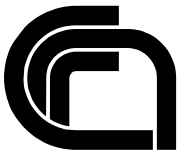


Consiglio Nazionale delle Ricerche

ZARS: a Software to extract information from Online Travel Agencies

A. Minissale, A. Marchetti, A. Lo Duca

IIT TR-09/2020

Technical Report

Maggio 2020

Istituto di Informatica e Telematica

ZARS: a Software to extract information from Online Travel Agencies

Alessia Minissale*, Andrea Marchetti, Angelica Lo Duca****

* Università di Pisa
a.minissale@studenti.unipi.it

** Istituto di Informatica e Telematica – Consiglio Nazionale delle Ricerche
andrea.marchetti@iit.cnr.it
angelica.loduca@iit.cnr.it

General index

1. Introduction 2

1.1 Context 2

1.2 Justification 2

1.3 Software overview 3

2. Audience 3

3. System requirements 4

4. Description of the software 5

4.1 Objective 5

4.2 Architecture 5

4.3. License 7

5. Using the software 7

5.1 Installation 7

5.2 Running the software 8

5.2.1. Hotel_Info.py 8

5.2.2. Hotel_Facilities.py 9

5.2.3. Hotel_Amenities.py 10

5.2.4. Hotel_Reviews.py 11

5.3 Test 12

5.3.1. Execution time 12

5.4 Software limitations 12

6. Conclusions 13

Abstract

Zars is a software completely written in Python for the extraction of information relating to the accommodation facilities of a given city, passed as an argument, from a tourist portal called OTA (Online Travel agencies). With regard to each accommodation facility, Zars allows you to extract its description, services and associated reviews. The results obtained are returned in the form of a SQL database. Zars is released under the GNU General Public License v 3.0.

Keywords

Data Collection, Web Scraping, Tourism, Online Travel Agencies

ACM Classification

D.2.13: Reusable Software

1. Introduction

1.1 Context

The software was implemented and developed during a curricular internship carried out at the University of Pisa (FILELI Department - Digital Humanities) in collaboration with Prof. Andrea Marchetti and Prof.ssa Angelica Lo Duca of the IIT (Institute of Informatics and Telematics) of the CNR of Pisa. The project is part of the Data Collection field of study, making use of automated data extraction techniques and using the database as a collection tool.

1.2 Justification

The collection process arises from the need to collect data relating to accommodation facilities and the goal is to acquire quality evidence and information that allow a subsequent analysis of the data with the aim of carrying out statistical surveys. The so-called OTAs are the main sources for travelers to leave and consult reviews, comments and feedback on hotels, tour operators and destinations in general. The growing importance of tourists, in the process of co-creating the tourist experience and the analysis of the experiences they have experienced, has also increased the importance of Big Data for a more effective and efficient process of destination management. Their use allows the management of the offer of tourist products and services more efficient, precisely because

a subsequent analysis of the data will allow the personalization of the goods and services according to the different types of demand.

1.3 Software overview

The goal pursued by the Zars software implementation[1] is to automate a data collection process for testing purposes by using a tool implemented as a browser driver. A remote-control interface was used that allows you to manipulate DOM elements in web pages and to control the behavior of user programs. The interface, in this work, has been implemented in the Google Chrome platform. The website on which the software was tested is the Trip Advisor tourism portal[2], the largest travel site in the world, with 260 million visitors every month, and more than 150 million reviews and 4 million offers of accommodation. The proposal is to search for the destination of interest and create a database that contains all the data relating to the user's needs. Four tables can be viewed in the database. These will contain the information of all the accommodation facilities of the city being researched: complete address, score assigned based on the position (to be understood in relation to the center and the attractions / restaurants within 500 meters), the services they offer and the reviews written by travelers following their stay. Having in possession the extracted data, it will be possible to carry out exploration and analysis of the data, in order to discover and study significant regularities, make predictions on the behavior of individuals and try to motivate their choices. It is important to remember that the software implemented can also be used with other OTA search engines, provided that the appropriate changes are made (e.g. the Xpath of the elements on the web page will be different)**.**

2. Audience

The intent is to be able to extend the use of the software to any type of user, from beginners to experts. By following the guidelines and downloading the necessary tools, you can easily access the execution of the software. The following are two types of users for whom it is intended:

[1] https://github.com/alessiamns/ZARS
[2] www.tripadvisor.com

- Accommodation: the software can be useful when you intend to improve the services offered according to the preferences of the traveler.
- Tourism sector operators: the software has been studied and implemented for data collection and, above all, to study and predict, through subsequent data analysis, the customer's behavior and interest when choosing the structure in which to stay.

3. System requirements

- Python 3 (https://www.python.org/downloads/)
- Xampp (https://www.apachefriends.org/download.html)
- Chrome Driver (https://chromedriver.chromium.org/)

Package Name	**Link**	**License**
Selenium	https://www.selenium.dev/downloads/	Free
MySQL connector	https://www.mysql.com/it/products/connector/	Free
Configparser	https://docs.python.org/3/library /configparser.html	Free
Argparse	https://docs.python.org/3/library/argparse.html	Free
Sys	https://docs.python.org/3/library/sys.html	Free
Time	https: // docs. python.org/3/library/time.html	Free
Re	https://docs.python.org/3/library/re.html	Free

4. Description of the software

4.1 Objective

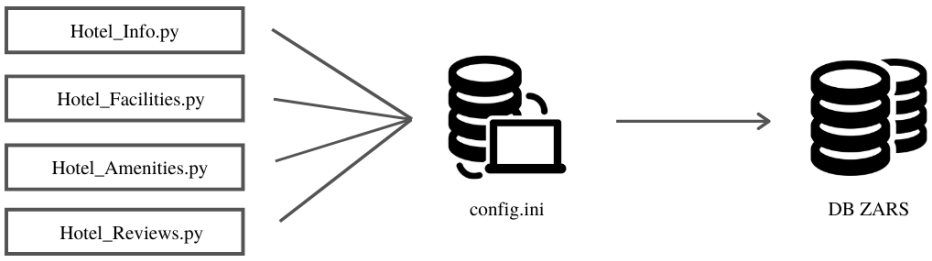
The objective of the Zars software is to extract information on accommodations of a given date cities by a specific search engine (tourist portal): it is, in fact, contact information, evaluation, services offered by the structures and reviews written by travelers.

This goal is achieved by implementing four different modules and inserting the relevant data into a database.

This is the procedure followed by the software:

1) connection to the OTA site via simulation of a browser (using Selenium)
2) simulation of typing the keyword relating to the city in the search engine text box and subsequent click (Key Enter)
3) extraction of information from the pages of the results
4) storage on database

4.2 Architecture



The software consists of four files with the extension *.py*, each of which is launched from the terminal to perform a specific task, and a configuration file to allow connection to the database:

- **Hotel_Info.py**: extracts the main information on accommodation facilities;
- **Hotel_Facilities.py**: extracts the score with which the structure is assessed (from 0 to 100) and the number of restaurants and attractions within 500 meters;
- **Hotel_Amenities.py**: extracts the services offered by the structures;
- **Hotel_Reviews.py**: extracts the reviews written by customers;

- **Config.ini**: configuration file containing the setting for the connection to the database and for the waiting time during program execution.

The program outputs a single database that will contain the following tables:

- Info table

#	Nome	Tipo
1	Name	varchar(64)
2	City	varchar(64)
3	Address	varchar(512)
4	Url	varchar(512)
5	Rating	float(2,1)
6	Review_Count	varchar(64)
7	Popular_Index	varchar(64)

address

1. Name
2. City
3. Full
4. Url
5. Overall rating
6. Number of reviews
7. Popular index

- Facilities table

#	Nome	Tipo
1	Name	varchar(64)
2	City	varchar(64)
3	Great_to_walkers	int(11)
4	Restaurants_500m	int(11)
5	Attractions_500m	int(11)

1. Name
2. City
3. Score from 0 to 100 on walking comfort
4. Number of restaurants within 0.5 km
5. Number of attractions within 0.5 km

- Amenities table

#	Nome	Tipo
1	Name	varchar(64)
2	City	varchar(64)
3	Amenity	varchar(64)

offered by the facility

1. Name
2. City
3. Service

- Reviews table

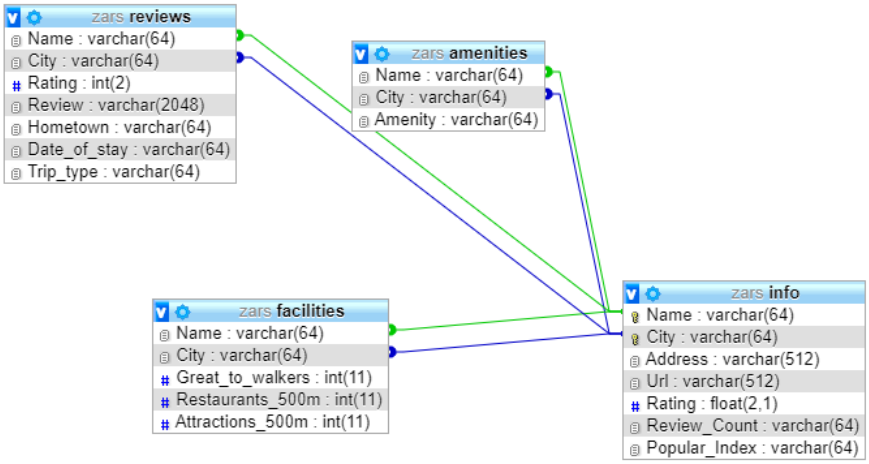
#	Nome	Tipo
1	Name	varchar(64)
2	City	varchar(64)
3	Rating	int(2)
4	Review	varchar(2048)
5	Hometown	varchar(64)
6	Date_of_stay	varchar(64)
7	Trip_type	varchar(64)

1. Name
2. City

3. Rating review
4. Text review
5. Origin of traveler
6. Date of stay
7. Type of travel

In the tables generated the primary key (Name, City) of the Info table is indicated with the symbol of a golden key, while, in the other tables, the external key is represented by a silver colored key which refers to the primary key.

The relationships between the tables are shown below:



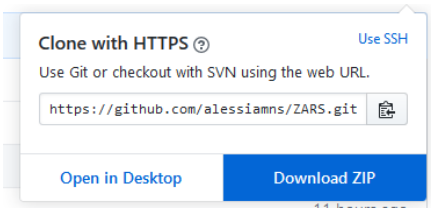
4.3. License

The software is released under the GNU General Public License v3.0[3]. The GNU General Public License is a free copyleft license for free software. It guarantees end users such as organizations, businesses or simple individuals to use, share and even modify the software.

5. Using the software

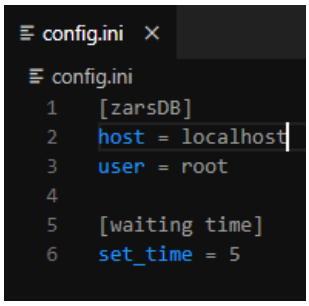
5.1 Installation

- Download zip from GitHub (https://github.com/alessiamns/ZARS)



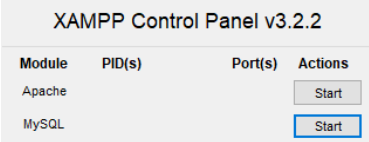
[3] https://www.gnu.org/licenses/gpl-3.0.html

- Config.ini file configuration: the configuration settings for the connection to the database and the initialization of the value of the function are entered inside the file, `time.sleep ()` which is initially set at 5 seconds; however, it will be possible to increase the value if the execution of the script is not successful, so as to allow the machine to wait longer before carrying out the test and recognizing the elements within the automated web page.

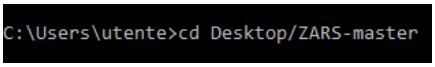


5.2 Running the software

- Extract the folder from the zip archive: this will contain the four files in Python and the configuration file
- Access the Xampp control panel: press *Start* for Apache and MySQL



- Open command prompt and go to the directory where the folder is saved (for example: Desktop)



The Hotel_Info.py file must be executed as the first script, which contains the so-called PRIMARY KEY. Consequently, the other scripts will contain the FOREIGN KEY associated with the corresponding value in the Info table, which will generate a link between the tables.

5.2.1. Hotel_Info.py

Executing the command:

```
python Hotel_Info.py -place [place_name] -pages [num_pages]
```

- Input:
    - `-place` **(mandatory)**: name of the city to be searched;

- `-pages` **(optional)**: number of pages on which to search - if omitted, the software will scroll through all the pages containing the accommodation facilities of the chosen location.

Example:

```
C:\Users\utente\Desktop\ZARS-master>python Hotel_Info.py -place Noto -pages 1
```

- Output:
  - the script populates the table *Info*.

Example:

Name	City	Address	Url	Rating	Review_Count	Popular_Index
AgriResidence Valle Degli Dei	Noto	Contrada Fiumare, 96017, Noto, Sicilia Italia	https://www.tripadvisor.it/Hotel_Review-g652022-d6...	4.5	159 recensioni	N. 18 di 38 hotel a Noto
Agriturismo A' Lumia	Noto	Contrada Madonna Marina, 96017, Noto, Sicilia Ital...	https://www.tripadvisor.it/Hotel_Review-g652022-d1...	4.5	155 recensioni	N. 9 di 152 altri alloggi a Noto
Agriturismo Saccollino	Noto	Strada Provinciale 51 C.da Saccollino, 96017, Noto...	https://www.tripadvisor.it/Hotel_Review-g652022-d2...	4.0	98 recensioni	N. 20 di 152 altri alloggi a Noto
Alma Suite	Noto	Via Pietro Micca 17, 96017, Noto, Sicilia Italia	https://www.tripadvisor.it/Hotel_Review-g652022-d1...	NULL		
Antica Masseria La Fiumara	Noto	Contrada Fiumara Snc, 96017, Noto, Sicilia Italia	https://www.tripadvisor.it/Hotel_Review-g652022-d1...	NULL		
B&B Casa de Rollo	Noto	Via Rocco Pirri 55, 96017, Noto, Sicilia Italia	https://www.tripadvisor.it/Hotel_Review-g652022-d1...	4.5	18 recensioni	N. 90 di 158 Bed & Breakfast a Noto
B&B Casa Giunchiglia	Noto	Vico Giunchiglia 2, 96017, Noto, Sicilia Italia	https://www.tripadvisor.it/Hotel_Review-g652022-d8...	5.0	155 recensioni	N. 4 di 158 Bed & Breakfast a Noto
B&B Novecento Siciliano	Noto	Via Silvio Spaventa 2, 96017, Noto, Sicilia Italia	https://www.tripadvisor.it/Hotel_Review-g652022-d8...	5.0	41 recensioni	N. 36 di 158 Bed & Breakfast a Noto

5.2.2. Hotel_Facilities.py

Executing the command:

`python Hotel_Facilities.py -place [place_name] -pages [num_pages]`

- Input:
  - `-place` **(mandatory)**: name of the city to be searched;
  - `-pages` **(optional)**: number of pages on which to search - if omitted, the software will scroll through all the pages containing the accommodation facilities of the chosen location.

Example:

```
C:\Users\utente\Desktop\ZARS-master>python Hotel_Facilities.py -place Noto -pages 1
```

- Output:
  - the script populates the table *Facilities*.

Example:

Name ▲ 1	City ▲ 2	Great_to_walkers	Restaurants_500m	Attractions_500m
AgriResidence Valle Degli Dei	Noto	*NULL*	*NULL*	*NULL*
Agriturismo A' Lumia	Noto	*NULL*	*NULL*	*NULL*
Agriturismo Saccollino	Noto	*NULL*	*NULL*	*NULL*
Alma Suite	Noto	96	79	38
Antica Masseria La Fiumara	Noto	*NULL*	*NULL*	*NULL*
B&B Casa de Rollo	Noto	96	104	43
B&B Casa Giunchiglia	Noto	96	75	34
B&B Novecento Siciliano	Noto	96	106	42
B&B Teatro	Noto	96	97	41
B&B Villa Ambra	Noto	81	11	5

5.2.3. Hotel_Amenities.py

Executing the command:

```
python Hotel_Amenities.py -place [place_name] -pages [num_pages]
```

- Input:
    - `-place` **(mandatory)**: name of the city to be searched;
    - `-pages` **(optional)**: number of pages on which to search - if omitted, the software will scroll through all the pages containing the accommodation facilities of the chosen location.

Example:

```
C:\Users\utente\Desktop\ZARS-master>python Hotel_Amenities.py -place Noto -pages 1
```

- Output:
    - the script populates the table *Amenities*.

Example:

Name ▲ 1	City ▲ 2	Amenity
AgriResidence Valle Degli Dei	Noto	Piscina
AgriResidence Valle Degli Dei	Noto	Internet ad alta velocità gratuito (WiFi)
AgriResidence Valle Degli Dei	Noto	Hotel per non fumatori
AgriResidence Valle Degli Dei	Noto	Animali domestici ammessi
AgriResidence Valle Degli Dei	Noto	Parcheggio gratuito
AgriResidence Valle Degli Dei	Noto	Vasca idromassaggio
AgriResidence Valle Degli Dei	Noto	Lavanderia self-service
AgriResidence Valle Degli Dei	Noto	Attività per bambini/famiglie
AgriResidence Valle Degli Dei	Noto	Servizio lavanderia
AgriResidence Valle Degli Dei	Noto	Piscina all'aperto
AgriResidence Valle Degli Dei	Noto	Wi-Fi
Agriturismo A' Lumia	Noto	Concierge
Agriturismo A' Lumia	Noto	Colazione gratuita
Agriturismo A' Lumia	Noto	Lavaggio a secco
Agriturismo A' Lumia	Noto	Animali domestici ammessi
Agriturismo A' Lumia	Noto	Hotel per non fumatori
Agriturismo A' Lumia	Noto	Parcheggio gratuito
Agriturismo A' Lumia	Noto	Internet ad alta velocità gratuito (WiFi)

5.2.4. Hotel_Reviews.py

Executing the command:

```
python Hotel_Reviews.py -place [place_name] -pages [num_pages] -pr [num_pages Tab]
```

- Input:
    - `-place` **(mandatory)**: name of the city to be searched;
    - `-pages` **(optional)**: number of pages on which to search - if omitted, the software will scroll through all the pages containing the accommodation facilities of the chosen location.
    - `-pr` **(optional)**: number of pages containing reviews - if omitted, the software will scroll through all the pages containing the reviews of the accommodation.

Example:

```
C:\Users\utente\Desktop\ZARS-master>python Hotel_Reviews.py -place Noto -pr 2 -pages 1
```

- Output:
    - the script populates the table *Reviews*.

Example:

Name ▲ 1	City ▲ 2	Rating	Review	Hometown	Date_of_stay	Trip_type
AgriResidence Valle Degli Dei	Noto	4	durante il ns tour abbiamo soggiornato per una not...	Milano, Italia	dicembre 2019	
AgriResidence Valle Degli Dei	Noto	4	Abbiamo scelto questo posto insieme a mio marito p...		settembre 2019	Ha viaggiato con la famiglia
AgriResidence Valle Degli Dei	Noto	5	Incantevole, Rilassante, Verde... tre aggettivi Pe...		settembre 2019	Ha viaggiato con la famiglia
AgriResidence Valle Degli Dei	Noto	5	Siamo stati 5 notti all agri resort valle degli de...	Palermo, Italia	settembre 2019	Ha viaggiato in coppia
AgriResidence Valle Degli Dei	Noto	5	e siamo qui per il tredicesimo anno ormai diventat...	Sicilia, Italia	settembre 2019	
AgriResidence Valle Degli Dei	Noto	3	Il posto si trova in una incantevole vallata, imme...	palermo	agosto 2019	
AgriResidence Valle Degli Dei	Noto	5	Polecam w 100% super miejsce bajeczne?????????ś...		agosto 2019	Ha viaggiato con la famiglia
AgriResidence Valle Degli Dei	Noto	5	Wat n geweldige plek om vakantie te houden. Zo rel...		luglio 2019	Ha viaggiato in coppia
AgriResidence Valle Degli Dei	Noto	5	La struttura è in una posizione fantastica immersa...	Marche, Italia	giugno 2019	Ha viaggiato in coppia
AgriResidence Valle Degli Dei	Noto	5	Przepiękna okolica do której napewno wrócimy.Apart...		maggio 2019	Ha viaggiato in coppia
Agriturismo A' Lumia	Noto	5	Ottima posizione, pergolato carinissimo con vista ...	Cardano al Campo, Italia	agosto 2019	Ha viaggiato in coppia
Agriturismo A' Lumia	Noto	5	Abbiamo trovato quasi per caso, la sig.ra Laura ci...	Torino, Italia	agosto 2019	Ha viaggiato con la famiglia
Agriturismo A' Lumia	Noto	5	Siamo stati in questo agriturismo a luglio e abbia...	Stoccolma, Svezia	luglio 2019	
Agriturismo A' Lumia	Noto	5	Un angolo di pace ritagliato tra le campagne a poc...	Sant'Elpidio a Mare, Italia	luglio 2019	
Agriturismo A' Lumia	Noto	4	Entouré de citronniers, d'amandiers, d'orangers et...	Le Mesnil-Aubry, Francia	giugno 2019	Ha viaggiato con amici
Agriturismo A' Lumia	Noto	3	L'agriturismo si trova in un punto strategico a me...	Catania, Italia	giugno 2019	Ha viaggiato con la famiglia

5.3 Test

The software execution times can vary in relation to various factors such as, for example, the Internet connection speed and the type of machine on which the script is run. In this regard, it was deemed necessary to insert the variable in the config.ini file `set_time` initialized to 5 (seconds): to adapt it to your machine, the value will be editable if necessary.

5.3.1. Execution time

A speed test was performed on a machine with a second-generation Intel Core i5 processor with Windows operating system and 8GB of RAM. Using the method `time.time ()` it was possible to calculate (in seconds) the execution time of each program:

Module	**Execution time (in sec.)**	**Number of records**
Hotel_Info.py -place Noto -pages 1	885.995007276535	30
Hotel_Facilities.py -place Noto -pages 1	746.2978010177612	30
Hotel_Amenities.py -place Noto -pages 1	2848.3869087696075	309
Hotel_Reviews.py -place Noto -pr 2 -pages 1	2465.1091043949127	248

5.4 Software limitations

As regards the limitations attributed to the software, this has some anomalies as for the rather slow execution time: this is mainly related to the exploration of the web page to be tested and the extraction of the data of interest. It is always recommended, however, to clear your browsing history and data before running the script. It is essential to pay attention to the version of the browser used, in order to download the driver that fits the updated version (in this case, it will be sufficient to find the version of Google Chrome from the browser settings and go to search for the compatible driver). Another factor that could affect the execution of the software is, of course, the Ram of the machine on which it is run. As regards, however, the implementation of the software on other OTA search engines, this must be adapted to the website used as a test, since in each web page the

elements may have a different position and it will therefore be necessary to identify the XPath of the site web being tested.

6. Conclusions

Data are also a strategic factor in the construction of the tourism product and therefore the management of these is becoming the future bet for the competitiveness of a territory; knowing how to put one's own potential online, in terms of resources, services offered, and reception capacity, is increasingly becoming a factor of attraction and possible development. This study, conducted precisely for the collection of data on accommodation facilities, making use of advanced data extraction technologies from websites, aims to create a single database that can be considered a starting point for future data analysis and statistical surveys to study and understand the customer's behavior when choosing the structure.