about the reviewer's demographics (Figure 2).

As a label, you will use the reviewer's age or gender. Your features will include a variety of factors, such as review text, user rating distribution, sentiment features (calculated in the previous step), and linguistic features (e.g., use of exclamation, emojis, etc.) that can be related to gender or age. Be creative during feature engineering! In the end, you need to evaluate the performance of your model and if acceptable, you can use it to predict the demographics of the remaining users.

[3] https://sentic.net/
[4] https://textblob.readthedocs.io/en/dev/index.html

Tip

If the subset of reviews is not large enough to train an ML model, you might want to scrape a few of the reviewer profiles (the ones with demographics) to collect all their other reviews to use for training purposes. The URL to a reviewer's profile (to scrape) is https://www.tripadvisor.com/Profile/[USERNAME]. For example: https://www.tripadvisor.com/Profile/Colmogorman

Table 4: Requested visualizations for user profiling

#	**Question**
1	Calculate the percentage of each class (e.g., male vs. female or age groups) in your training data users and in the whole dataset (after you use your model for prediction).
2	Visualize the most common words per class (e.g., male vs. female or age groups) and sentiment/polarity (negative vs. positive) in a bar chart or word cloud.
3	Create a box plot of user ratings visualizing the differences in ratings between user classes. Also, create a similar box plot for sentiment/polarity scores.