



UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

**Progettazione e sviluppo di una piattaforma low-cost  
per l'acquisizione e l'analisi di dati sul turismo**

**Candidato:** *John Bianchi*

**Relatore:** *Andrea Marchetti*

**Correlatore:** *Paolo Macchia*

Anno Accademico 2017-2018

# Indice

1. Introduzione .....	4
2. Cosa s'intende per "fenomeno turistico" .....	6
2.1 Esigenza di dati riguardanti l'offerta turistica .....	7
2.1.1 Definizione delle unità e zone di analisi.....	7
2.2 Scelta della sorgente dati .....	9
2.2.1 La scelta di TripAdvisor rispetto ad altri siti di viaggi .....	11
2.2.2 Caratteristiche generali del sito web TripAdvisor .....	13
2.2.3 Tipologia dei dati necessari all'indagine .....	14
3. Metodo di raccolta dati .....	15
3.1 Scelta della tipologia di software da utilizzare .....	15
3.1.1 Linguaggio di programmazione Python .....	16
3.1.2 Libreria Selenium .....	17
3.1.3 WebDriver PhantomJS .....	18
3.1.4 WebDriver ChromeDriver .....	19
3.1.5 Salvataggio dati su un database SQL .....	21
3.1.6 Software per la pianificazione di comandi .....	23
3.2 Scelta della tipologia di hardware da utilizzare .....	24
3.2.1 Economia del prodotto in relazione alle caratteristiche .....	25
3.2.2 Potenzialità del sistema operativo Raspbian .....	26
3.2.3 Requisiti di alimentazione .....	27
3.3 Costruzione del dispositivo di web scraping .....	29
3.3.1 Accessori aggiuntivi .....	30
3.3.2 Installazione hardware .....	32
3.3.3 Installazione del sistema operativo .....	33
3.3.4 Installazione del software necessario al progetto .....	35
3.3.5 Progettazione e sviluppo del software di web scraping .....	36
3.3.6 Definizione della durata delle acquisizioni .....	42
3.3.7 Manutenzione del dispositivo durante la durata delle acquisizioni .....	42
4. Analisi dei dati ottenuti .....	44
4.1 Cosa s'intende per tipologia di clienti .....	44

4.1.1 Definizione delle varie tipologie di clienti .....	45
4.1.2 Analisi e considerazioni sui dati ottenuti .....	46
4.1.3 Associazione delle tipologie di clienti alle zone d'interesse .....	47
4.2 Visualizzazione dei dati ottenuti .....	48
4.2.1 Tecniche utilizzate per la visualizzazione geografica delle strutture .....	48
4.2.2 Tecniche utilizzate per la visualizzazione dei dati delle strutture .....	50
5. Conclusioni .....	52
5.1 Sviluppi futuri .....	53
6. Bibliografia .....	54
7. Sitografia .....	54

# 1. Introduzione

I dati presenti nel web sono elementi importanti per l'indagine e lo studio di svariati fenomeni. Grazie alla tecnica di web scraping questi ultimi possono essere raccolti in grandi quantità e in maniera totalmente automatica.

In questa tesi è stato progettato e realizzato un software di web scraping in grado di ottenere dati dal portale web di viaggi TripAdvisor, al fine di condurre un'analisi sulla tipologia di clientela delle strutture alberghiere di quattro comuni specifici: Cascina, Livorno, Lucca e Pietrasanta.

Le analisi sono state eseguite su dati raccolti in un mese a partire dall'11 febbraio 2019 fino all'11 marzo 2019.

Ogni aspetto tecnico della tesi è stato realizzato dalle risorse personali del candidato, pertanto per ovviare al problema della gestione del software di scraping durante la durata delle acquisizioni dati, è stato scelto di utilizzare un single board computer il quale ha reso possibile questo procedimento che di norma imporrebbe costi di alimentazione elevati a causa della lunga durata del processo di raccolta dati.

Queste tipologie di computer di piccola dimensione necessitano di una quantità di energia inferiore a quella di un computer desktop, e hanno una potenza di calcolo sufficiente per eseguire un programma di web scraping.

Nella prima parte del secondo capitolo verrà trattato in modo globale il fenomeno turistico, definendone aree e unità di ricerca; nella seconda, sarà giustificata la scelta dell'utilizzo del sito web di viaggi TripAdvisor elencandone infine le principali caratteristiche.

Il terzo capitolo riguarderà più specificamente la parte tecnica della raccolta dati: sarà elencata la tipologia di software e hardware utilizzato e come essi sono stati configurati per questo scopo.

Nel quarto capitolo verrà affrontato il metodo di trattamento dei dati ottenuti, verranno poi indagate le varie tipologie di clienti in relazione alle quattro differenti aree di ricerca. Nella seconda parte del medesimo capitolo verranno illustrate le tecniche di visualizzazione dei dati impiegate.

Seguirà una conclusione in cui verranno indicati i possibili sviluppi futuri.

## 2. Cosa s'intende per "fenomeno turistico"

L'enciclopedia Treccani definisce il fenomeno turistico come:

*"L'insieme di attività e di servizi a carattere polivalente che si riferiscono al trasferimento temporaneo di persone dalla località di abituale residenza ad altra località per fini di svago, riposo, cultura, curiosità, cura, sport ecc. Il turismo è pertanto trasferimento ciclico: partenza dal domicilio abituale, arrivo ed eventuale soggiorno nella località di destinazione, ritorno alla località di partenza."* (Treccani, sito web)

Negli anni l'importanza di questo fenomeno si è ampliata tanto che esso è divenuto uno dei punti cardine del settore terziario.

Anche a livello politico l'interesse è andato crescendo; a dimostrazione di questa tesi vi è anche la dichiarazione dell'attuale vicepresidente della Commissione Europea Antonio Tajani che lo descrive come:

*"Un motore di innovazione sviluppo sostenibile, di integrazione economica e sociale in regioni rurali, periferiche o in ritardo di crescita e un mezzo per promuovere un'occupazione stabile, consentendo allo stesso tempo di salvaguardare e valorizzare il patrimonio culturale, storico e ambientale."* (Tajani, 2011)

Egli suggerisce un'interpretazione di questo fenomeno come un'opportunità di valorizzazione delle risorse territoriali, nella salvaguardia della storia e dell'economia locali, consentendo la creazione di nuovi posti di lavoro in maniera totalmente sostenibile.

Questa definizione tiene conto dei potenziali effetti che possono essere prodotti da un flusso di persone in ingresso e uscita da un luogo.

## **2.1 Esigenza di dati riguardanti l'offerta turistica**

La possibilità di consultare dati costantemente aggiornati sul fenomeno turistico in una determinata area è un fattore importante per lo sviluppo dell'offerta turistica. Essi rappresentano un solido punto di partenza per lo sviluppo del settore.

Tuttavia, i principali servizi di statistica, come ad esempio l'Istituto Centrale di Statistica oppure il Portale della Regione Toscana, forniscono dati soltanto fino a due anni precedenti al corrente, fattore che va ad incidere in maniera negativa sulla loro fruibilità.

Da ciò nasce dunque l'esigenza di condurre un'indagine che miri all'acquisizione di queste informazioni.

### **2.1.1 Definizione delle unità e zone di analisi**

Per condurre l'indagine svolta da questa tesi, si è scelto di considerare come unità di ricerca le strutture ricettive comunemente disponibili nei principali siti di viaggi online.

Ne fanno quindi parte le seguenti categorie: residenze turistiche alberghiere, alloggi agrituristici, affittacamere, alloggi privati, case e appartamenti per vacanze, case per ferie, residence e residenze d'epoca.

Per quanto riguarda la scelta delle zone di analisi si è deciso per la selezione di quattro Comuni differenti: Lucca, Pietrasanta, Cascina, Livorno.

La scelta è dovuta da un lato alla loro comune appartenenza alla Regione Toscana e dell'altro alle loro molteplici differenze.

Esse sono facilmente deducibili a partire dalla loro posizione nel territorio, basti considerare ad esempio Cascina - situata in una pianura alluvionale adiacente al fiume Arno - a confronto con Pietrasanta, posta ai piedi delle Alpi Apuane al margine della pianura costiera della Versilia, a pochi chilometri dal mare.

Queste differenze sono rese ancora più evidenti dai dati forniti dall'Ufficio Regionale di Statistica riguardanti informazioni dell'anno 2017. Nonostante che al momento non siano disponibili dati più recenti, rimangono ugualmente validi quali oggetto d'indagine.

<b>Comune</b>	<b>Provincia</b>	<b>Provenienza</b>	<b>Arrivi</b>	<b>Presenze</b>	<b>N° di presenze per arrivo</b>
Livorno	Livorno	Italiani	94.330	199.631	2,11
Livorno	Livorno	Stranieri	60.559	129.887	2,14
Lucca	Lucca	Italiani	105.579	191.839	1,81
Lucca	Lucca	Stranieri	135.970	319.338	2,34
Pietrasanta	Lucca	Italiani	68.538	289.606	4,22
Pietrasanta	Lucca	Stranieri	55.588	223.534	4,02
Cascina	Pisa	Italiani	5.919	11.556	1,95
Cascina	Pisa	Stranieri	4.756	8.934	1,87
TOSCANA		Italiani	6.198.386	21.214.573	3,44
TOSCANA		Stranieri	7.573.788	25.215.793	3,32
TOSCANA		Totale	13.772.174	46.430.366	3,37

*Tabella 1. Arrivi e presenze divise per italiani e stranieri delle strutture ricettive per comune - Regione Toscana 2017*

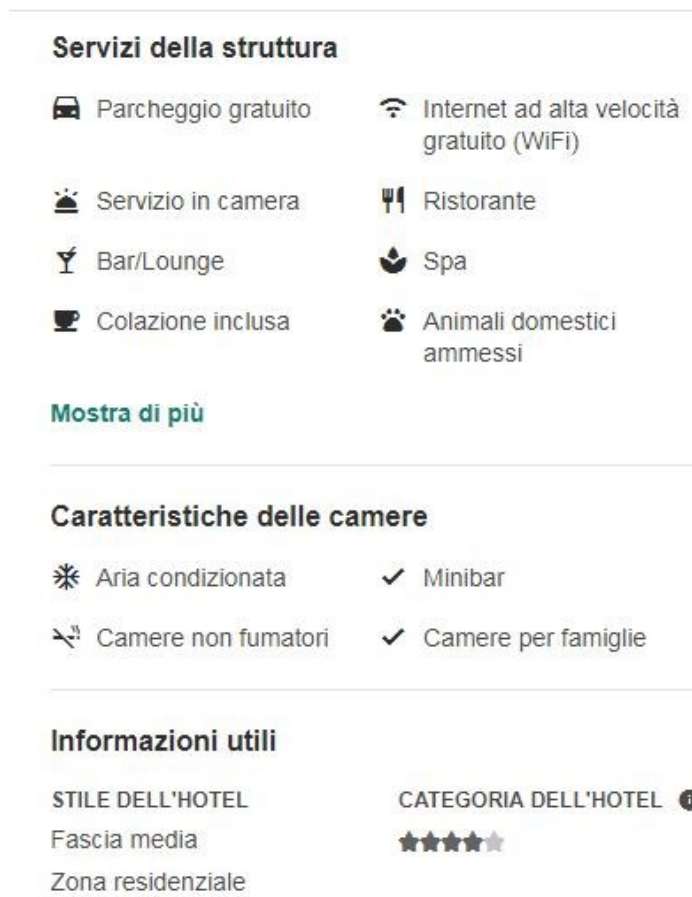


Nella tabella 1 appare chiara la differenza tra i vari comuni, in particolare tra la quello di Livorno e di Lucca, dove, se nella prima il numero di arrivi di italiani è superiore a quello di stranieri, nel secondo la situazione è totalmente invertita. Questo fattore avrà sicuramente un impatto influente sulla tipologia di offerta turistica.

## **2.2 Scelta della sorgente dati**

Oggigiorno la quasi totalità delle strutture alberghiere è registrata su molteplici siti di viaggi. Essi forniscono informazioni non solo riguardanti il costo della prenotazione, ma anche sulle caratteristiche delle strutture stesse come illustrato dalla figura 1.

I dati sono forniti direttamente dai titolari delle strutture i quali scelgono liberamente di registrarsi, fattore che rafforza la veridicità delle informazioni.



*Figura 1. Screenshot ottenuto dal sito web TripAdvisor, elenco servizi e informazioni di una struttura*

Questi siti permettono anche agli utenti di recensire le strutture e di poterle valutare con un giudizio espresso in numeri.

### **2.2.1 La scelta di TripAdvisor rispetto ad altri siti di viaggi**

Per compiere questo tipo di scelta si è deciso di optare per un sito di viaggi che avesse le caratteristiche di un metamatore, ovvero che riuscisse a fornire prezzi da più di un sito di prenotazione come Booking, Agoda, Expedia, Hotels.com.

Infatti se l'indagine fosse stata compiuta su uno di questi ultimi, i prezzi forniti sarebbero stati fortemente condizionati dalle politiche interne di queste società.

Un sito di viaggi che ha funzione di metamatore non permette di completare alcuna transazione di denaro all'interno del suo dominio ma reindirizza l'utente verso il sito che offre il servizio di prenotazione dal quale è riuscito a trovare il prezzo.

A partire dalla ricerca di un utente, il metamatore compie altre ricerche sui vari servizi online di prenotazione, per arrivare a fornire il prezzo più basso disponibile.

La figura 2 è uno screenshot ottenuto direttamente dal sito web di viaggi TripAdvisor il quale, grazie alla sua funzione di metamatore, fornisce prezzi da più servizi di prenotazione.


Ad esempio per quanto riguarda la prima struttura elencata ovvero “B&B Arena di Lucca” il prezzo di 98€ è fornito dal provider Agoda, mentre Booking non fornisce alcun prezzo. Per il secondo risultato, quello riguardante “Best Western Grand Hotel Guinigi”, Booking offre il prezzo migliore, assieme ad Agoda e Hotels.com.

Prezzi più bassi

Arrivo  
dom 07/04/19

Partenza  
lun 08/04/19

Ospiti  
1 camera, 2 adulti, 0 bambini



**B&B Arena di Lucca**

agoda

98 €

Vedi l'offerta

Hotels.com  
Booking.com  
Expedia.it


Vedi tutte le 5 offerte da 98 €

Bed & Breakfast

168 recensioni

N. 8 di 346 alloggi con il miglior rapporto qualità-prezzo a Lucca

Connessione Wi-Fi gratuita  
Colazione inclusa



**Best Western Grand Hotel Guinigi**

Booking.com

124 €  
103 €

Vedi l'offerta

Agoda.com  
Hotels.com  
Lol.travel


Vedi tutte le 8 offerte da 103 €

1.142 recensioni

N. 9 di 346 alloggi con il miglior rapporto qualità-prezzo a Lucca

Connessione Wi-Fi gratuita  
Parcheggio gratuito


Figura 2. Screenshot ottenuto dal sito web TripAdvisor, risultati di ricerca



**L'Iris B&B**

Centro storico, Lucca · Mostra sulla mappa · 700 m dal centro

Super Affare


Camera Tripla - 

Solo 1 camera rimasta sul nostro sito!


Eccellente **9,3**  
427 recensioni  
Posizione **9,6**

€ 80  
include tasse e costi

Vedi disponibilità >

 Anche una casa per le vacanze potrebbe fare al caso tuo.  
Trova il posto perfetto per il tuo viaggio!

Mostra le case vacanze




**Hotel Napoléon** ★★★★★

Lucca · Mostra sulla mappa · 1 km dal centro

La prenotazione può essere effettuata senza carta di credito

Super richiesta! 1 prenotazione per le tue date nelle ultime 24 ore sul nostro sito

Camera Matrimoniale Comfort - 

Solo 4 camere rimaste sul nostro sito!

Zero rischi: puoi cancellare più tardi, quindi assicurati questo ottimo prezzo oggi stesso.

Buono **7,6**  
1.143 recensioni

€ 84  
include tasse e costi  
Colazione inclusa  
Cancellazione GRATUITA

Vedi le nostre ultime camere disponibili >

Figura 3. Screenshot ottenuto dal sito web Booking, risultati di ricerca

La figura 3 invece è uno screenshot ottenuto da un risultato di ricerca proveniente dal servizio di prenotazione Booking. Qui i prezzi elencati sono unicamente quelli dello stesso elemento che come già anticipato, comprometterebbe la veridicità di uno studio sul territorio.

Infine, la preferenza del sito di viaggi TripAdvisor rispetto all'altro principale metamatore Trivago, è stata dettata da motivazioni dal carattere pratico, in quanto le tempistiche dedicate all'indagine non hanno permesso un approccio completo anche su quest'ultimo, motivo per cui è stato deciso di non trattare di esso nella suddetta relazione.

### **2.2.2 Caratteristiche generali del sito web TripAdvisor**

*“TripAdvisor® è il sito di viaggi più grande del mondo\*, che permette ai viaggiatori di sfruttare al massimo il potenziale di ogni viaggio. Con più di 730 milioni di recensioni e opinioni relative alla più grande selezione di business di viaggio a livello mondiale – circa 8.1 milioni di alloggi, compagnie aeree, esperienze e ristoranti – TripAdvisor offre ai viaggiatori le esperienze della community per aiutarli a decidere dove soggiornare, come volare, cosa fare e dove mangiare. TripAdvisor inoltre confronta i prezzi di oltre 200 siti di prenotazione per permettere ai viaggiatori di trovare la tariffa più conveniente dell'hotel adatto a loro. I siti a marchio TripAdvisor sono presenti in 49 mercati e rappresentano la più grande community di viaggiatori del mondo: 490 milioni di visitatori unici al mese” (TripAdvisor, sito web)*

TripAdvisor è, grazie alle sue caratteristiche appena citate, la fonte di dati più completa per condurre l'indagine obiettivo della tesi, come illustrato chiaramente dalla pagina dedicata all'interno del portale stesso.

### **2.2.3 Tipologia dei dati necessari all'indagine**

Oltre alle caratteristiche sopraelencate, il portale fornisce svariate informazioni, ma per il raggiungimento dell'obiettivo dell'indagine, sono stati selezionati dati riguardanti:

- posizione geografica esatta della struttura
- numero di stelle
- indirizzo link della pagina dedicata alla struttura all'interno del sito stesso
- prezzo relativo alla permanenza di una notte per due persone nel giorno odierno all'acquisizione
- provider che fornisce tale prezzo
- numero di recensioni
- media del punteggio delle recensioni
- elenco di tutti i servizi disponibili della struttura

Al fine di ottenere un campione sufficientemente rappresentativo si è scelto, come già anticipato nell'introduzione, di raccogliere questi dati per un lasso di tempo di 30 giorni a partire dall'11 febbraio 2019, eseguendo l'acquisizione dati per due volte al giorno. Questo passo è illustrato nel dettaglio nella sezione 3.3.6.

### **3. Metodo di raccolta dati**

I siti web rappresentano un'ampia risorsa di dati. Tuttavia questi ultimi sono racchiusi all'interno di codici di vario tipo, che rendono ardua la creazione di un software che ne estrapoli solo la parte d'interesse.

*“Alcune tipologie di software riescono a simulare la navigazione umana”* (Webopedia, sito web).

Con essi infatti è possibile interagire con la pagina web in questione, ad esempio premendo pulsanti come quello del cambio della pagina, in modo tale da riuscire a rendere la raccolta dati ancora più efficiente.

Fare un'operazione di questo tipo significa eseguire un compito di “web scraping”.

In questo capitolo saranno illustrati i software e l'hardware utilizzato per condurre l'indagine.

#### **3.1 Scelta della tipologia di software da utilizzare**

I software reperibili oggi per effettuare un compito di web scraping sono numerosi e la scelta di quale impiegare deve tenere conto di più fattori. Nel caso in questione, è stato necessario utilizzarne una tipologia che fosse gratuita (free-software) ed eventualmente aperta alla modifica dei parametri interni (open-source), ma soprattutto che potesse essere eseguita da un dispositivo single board computer.

Per questa tesi si è optato per una delle configurazioni più comuni: linguaggio di programmazione Python affiancato dalla libreria Selenium e due tipologie di WebDriver: ChromeDriver e PhantomJS.

Tutti questi verranno spiegati nel dettaglio nelle sezioni successive.

### **3.1.1 Linguaggio di programmazione Python**

Python è un linguaggio di programmazione sviluppato dall'olandese Guido van Rossum e pubblicato il 20 febbraio 1991. (Python Software Foundation, sito web)

È un linguaggio di programmazione ad alto livello: supporta il paradigma orientato ad oggetti, la programmazione strutturata e molte caratteristiche di programmazione funzionale.

Deriva il suo sistema di indentazione dal meno noto linguaggio di programmazione Occam: invece di usare parentesi o parole chiave, usa l'indentazione stessa per indicare i blocchi nidificati.

La grande quantità di librerie disponibili e la facilità con cui questo linguaggio permette di scrivere software modulare, favoriscono anche lo sviluppo di applicazioni molto complesse come nel caso della libreria Selenium la quale rende possibile utilizzare questo linguaggio nelle procedure di web scraping.



### 3.1.2 Libreria Selenium

*“Selenium automatizza i browser. Questo è tutto! Quello che fai con quel potere dipende interamente da te. In primo luogo, è per automatizzare le applicazioni Web a scopo di testing, ma non è certo limitato a questo. Anche le noiose attività di amministrazione basate sul web possono (e dovrebbero!) essere automatizzate”*

(Selenium, sito web)

La porzione di testo sovrastante è la semplice definizione del software Selenium fornita nella homepage del sito internet ufficiale. Selenium è un framework portatile per il testing di applicazioni web; i test possono essere eseguiti sulla maggior parte degli web browser moderni.

È un software Open Source, rilasciato con la licenza Apache 2.0: chiunque quindi può scaricarlo, utilizzarlo gratuitamente ed eventualmente consultare e modificare il codice sorgente.

Esistono numerose versioni di questo software e in questa indagine è stato scelto di utilizzare Selenium WebDriver. Esso riesce ad estrapolare dati e ad interagire con la pagina web in questione grazie all'avvio di un'istanza di un web browser compatibile con esso.

Questa versione specifica è disponibile come libreria di Python e, per quanto riguarda la parte dell'avvio dell'istanza del browser, è necessario utilizzare un software esterno. Di essi esistono numerose tipologie e in questa indagine si è scelto di utilizzarne due: ChromeDriver e PhantomJS i quali verranno illustrati nelle sezioni immediatamente successive.

### 3.1.3 WebDriver PhantomJS

PhantomJS è un WebDriver che ha la particolare caratteristica di essere “Headless” ovvero di non possedere un’interfaccia grafica. Ciò comporta numerosi vantaggi, in primis il fatto che non esegua il download delle parti grafiche di un sito, le quali molto spesso hanno dimensioni maggiori in termini di byte. Questo migliora significativamente la velocità di esecuzione e diminuisce il traffico di dati in entrata, fattore molto importante in una procedura di web scraping.

Infatti, durante l’esecuzione di un processo di questa tipologia, si crea una grande quantità di traffico dati tra il dispositivo che esegue il software di web scraping e il sito in questione, molte più di quante una singola persona possa realizzare.

Ciò può portare all’interruzione dello scambio di dati da parte del sito. Questo accade perché la maggior parte dei siti web esegue controlli automatizzati che mirano a bloccare chiunque operi uno spropositato traffico di dati, misura adottata anche da TripAdvisor.

Grazie quindi alla caratteristica “Headless”, il WebDriver PhantomJS riesce a compiere molte più richieste senza innescare questo blocco, cosa non possibile per gli altri WebDriver senza questa caratteristica.

D’altra parte, sempre per questa peculiarità, PhantomJS non è in grado di interagire con le pagine web, difatti non è possibile eseguire un click su un pulsante, come nell’esempio di questa indagine in cui è stato necessario eseguire un evento di tipologia “click” sul calendario fornito nel sito web di Tripadvisor per ottenere dati riguardanti il periodo d’interesse.

Nonostante questa limitazione, PhantomJS è stato utilizzato per ottenere tutti quei dati necessari che non mutano a seconda della selezione della data.

Questi sono i seguenti:

- coordinate geografiche
- numero di stelle
- numero di recensioni
- media del punteggio delle recensioni
- elenco dei servizi disponibili
- nome della struttura
- link della pagina dedicata alla struttura all'interno del sito web Tripadvisor.

### **3.1.4 WebDriver ChromeDriver**

È stato necessario utilizzare questo WebDriver per tutte le informazioni che variano a seconda della selezione della data all'interno del sito web TripAdvisor, esse sono:

- prezzo
- provider che offre tale prezzo.

Infatti per questa indagine si è scelto di ottenere dati sul prezzo del giorno stesso dell'acquisizione riguardo la permanenza di una notte per due persone.

Per fare ciò è stato necessario eseguire un evento di tipologia “click” all'interno del calendario fornito dal sito web TripAdvisor, sulla casella che indica la data odierna.

La figura 4 è uno screenshot eseguito il 31 marzo 2019 sul risultato di una comune ricerca. Non è stato eseguito ancora alcun evento di tipologia “click”, ma il sito web sta comunque visualizzando dei risultati. Questo accade perché come impostazione predefinita, il sito TripAdvisor mostra i prezzi riguardanti le date del fine settimana successivo a quello odierno.

Per riuscire quindi a selezionare la data odierna è stato necessario indicare a ChromeDriver di eseguire un evento di tipologia “click” sulla casella che la contraddistingue, in modo tale da poter accedere ai dati d’interesse.

Da notare che, come nel caso illustrato dalla figura 5, quando il fine settimana successivo a quello della data odierna è collocato nel mese successivo a quello corrente, è necessario anche eseguire un altro evento della medesima tipologia sul pulsante che offre il cambio del mese, indicato dalla freccia posta alla sinistra del nome del mese in questione.

Così facendo comparirà il mese appartenente alla data odierna e sarà possibile eseguire l’evento “click” su di essa.

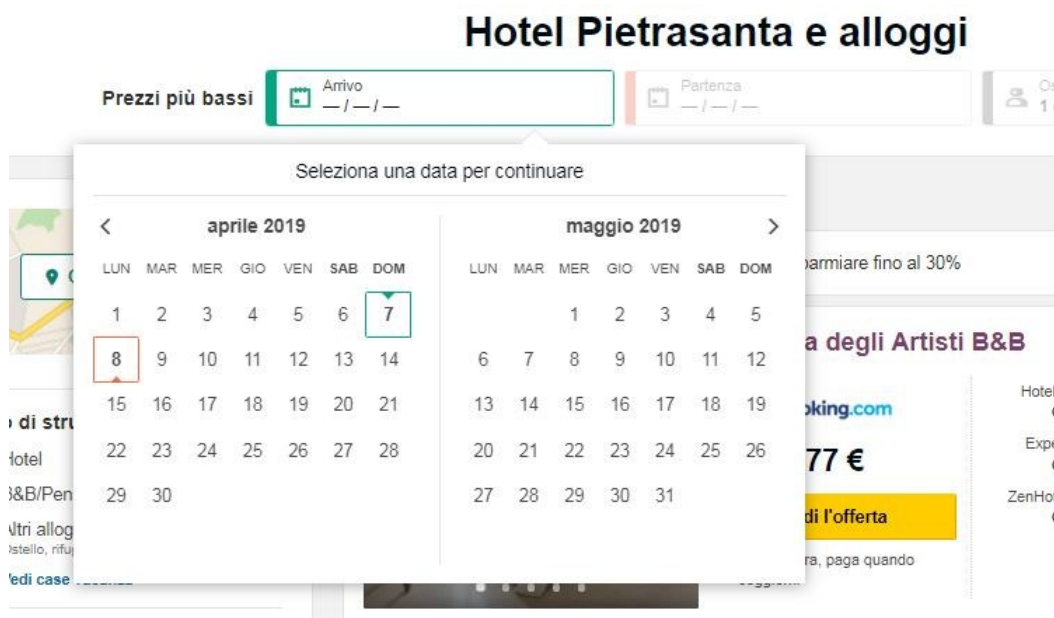


Figura 4. Screenshot ottenuto dal sito web TripAdvisor, selettore data

## Hotel Pietrasanta e alloggi

Prezzi più bassi

Arrivo  
dom 31/03/19

Partenza  
lun 01/04/19

Seleziona una data per continuare

marzo 2019							aprile 2019						
LUN	MAR	MER	GIO	VEN	SAB	DOM	LUN	MAR	MER	GIO	VEN	SAB	DOM
				1	2	3	1	2	3	4	5	6	7
4	5	6	7	8	9	10	8	9	10	11	12	13	14
11	12	13	14	15	16	17	15	16	17	18	19	20	21
18	19	20	21	22	23	24	22	23	24	25	26	27	28
25	26	27	28	29	30	31	29	30					

Figura 5. Screenshot ottenuto dal sito web TripAdvisor, selettore data con interazione

### 3.1.5 Salvataggio dati su un database SQL

Per il salvataggio dei dati è stato creato un database specifico attraverso il software phpMyAdmin.

Come illustrato dalla figura 6, il database è costituito da due tabelle. La prima, chiamata strutture, al suo interno possiede tutte le informazioni non variabili di una struttura ovvero:

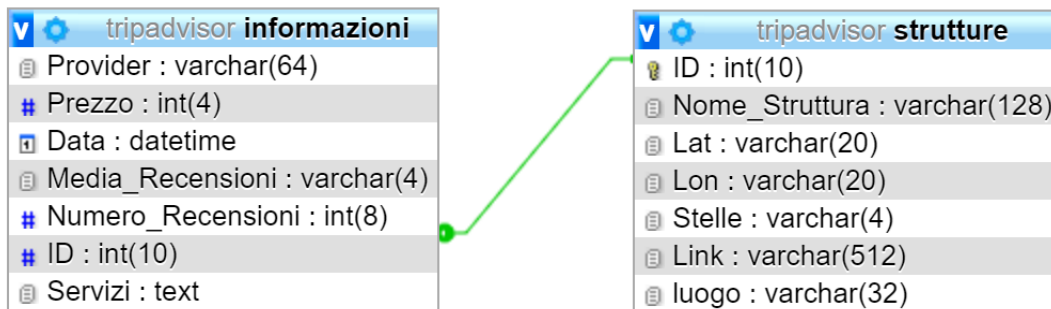
- ID all'interno del sito web TripAdvisor (impostato come chiave primaria)
- nome della struttura
- coordinate geografiche espresse in latitudine e longitudine
- numero di stelle
- link della pagina dedicata ad essa all'interno del sito web TripAdvisor
- comune di appartenenza

Ogni qualvolta che durante l'acquisizione dati verrà acquisita una struttura sarà controllato se essa fa parte o meno della suddetta tabella; se mancante verrà aggiunta mentre se presente ignorata.

La seconda tabella che fa parte del database è quella che contiene tutte le informazioni variabili nel tempo della struttura, esse sono le seguenti:

- Prezzo
- Provider che propone il prezzo indicato
- Numero di recensioni
- Media dei punteggi espressi dagli utenti
- Elenco di tutti i servizi disponibili

Per rendere ogni record di questa tabella unico si è scelto di riutilizzare il campo "ID" della precedente come chiave esterna e di aggiungere un campo chiamato "Data" dove vi verrà impressa l'ora e la data esatta di ogni acquisizione dati. Questi due campi costituiscono la superchiave della tabella.



*Figura 6. Screenshot ottenuto da phpMyAdmin, database visualizzato con la modalità designer*

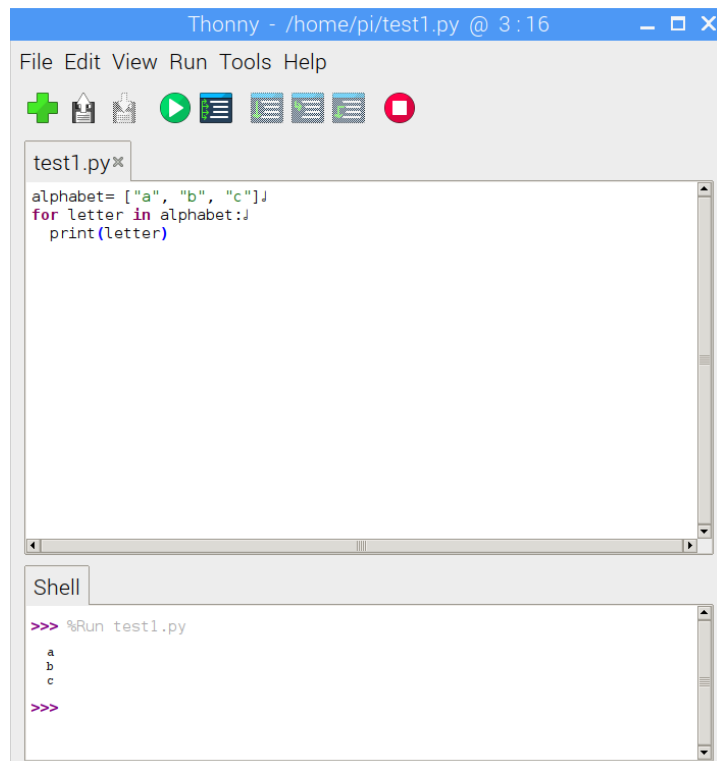
Questi dati vengono aggiunti al database dal programma di web scraping grazie al software MySQL e MySQL Connector.

### 3.1.6 Software per la pianificazione di comandi

Per non dover delegare all'utente il compito di eseguire manualmente l'avvio del processo di acquisizione dati, si è adottato un software per la pianificazione di comandi.

La variante più comune di questa tipologia è senz'altro "Crontab" il quale però non è in grado di eseguire processi che richiedano la visualizzazione di un'interfaccia grafica, fattore che andrebbe a rendere impossibile l'esecuzione del software "ChromeDriver".

Si è quindi optato per "xdotool", un semplice software che riesce a simulare il movimento e click del mouse, quello della digitazione di un pulsante dalla tastiera e a programmare la ripetizione di questi compiti nel tempo.



*Figura 7. Ambiente di sviluppo integrato per il linguaggio di programmazione Python Thonny*

Esso è stato applicato in combinazione con l'ambiente di sviluppo integrato per il linguaggio di programmazione Python Thonny, illustrato nella figura 7.

Se infatti si esegue un evento di tipologia click sul pulsante contrassegnato dall'icona con la freccia verde dell'interfaccia, sarà possibile avviare l'esecuzione del file in formato Python in input al programma.

È stato quindi sufficiente indicare a xdotool, di eseguire tale azione autonomamente nei tempi prestabiliti.

### **3.2 Scelta della tipologia di hardware da utilizzare**

Come già illustrato precedentemente in questa relazione, la singola operazione di web scraping deve essere eseguita più di una volta al giorno e la sua durata varia da due alle tre ore di tempo. Inoltre, un software per la pianificazione di eventi, deve operare senza accensioni o spegnimenti da parte del dispositivo sul quale è in esecuzione.

Questi fattori rendono indispensabile l'utilizzo di un computer che non venga mai riavviato durante l'intero periodo d'acquisizione dei dati, che nel caso in questione, ha la durata di un mese.

Utilizzare un normale computer fisso o laptop di ultima o penultima generazione coprirebbe sicuramente la potenza di calcolo necessaria per eseguire software di questo tipo, ma rappresenterebbe un costo di alimentazione eccessivo per le risorse di un privato, senza considerare l'inevitabile usura degli strumenti che in questo caso possono avere costi elevati, sempre in rapporto alle risorse di un privato.



Anche il suono prodotto dalle ventole di raffreddamento e la loro considerevole produzione termica, richiederebbero inevitabilmente l'utilizzo di un'apposita area adibita a tale scopo.

La scelta quindi è stata quella di utilizzare un single board computer, ovvero un dispositivo, nella fattispecie di piccole dimensioni, costruito interamente su di un singolo circuito stampato.

Nel mercato di oggi sono disponibili numerose varianti a basso prezzo, molte delle quali richiedono una quantità di alimentazione considerevolmente inferiore a quella di un laptop o di un pc fisso.

Nello specifico è stato scelto un Raspberry Pi 3 Model B+. I motivi della scelta verranno illustrati nelle sezioni sottostanti.

### **3.2.1 Economia del prodotto in relazione alle caratteristiche**

Questo prodotto è acquistabile presso il sito ufficiale ad una cifra pari a 35,95€, quantità consona alle risorse disponibili dal candidato.

Le sue caratteristiche hardware sono sufficienti per riuscire ad eseguire fluidamente i software necessari per condurre l'indagine. Infatti esso vanta le seguenti caratteristiche (Raspberry, sito web):

- Broadcom BCM2837B0, Cortex-A53 (ARMv8) 64-bit SoC @ 1.4GHz
- 1GB LPDDR2 SDRAM
- 2.4GHz and 5GHz IEEE 802.11.b/g/n/ac wireless LAN, Bluetooth 4.2, BLE
- Gigabit Ethernet over USB 2.0 (maximum throughput 300 Mbps)

- Extended 40-pin GPIO header
- Full-size HDMI
- 4 USB 2.0 ports
- CSI camera port for connecting a Raspberry Pi camera
- DSI display port for connecting a Raspberry Pi touchscreen display
- 4-pole stereo output and composite video port
- Micro SD port for loading your operating system and storing data
- 5V/2.5A DC power input
- Power-over-Ethernet (PoE) support (requires separate PoE HAT)

### 3.2.2 Potenzialità del sistema operativo Raspbian

*“Raspbian è un sistema operativo gratuito basato su Debian ottimizzato per l'hardware Raspberry Pi. Un sistema operativo è l'insieme di programmi e utilità di base che esegue il tuo Raspberry Pi. Tuttavia, Raspbian fornisce più di un sistema operativo puro: viene fornito con oltre 35.000 pacchetti, software precompilato in bundle in un formato piacevole per una facile installazione sul tuo Raspberry Pi ...*

*Nota: Raspbian non è affiliato con la Raspberry Pi Foundation. Raspbian è stato creato da un piccolo team dedicato di sviluppatori appassionati dell'hardware Raspberry Pi, dagli obiettivi educativi della Raspberry Pi Foundation e, naturalmente, dal Progetto Debian.”* (Raspbian, sito web)

Raspbian è un sistema operativo costruito su misura per la famiglia dei single board computer marchio Raspberry, che trova le sue massime prestazioni proprio nel modello utilizzato in questo progetto.

Infatti è possibile eseguire operazioni come: navigare sul web, utilizzare il software LibreOffice, eseguire operazioni su file e molto altro ancora con estrema facilità. Il prodotto in questione risulta essere il più valido del suo mercato non per le sue caratteristiche hardware ma proprio per la sua perfetta integrazione con questo sistema operativo.

Esso contiene inoltre numerosi software preinstallati, alcuni dei quali sono stati utilizzati per la progettazione e lo sviluppo del software di scraping come ad esempio l'ambiente di sviluppo integrato per il linguaggio di programmazione Python Thonny, il software Python in versione 3.7, il web browser Chromium e l'emulatore di terminale LXTerminal.

### 3.2.3 Requisiti di alimentazione

Raspberry Pi 3 Model B+ necessita di essere alimentato da un alimentatore da 5V, 2.5A, quindi da un alimentatore di un formato simile a quello di uno per smartphone.

Per quanto riguarda il consumo effettivo del dispositivo, sono disponibili sul web vari studi effettuati a riguardo. Uno di essi (RasPi.TV, sito web) comunica i seguenti dati:

	Zero	Zero W	A+	A	B+	B	Pi2B	Pi3B	Pi3B+
	mA	mA	mA	mA	mA	mA	mA	mA	mA
<b>Idling</b>	100	120	100	140	200	360	230	230	400
<b>Loading LXDE</b>	140	160	130	190	230	400	310	310	690
<b>Watch 1080p Video</b>	140	170	140	200	240	420	290	290	510
<b>Shoot 1080p Video</b>	240	230	230	320	330	480	350	350	520

*Tabella 2. Screenshot ottenuto dal sito web RasPiTV, consumi a confronto modelli di Raspberry*

La tabella 2 indica la quantità di utilizzo dell'energia elettrica da parte dei vari modelli di single board computer prodotti dalla casa produttrice Raspberry, espressi con l'unità di misura dei milliampere. L'ultima colonna indica il modello corrispondente a quello utilizzato.

I dati ci comunicano che in stato di “Idling” - termine che inglese significa “minimo”, ovvero quando la macchina è accesa ma in esecuzione non c'è alcun tipo di processo - il consumo è di 400 mA.

In stato di caricamento del sistema operativo, il consumo è di 690 mA ed è l'operazione in cui la macchina richiede il maggior quantitativo di energia.

Mentre si riproduce un video di qualità medio-alta, ovvero di 1080p (operazione che richiede una moderata potenza computazionale sui computer fissi o laptop di ultima o penultima generazione) il consumo è di 510 mA, se invece si sta producendo un file video con le stesse caratteristiche (operazione più pesante di quella precedente), il consumo è di poco superiore, ovvero di 520 mA.

Per convertire questi risultati in Watt è sufficiente dividere ciascuno di questi dati per il numero 1000 - in modo tale da trasformarli da milliampere ad ampere - e successivamente moltiplicarli per il numero di Volt in uscita dell'alimentatore, che in questo è pari a 5. Otteniamo quindi rispettivamente: 2W per lo stato di “Idling”, 3.45W per quello del caricamento del sistema operativo, 2.55W durante la riproduzione di un video in 1080p e 2.6W durante la produzione di un file equivalente.

Per stimare, a grandi linee, il consumo di un pc fisso e di un laptop di ultima o penultima generazione si è deciso di prendere in analisi i dati forniti dal sito (Navigaweb.net, sito web). Esso ci comunica che i pc fissi hanno un consumo in media di 100W in stato di “Idling”, 200W quando in sistema operativo è in esecuzione, e dai 300 ai 350W quando si utilizza a pieno regime.

Per i laptop i consumi sono molto più moderati rispetto a questi ultimi, ma comunque di gran lunga superiori a quelli del Raspberry Pi B3+. Sempre lo stesso sito ci comunica

nello stato di “Idling” il consumo è di 20W, 30W durante l’esecuzione del sistema operativo e, durante l’uso intensivo, si possono registrare picchi fino ai 50W.

In conclusione questi dati ci confermano che il consumo del Raspberry Pi 3 Model B+ è nettamente inferiore a quello di un pc fisso o di un laptop e che il suo utilizzo per tutta la durata delle acquisizioni non comporterà costi di alimentazione elevati.

Il confronto è reso esplicito nella seguente tabella:

<b>Dispositivo</b>	<b>Watt minimi</b>	<b>Watt massimi</b>
Computer fisso	100	350
Laptop	20	50
Raspberry Pi 3 Model B+	2	3.45

*Tabella 3, consumi a confronto espressi in Watt dei tre sistemi*

### **3.3 Costruzione del dispositivo di web scraping**

Nelle prossime sezioni verranno illustrati tutti i procedimenti che hanno portato alla costruzione del dispositivo di web scraping. Inizieremo con la parte dedicata all’hardware, per poi passare al lato software per concludere con la manutenzione di quest’ultimo durante il periodo delle acquisizioni.

### 3.3.1 Accessori aggiuntivi

Il dispositivo Raspberry Pi B3+ necessita, per il corretto funzionamento, dell'aggiunta di componenti supplementari. In questa relazione, per motivi di praticità, sono stati divisi in due categorie: necessari e non necessari.

Per quanto riguarda i componenti necessari è opportuno chiarire che, se si acquista un Raspberry Pi 3 Model B+ presso il sito ufficiale del produttore, nella confezione sarà presente soltanto il single board computer stesso.

Esso necessita di altri due componenti: una scheda di memoria Micro SD, che svolgerà la funzione analoga a quella di un disco rigido e di un alimentatore correlato ad un cavo micro USB di tipo b.

Entrambi sono quindi stati acquistati separatamente. Per la scheda di memoria Micro SD si è optato per una che potesse lavorare ad alte velocità, in modo da non interferire con il corretto funzionamento del dispositivo e che non ponesse limiti di spazio invalidanti per l'obiettivo del progetto. Quindi ne è stata acquistata una con 32 GB di memoria e velocità massima di scrittura e lettura pari a 95 MB/s.

Per quanto riguarda il sistema di alimentazione è stato semplicemente acquistato un alimentatore che fosse corrispondente alle richieste di energia del dispositivo indicate nel sito del produttore e come cavo, è stato utilizzato quello in dotazione all'alimentatore.

Passando ai componenti non necessari si è deciso di acquistare, per migliorare la solidità tecnica del dispositivo, un case fornito presso rivenditori terzi alla società Raspberry.



*Figura 8. Case applicato ad un Raspberry Pi 3 Model B+*

Esso permette, come illustrato dalla figura 8, anche l'installazione di una ventola di raffreddamento la quale risulta essere un componente essenziale in situazioni in cui si deve mantenere in funzione un dispositivo per molto tempo ad eseguire operazioni che spingono al limite la potenza computazionale dello stesso.

L'ultimo componente di questo elenco è un set di tre dissipatori che sono stati inseriti in corrispondenza delle parti più soggette a sbalzi di calore del dispositivo; essi sono illustrati nella figura 9.



*Figura 9. Dissipatori compatibili con Raspberry Pi 3 Model B+*

### 3.3.2 Installazione hardware

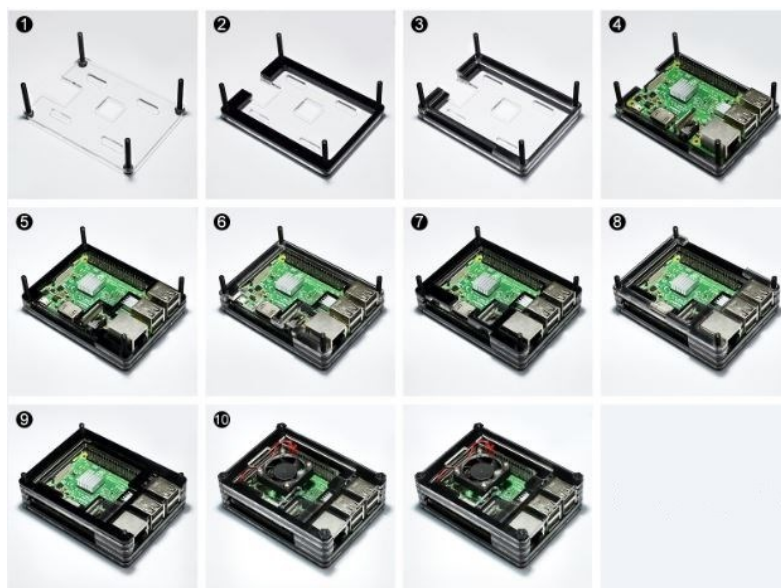
Il primo passo di questa procedura è stato applicare i tre dissipatori sulle parti più soggette a sbalzi di calore del dispositivo, come indicato nella figura 10.



*Figura 10. Applicazione dissipatori sul Raspberry Pi 3 Model B+*

Successivamente si è passati all'integrazione del computer con il case e all'installazione della ventola di raffreddamento. Questo passaggio è illustrato nella figura 11.





*Figura 11. Applicazione del case e della ventola di raffreddamento sul Raspberry Pi 3 Model B+*

### **3.3.3 Installazione del sistema operativo**

Si è deciso, come già spiegato nella sezione 2.2.2, per l'utilizzo del sistema operativo Raspbian. L'installazione di questo sistema operativo ha seguito i seguenti passi:

In primo luogo è stata scaricata una copia dello stesso, presente nel sito web della casa produttrice Raspberry, attraverso un computer che è stato utilizzato unicamente per compiere questa e le prossime tre operazioni.

Successivamente è stato necessario inserire la scheda di memoria SD all'interno del computer attraverso un apposito adattatore fornito nella confezione della memoria al momento dell'acquisto.

In seguito è stato scaricato il software Etcher presso il sito <https://etcher.io/> grazie al quale è stato possibile inserire il file contenente il sistema operativo all'interno della memoria.

Una volta terminato il processo, la memoria è stata estratta dal computer e inserita all'interno del Raspberry Pi 3 Model B+ presso l'apposito lettore di schede di memoria come indicato nella figura 12.



*Figura 12. Scheda di memoria SD inserita nel lettore di schede del Raspberry Pi 3 Model B+*

A questo punto il dispositivo è stato collegato alla rete elettrica attraverso l'alimentatore fornito, ad un mouse, ad una tastiera e ad un monitor attraverso un cavo HDMI.

Il dispositivo è stato avviato con il sistema operativo già pronto all'uso, senza ulteriori installazioni o configurazioni.

Questa procedura è illustrata nel dettaglio presso l'indirizzo web "[https://www.w3schools.com/nodejs/nodejs\\_raspberrypi.asp](https://www.w3schools.com/nodejs/nodejs_raspberrypi.asp)".

### **3.3.4 Installazione del software necessario al progetto**

In primo luogo è stato necessario installare il software Selenium. Ciò è stato possibile con l'esecuzione del comando `"sudo pip install selenium"` all'interno del terminale.

Anche l'installazione del software PhantomJS è avvenuta tramite l'esecuzione di un comando nel terminale ovvero: `"sudo apt-get install phantomjs"`.

Per quanto riguarda il secondo WebDriver utilizzato, ChromeDriver, è stato necessario recarsi presso il sito web "<http://chromedriver.chromium.org/>" e scaricare l'apposito file. Una volta eseguita questa procedura si è dovuto collocare tale file all'interno della cartella situata nel percorso `"usr/local/bin"` all'interno del dispositivo.

Per la parte dell'installazione del software di gestione del formato SQL il primo passo è stato quello di installare MySQL ed è stato fatto con il comando: `"sudo apt-get install mysql-server php-mysql -y"`; successivamente è stato installato MySQL Connector, una libreria di Python in grado di far interagire quest'ultimo con il

linguaggio SQL. È stato installato attraverso l'esecuzione del comando `"python -m pip install mysql-connector"`.

Si è poi proseguito con l'installazione di phpMyAdmin, per il quale è stato necessario svolgere preventivamente l'installazione di due componenti: per primo Apache2 installato con il comando: `"sudo apt-get install apache2"` e per secondo il linguaggio PHP vero e proprio, installato con `"sudo apt-get install php libapache2-mod-php"`.

Quindi è stato installato phpMyAdmin con `"sudo apt-get install phpmyadmin"`.

Infine è stato installato il software per l'automazione dei comandi della tastiera `"xdotool"` con il seguente comando `"sudo apt-get install xdotool"`

### **3.3.5 Progettazione e sviluppo del software di web scraping**

Prima di illustrare i vari passaggi eseguiti dal software, è opportuno chiarire la tecnica di raccolta dati utilizzata per questa indagine.

Per mettere in atto una procedura di web scraping, Selenium dispone varie tecniche di ricerca delle informazioni, le principali sono:

- XPATH
- CSS Selector
- ID
- Classi
- Ricerca tramite espressioni regolari.

I siti web, per tutelarsi da chiunque esegua procedure di questa tipologia, mutano spesso la struttura delle loro pagine, rendendo quindi inefficaci la maggior parte delle tecniche sopracitate.

In questi casi risulta più conveniente la ricerca per espressioni regolari.

Ciò è dato dal fatto che, a differenza delle altre citate, questa tecnica non è dipendente dalla struttura gerarchica della pagina web che analizza e, in egual modo, dal fatto che ogni informazione che si consulta normalmente su un sito web è contenuta anche all'interno del codice HTML.

Esse sono situate in mezzo ad altre informazioni (codici in formato HTML) che ne indicano le varie caratteristiche di visualizzazione.

Queste, a differenza del dato da estrapolare, tendono ad essere sempre le stesse per ogni tipologia di informazione ricercata. Ciò le rende degli importanti punti di riferimento per la ricerca dei dati poiché non sono quindi ignote. Ergo sarà sufficiente individuare i codici HTML adiacenti ai dati d'interesse e indicare al software di ottenerli a partire dalla posizione ottenuta.

Un esempio dell'applicazione di questa tecnica è fornito dalla figura 13 la quale è stata realizzata tramite uno screenshot eseguito dal web browser Google Chrome.

Essa ci mostra come il nome della struttura sia presente anche all'interno del codice HTML della pagina web.

Esso infatti è delimitato da una parte dalla stringa:

```
"<h1 id="HEADING" class="ui_header h1">"
```

e dall'altra da:

```
"</h1>"
```

Questo è valido per tutte le pagine dedicate alle singole strutture, come mostrato dalla figura 14 dove, nonostante che la pagina web sia dedicata ad un'altra struttura, la stringa che ne delimita il nome all'interno del codice HTML rimane la stessa.

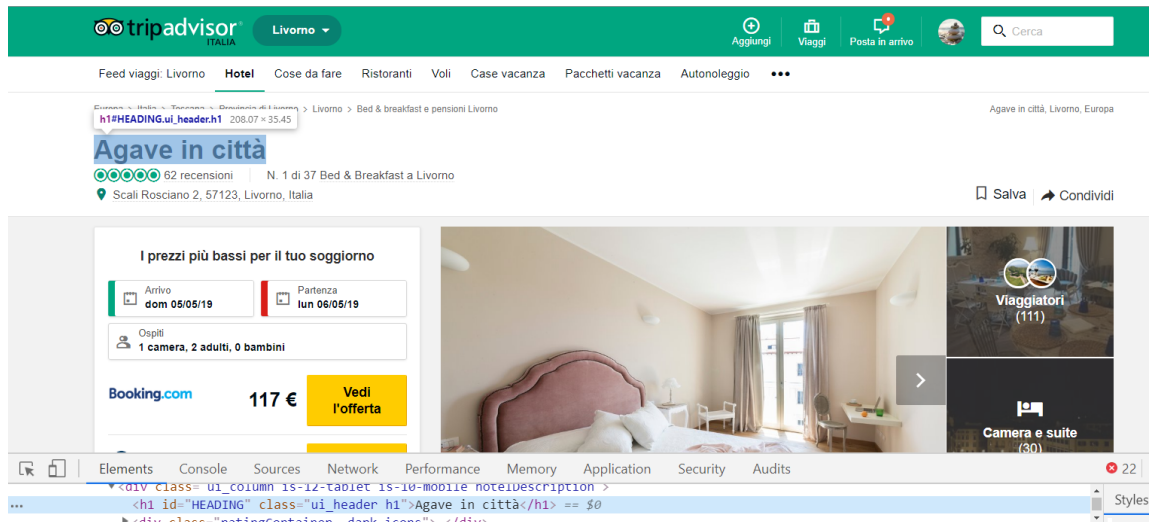


Figura 13. Screenshot ottenuto del sito web TripAdvisor, titolo della struttura nel codice HTML

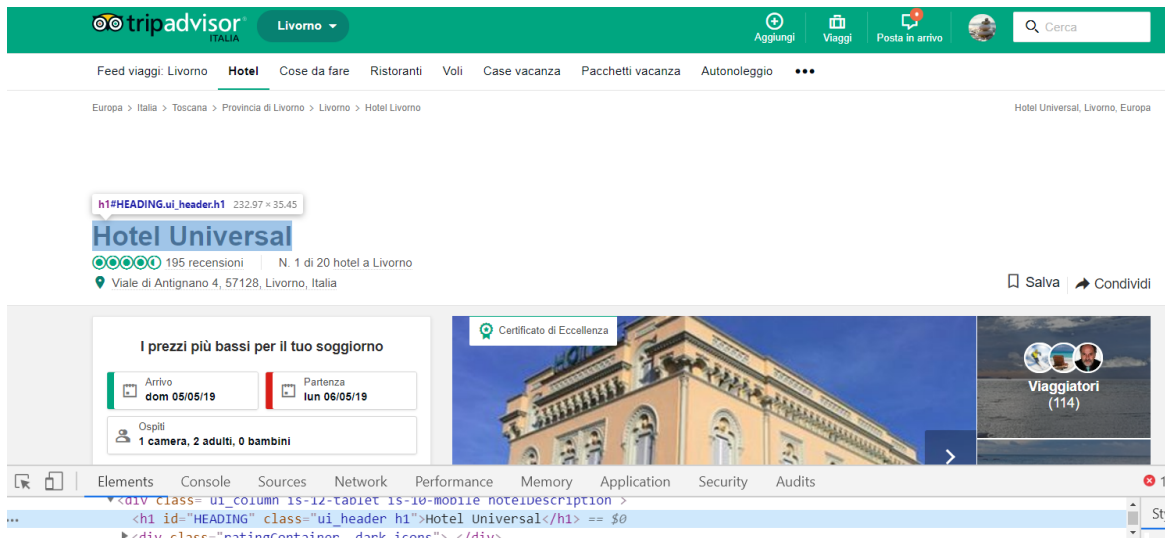


Figura 14. Screenshot ottenuto del sito web TripAdvisor, titolo della struttura nel codice HTML

Procediamo adesso con la descrizione dei vari passaggi eseguiti dal software.

Esso opera a partire da un link che mostra i risultati della prima pagina di una ricerca sul sito web TripAdvisor in merito ad uno dei quattro comuni, dal quale riesce ad ottenere tutti i link relativi alle singole pagine dedicate alle strutture.

Dopodiché, provvede con l'avanzare tra le molteplici pagine dei risultati grazie ad un cambio nella dicitura del link iniziale delle stesse.

Infatti, se nel link della prima pagina dei risultati:

```
"https://www.tripadvisor.it/Hotels-g187897-oa00-Livorno_Province_of_Livorno_Tuscany-Hotels.html"
```

contenente la dicitura "oa00", ai doppi zeri viene sommato il numero 30 e ricaricata la pagina, allora verranno visualizzati i risultati relativi alla seconda, la quale stavolta avrà la dicitura "oa30" e così via fino ad arrivare all'ultima.

Il numero di pagine disponibili invece, che corrisponde quindi al numero di somme effettuate, è indicato al termine di ogni pagina di ricerca, come illustrato dalla figura 15, dove in questo caso il numero in questione va incrementato 12 volte di 30 unità, fino ad arrivare alla dicitura di "oa330".



*Figura 15. Screenshot ottenuto del sito web TripAdvisor, selettore della pagina dei risultati*

Successivamente i link delle singole strutture vengono analizzati dal WebDriver PhantomJS, il quale provvede a ricercare tutte le informazioni d'interesse.

Grazie alla sua caratteristica di "Headless" riesce a compiere numerose operazioni senza oltrepassare il limite di traffico dati consentito dal sito web in questione.

Per questa parte dell'operazione risulta essere ideale, in quanto non si necessita di generare un evento di tipologia “click” atto all'impostazione della data di arrivo e partenza, poiché le informazioni che si vanno ad estrapolare non sono soggette al cambiamento dovuto a questo tipo di impostazione.

Questa operazione ha creato diversi problemi poiché, a causa di dinamiche interne al sito web TripAdvisor, la struttura del codice HTML dello stesso tendeva a cambiare imprevedibilmente. Questo accade per la ricerca di diverse tipologie di informazioni, come ad esempio per l'elenco dei servizi della struttura, i quali sono preceduti da codice HTML che si presenta casualmente in due diverse forme.

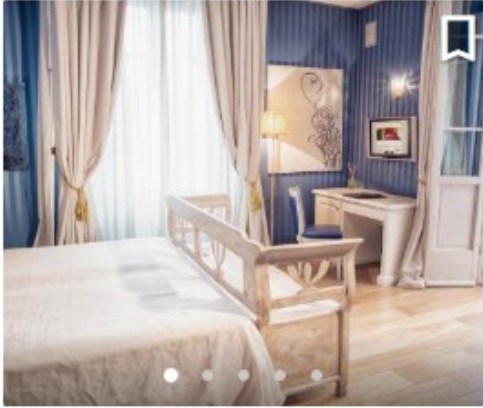
Sfortunatamente non esiste una procedura risolutiva per riuscire a prevedere questi cambiamenti, semplicemente perché è un fattore gestito dal sito web in questione. L'unico modo possibile per fronteggiare adeguatamente questo inconveniente, è quello di eseguire molte prove del software per lunghi periodi e molte volte nella durata di uno stesso giorno, in modo tale da potersi scontrare contro ogni possibile alterazione della struttura e rendere il software di web scraping più tollerante possibile.

I dati ottenuti con il WebDriver PhantomJS vengono poi aggiunti al database precedentemente creato, grazie alla libreria Python “MySQL connector” che permette al software Python di interagire con il database. Questo passo è illustrato nel dettaglio nella sezione 3.1.5.


La seconda parte del software è quella in cui si utilizza il WebDriver ChromeDriver. Esso, come descritto nella sezione 3.1.4, ha la possibilità di generare un evento di tipologia “click” all'interno del sito web che analizza.

Per poter limitare l'ammontare di traffico dati generato da questo WebDriver, si è scelto di operare non all'interno delle pagine dedicate alle singole strutture, ma in quelle di ricerca, poiché esso vi è disponibile sotto forma di anteprima come mostrato dalla figura 16.





### Hotel Palazzo Guiscardo



**168 €**


Vedi l'offerta

Expedia.it ↗  
159 €


ZenHotels.com ↗  
146 €

Booking.com ↗  
168 €

Vedi tutte le 8 offerte da 146 € ▼



### Albergo Pietrasanta



**247 €**

Vedi l'offerta

Agoda.com ↗  
247 €

Hotels.com  
🔍

Roomdi.com  
🔍

Vedi tutte le 5 offerte da 247 € ▼

Figura 16. Screenshot ottenuto dal sito web TripAdvisor, risultati di ricerca.

Per ottenere questo dato è necessario però prima selezionare la data d'interesse, che corrisponde a quella odierna, dall'apposito calendario disponibile nel sito web in questione come già illustrato ampiamente nella sezione 3.1.4.

Per poter generare un evento di tipologia “click” e simulare quindi l'intervento di un utente umano, è necessario conoscere il percorso XPATH della casella che indica la data della giornata odierna (per percorso XPATH s'intende la semplice espressione del codice HTML sotto forma di una gerarchia).

Questa necessità ha generato molti problemi, primo tra tutti il fatto che il percorso XPATH è un elemento molto esposto al cambiamento: la sua struttura muta a seconda dell'aggiunta o rimozione di un qualsiasi elemento all'interno del codice HTML che lo

precede gerarchicamente, fattore che, come già discusso, avviene per dinamiche interne al sito e quindi impossibili da prevedere.

Una volta estrapolati, anche questi dati vengono aggiunti al database tramite la medesima tecnica e al termine della procedura si passa poi ad un nuovo link che indica i risultati di ricerca del successivo comune.

### **3.3.6 Definizione della durata delle acquisizioni**

Si è deciso di avviare il software per due volte al giorno: alle ore 11:00 e alle ore 18:00. Questo lo si è fatto non solo per monitorare eventuali cambiamenti di prezzi durante la durata del singolo giorno, ma anche per prevenire il possibile fallimento di un'esecuzione.

Così facendo, se ad esempio nella prima acquisizione delle 11:00, si dovesse riscontrare un problema, sarà possibile correggere il software in modo tale da far sì che non comprometta anche l'acquisizione delle 18:00 durante il lasso di tempo che c'è tra la fine della prima e l'inizio della seconda.

Al contrario, se un errore dovesse accadere nella seconda acquisizione, allora avremmo comunque a disposizione quella precedente e i dati relativi alla singola giornata non andrebbero perduti.

La durata complessiva delle acquisizioni è avvenuta dal 11 febbraio fino all'11 di marzo.

### **3.3.7 Manutenzione del dispositivo durante la durata delle acquisizioni**

Durante la durata delle acquisizioni il software "xdotool" ha permesso l'avvio automatico del dispositivo.

Tuttavia sono stati eseguiti test prima dell'inizio della maggior parte di esse per verificare il corretto funzionamento del dispositivo e lo stato del sito web in questione, in modo tale da fronteggiare anche eventuali cambiamenti di quest'ultimo.

Al termine di ogni esecuzione sono stati verificati manualmente i dati ottenuti attraverso numerosi controlli a campione.

## **4. Analisi dei dati ottenuti**

Nella prima parte di questo capitolo verranno definite, a grandi linee, le varie tipologie di clienti frequentatori di strutture alberghiere, in seguito verranno illustrati i risultati dei dati ottenuti suddivisi per area e infine verrà associata la tipologia di clienti alla suddetta area.

Nella seconda parte del capitolo ci si occuperà delle strategie di visualizzazione dei dati ottenuti.

### **4.1 Cosa s'intende per tipologia di clienti**

Nel testo “marketing del turismo” di Philip Kotler, John Bowen e James Makens, sono indicati come “segmenti di visitatori” e con questo termine viene fatto riferimento alle varie tipologie di clienti che alloggiano in una determinata struttura.

Esse non sono universalmente definite ma, ai fini dell'indagine svolta, verranno prese come esempio quelle descritte nel testo sopracitato. Esse verranno illustrate nella sezione successiva.

### **4.1.1 Definizione delle varie tipologie di clienti**

In questa relazione è stato scelto di definire le differenti tipologie di clientela, tramite il testo (Philip Kotler, John Bowen, James Makens, 1996, p.361).

Le suddivisioni sono le seguenti:

- **Turismo di massa organizzato:** ovvero gruppi di persone dove si viaggia non in cerca di una specifica meta, ma per l'acquisto di un pacchetto rispetto ad un altro. Il viaggio è in pullman e lo shopping nei mercatini locali rappresenta l'unica forma di contatto con la popolazione del luogo. La vacanza è già interamente decisa a priori compresa della scelta dell'alloggio.
- **Turismo di massa individuale:** ovvero gruppi di persone con maggiore autonomia e controllo sull'itinerario del viaggio. Spesso si spostano noleggiando auto per visitare le diverse mete e attrazioni.
- **Esploratori:** sono coloro che pianificano autonomamente il loro viaggio prenotando singolarmente i pernottamenti. Interagiscono con la popolazione locale.
- **Girovaghi:** s'intendono persone, per lo più di età giovanile, che viaggiano in gruppo. Si mischiano facilmente ai gruppi locali dal punto di vista socio-economico più basso e si spostano con mezzi pubblici nella classe più economica.

### **4.1.2 Analisi e considerazioni sui dati ottenuti**

I dati ottenuti, dal punto di vista della media prezzo relativa alla prenotazione della permanenza per la durata di un giorno per due persone in merito alle quattro aree d'interesse, vedono Pietrasanta come il Comune più costoso con una media prezzo di 92,2€, seguita da Cascina e Lucca con rispettivamente 74,43€ e 73,71€.

Il Comune più economico è quello di Livorno con 70,80€.

Per quanto riguarda la media dei punteggi ottenuti, che si basa da un voto che va da zero a cinque, vede il Comune di Pietrasanta con una media di 4.10, seguito da Lucca con 4.07, Cascina con 4.05 e come ultimo il Comune di Livorno con 3.72.

Per i servizi occorre asserire che il confronto è stato fatto dove vi sono differenze più significative. Segue quindi una panoramica suddivisa per ognuno dei quattro comuni.

Per il Comune di Pietrasanta, i servizi più evidenti risultano essere quelli di: servizio in camera, frigorifero e minibar in camera, servizio babysitter, suite, vasca con jacuzzi e campo da tennis.

Per il Comune di Cascina tra i servizi più frequenti si riscontra: disponibilità di parcheggio gratuito, colazione inclusa, servizio navetta, trasporti per l'aeroporto, camere per famiglie, sale riunioni e centro congressi con accesso ad internet, sala banchetti e attività per bambini e famiglie.

Nel Comune di Livorno i servizi maggiormente presenti sono: animali domestici ammessi, presenza di un bar, accessibilità in sedia a rotelle, servizio lavanderia, presenza di concierge, la presenza di ristorante, microonde in camera ed è l'unico Comune ad avere una struttura a possedere una colonnina di ricarica per veicoli elettrici.

Per il Comune di Lucca il servizio più frequente è personale multilingue. Per il resto si ha una situazione di costante media con le altre città, senza mai avere la presenza di un servizio in difetto o esubero rispetto alle altre zone.

#### **4.1.3 Associazione delle tipologie di clienti alle zone d'interesse**

Il Comune di Pietrasanta risulta essere una meta eccellente per la tipologia di turismo di massa individuale.

Questo si evince dalla media prezzo sensibilmente più alta tra i quattro Comuni analizzati e anche dal fatto che i suoi servizi caratterizzanti si rivelano essere ridondanti per gli altri segmenti di visitatori illustrati, in particolare per la categoria dei girovaghi.

Il Comune di Cascina si presta perfettamente per la categoria di turismo di massa, ciò è dovuto alla sua media prezzo moderata e ai servizi che lo caratterizzano come ad esempio camere per famiglie e parcheggio gratis.

In generale è altrettanto corretto affermare che si presenta fruibile per tutte le categorie di clienti citate.

Il Comune di Livorno, grazie alla sua media prezzo tendenzialmente più bassa, e ai servizi quali la disponibilità di forno a microonde in camera e la possibilità di ospitare animali domestici, si presta principalmente alla tipologia dei girovaghi e degli esploratori. Da segnalare la presenza di una colonnina per la ricarica dei veicoli elettrici, fattore importante per la sostenibilità ambientale.

Infine il Comune di Lucca, con la una media prezzo moderata e un'equa distribuzione dei servizi, si presta generalmente per tutte le tipologie di clientela descritte.

In particolare l'unico servizio predominante, ovvero quello della presenza di personale multilingue, fa intendere che l'offerta turistica del luogo è improntata verso l'accoglienza di un turismo di origine estera. A supporto di questa tesi si hanno anche i dati forniti dal sito web della Regione Toscana (consultabili nella tabella 1).

## **4.2 Visualizzazione dei dati ottenuti**

Nelle prossime sezioni verranno elencate le metodologie di visualizzazione dei dati ottenuti. Essi sono stati inclusi in un sito web che, grazie a varie tecnologie di visualizzazione, ne rende maggiormente fruibile il contenuto.

La prima parte sarà dedicata all'illustrazione di come sono state visualizzate geograficamente le strutture, nella seconda della disposizione delle informazioni relative a prezzo, punteggio e servizi.

### **4.2.1 Tecniche per la visualizzazione geografica delle strutture**

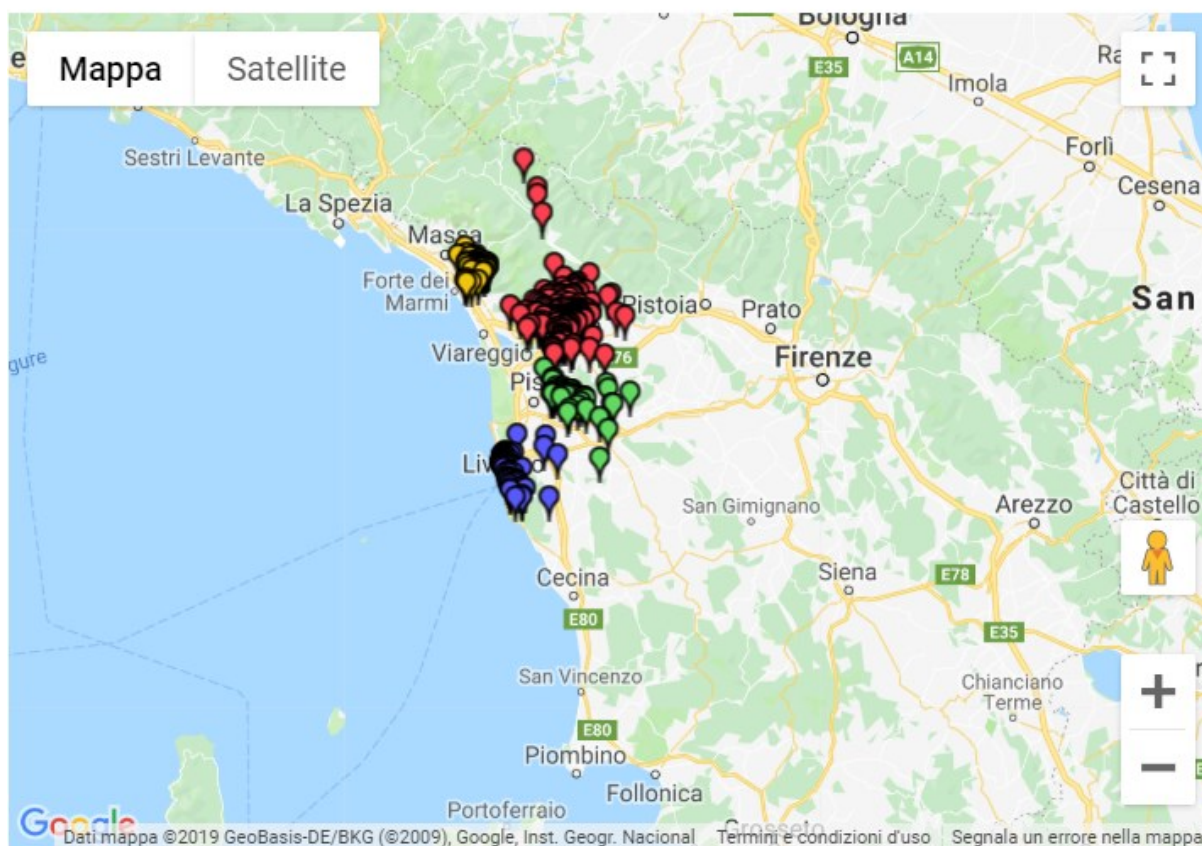
Per ottimizzare la visualizzazione geografica delle strutture, si è deciso di utilizzare la potenzialità offerta da Google Maps.

La compagnia leader nel settore delle mappe online ne consente il libero utilizzo, da parte di privati, grazie alla tecnologia delle Web API. Essa ci permette da un lato di poter delegare il costo computazionale a Google e dell'altro di poter operare significative personalizzazioni del servizio.



## Seleziona una zona

☒ Cascina ☒ Livorno ☒ Lucca ☒ Pietrasanta



*Figura 17. Screenshot ottenuto dal sito web realizzato per la visualizzazione dei risultati, cartina della zona di analisi.*

La figura 17 è uno screenshot ottenuto direttamente dal sito web realizzato per la visualizzazione dei dati. Esso ci mostra una mappa generata direttamente dal servizio Google Maps con impressi dei marcatori.

Essi rappresentano le posizioni geografiche delle strutture ottenute. La loro resa sulla mappa è stata realizzata grazie all'utilizzo del dato relativo alla longitudine e latitudine delle strutture. Le divisioni tra le quattro aree d'indagine sono state evidenziate grazie

all'utilizzo di colori diversi per i marcatori. È possibile selezionare le zone che si intendono visualizzare grazie all'utilizzo dei bottoni sovrastanti alla mappa, sempre come indicato nella figura.

#### 4.2.2 Tecniche per la visualizzazione dei dati delle strutture

Se si esegue un click su di uno dei marcatori illustrati nella figura 18, potremmo accedere all'andamento dei prezzi della singola struttura.

L'elenco di date, che si nota sotto il grafico, rappresenta la data e l'ora esatta di quando è stato ottenuto quel dato, che corrisponde a quando è stata effettuata l'acquisizione da parte del software di web scraping. Se si esegue un click sul nome della struttura, si verrà reindirizzati verso pagina di TripAdvisor dedicata ad essa.

Il grafico è stato realizzato dal software Highcharts, che massimizza la resa visiva dei dati.

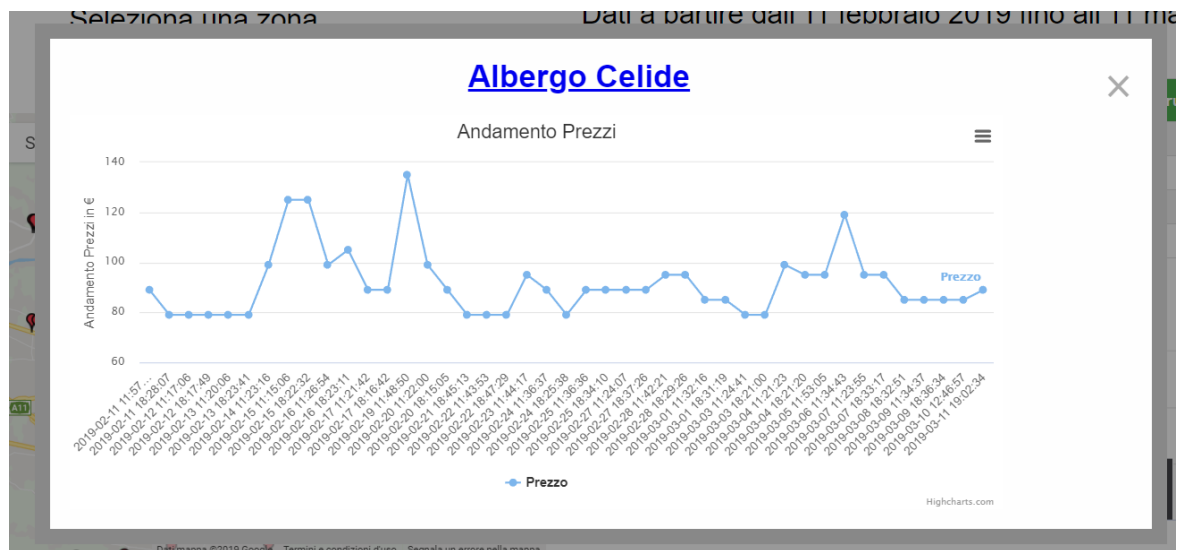


Figura 18. Screenshot ottenuto dal sito web realizzato per la visualizzazione dei risultati, dati riguardanti il prezzo di una struttura

Per la parte dei dati relativi ai servizi, si è scelto di visualizzare i dati in una tabella utilizzando il software Tabulator.

Esso rende la tabella interattiva, infatti è prevista un'azione per il click di una qualsiasi delle celle. La tabella 4 è stata ottenuta grazie ad uno screenshot rappresentante un frammento della stessa che mostra la realizzazione di quanto affermato.

Nome Servizio ▲	Lucca ▲	Livorno ▲	Cascina ▲	Pietrasanta
Piscina	15.4	11.6	17.5	16.8
Internet ad alta velocità gratuito (WiFi)	93.6	92.7	83.9	90.2
Parcheggio gratis	37.8	37.9	48.8	40.8
Colazione disponibile	8.4	6.5	5.0	11.1
Concierge	8.3	16.9	8.4	14.8
Hotel per non fumatori	49.7	45.1	40.0	53.8
Personale multilingue	18.9	17.0	13.2	14.8
Piscina all'aperto	7.4	7.7	3.2	9.3
Servizio lavanderia	14.5	20.1	17.9	18.6

*Tabella 4. Screenshot ottenuto dal sito web realizzato per la visualizzazione dei risultati, distribuzione dei servizi nelle 4 aree di analisi.*

## 5. Conclusioni

Nel corso di questa tesi si è studiata l'importanza, per lo sviluppo del settore turistico, della presenza di dati costantemente aggiornati riguardanti tale offerta. (sezione 2.1).

In questo contesto è stata giustificata la scelta di focalizzare l'indagine su quattro specifiche aree d'interesse e di utilizzare il sito web di viaggi TripAdvisor come fonte di essi (sezione 2.2).

È stato quindi realizzato un software di web scraping ed eseguito durante il lasso di tempo di un mese su di un dispositivo di tipologia single board computer appositamente configurato (capitolo 3).

Sono state elaborate tecniche di visualizzazione efficaci al fine di poter interpretare correttamente i dati ottenuti (sezione 4.2).

Le zone d'interesse sono poi state associate a diverse tipologie di clientela utilizzando, come parametri di confronto, le categorie descritte nel testo "marketing del turismo" (Philip Kotler, John Bowen, James Makens, 1996, p.361) (sezione 4.1).

Ogni aspetto tecnico della tesi è stato concepito - e successivamente realizzato - dalle risorse personali del candidato a dimostrazione che un'indagine di questo tipo oggi è possibile anche senza l'ausilio di strumentazione professionale. Questo lo si deve da un lato al software disponibile - interamente gratuito e open source - e dall'altro alla qualità, in relazione al prezzo, del single board computer utilizzato.

## 5.1 Sviluppi futuri

Il lavoro compiuto per questa indagine offre numerose possibilità di implementazione.

In primo luogo, riguardo il software realizzato per la raccolta dati, potrebbe risultare utile perfezionare la sua efficacia dal punto di vista della tolleranza ai cambiamenti interni del sito web che si decide di analizzare, possibile grazie all'applicazione di nuove tecniche di web scraping.

Dal lato hardware, il mercato dei single board computer è in costante evoluzione e la potenza computazionale di questi ultimi è destinata a crescere.

Utilizzare un dispositivo con capacità di calcolo maggiore incrementerebbe significativamente l'efficacia del processo di web scraping e semplificherebbe la realizzazione dello stesso da parte dell'utente.

## 6. Bibliografia

Antonio Tajani, 2011, 'Prefazione'. In Antonioli Corigliano M. e Baggio R. (a cura di), *Internet e Turismo 2.0*, Egea

Philip Kotler, John Bowen, James Makens, *marketing del turismo*, McGraw-Hill, 1996

## 7. Sitografia

Treccani, voce Turismo

(<http://www.treccani.it/enciclopedia/turismo>)

Webopedia, voce Web Scraping

([https://www.webopedia.com/TERM/W/Web\\_Scraping.html](https://www.webopedia.com/TERM/W/Web_Scraping.html))

Python Software Fondation, History of the software

(<https://docs.python.org/3/license.html>)

TripAdvisor, A proposito di TripAdvisor

(<https://tripadvisor.mediaroom.com/IT-about-us>)

Selenium, Homepage

(<https://www.seleniumhq.org/>)

Raspberry, Raspberry Pi Model B+

(<https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/>)

Raspbian, Homepage

(<https://www.raspbian.org/>)

RasPi.TV, How Much Power Does Raspberry Pi 3B+ Use? Power Measurements

(<https://raspi.tv/2018/how-much-power-does-raspberry-pi-3b-use-power-measurements>)

Navigaweb.net, Quanto consuma un pc e di quanta energia ha bisogno?

(<https://www.navigaweb.net/2012/11/quanto-consuma-un-pc.html>)