

Stelle Michelin e TripAdvisor

Valutazioni ai migliori ristoranti italiani

Università degli Studi Milano Bicocca, corso di Laurea Magistrale in Data Science.

Autori

Galimberti Dario, matricola n°826090, d.galimberti12@campus.unimib.it

Nardi Sara, matricola n° 876777, s.nardi1@campus.unimib.it

Stoffa Giacomo, matricola n°830159, g.stoffa1@campus.unimib.it

1 Introduzione

L'opinione dei clienti comuni è la stessa di quella degli esperti gastronomi? E gli strumenti a disposizione per effettuare la scelta relativa a dove sedersi a tavola sono soddisfacenti?

È da queste domande e dalla volontà di indagare il panorama della ristorazione italiana che è nato questo progetto.

Sul Il Corriere della Sera Danilo Taino scrisse: “Se la cucina è soft-power, l'Italia è di gran lunga il maggiore influencer a livello mondiale” e l'articolo chiudeva così: “Cosa sarebbe il mondo senza la cucina italiana?” [1]

Questo è soltanto un esempio per ribadire quanto in Italia si sostiene da sempre, ovvero che il cibo e in particolare la ristorazione sono un valore economico e culturale inestimabile per il Paese.

Pertanto, al fine di rispondere alle domande precedentemente formulate e analizzare la tematica citata, dapprima si sono acquisiti dati relativi a tutte le attività di ristorazione presenti nel nostro territorio, indipendentemente dalla loro tipologia e dalle loro caratteristiche e, in seguito, si sono effettuate delle operazioni di web scraping per l'ottenimento di informazioni relative a ristoranti e trattorie omaggiati di riconoscimenti (posizione in classifica 50 Top Italy e assegnazione della Stella Verde Michelin). L'analisi è proseguita integrando tali fonti, mediante l'acquisizione di dati relativi alle opinioni espresse dai clienti, in relazioni ai ristoranti e alle trattorie dotate dei riconoscimenti preliminarmente accennati. Tali informazioni si sono valutate in termini qualitativi e sono state inserite in un database, al fine della memorizzazione e di una futura consultazione.

Infine, si è convenuta la necessità di realizzare un dataset che riuscisse ad integrare le opinioni formulate dai clienti comuni con quelle formulate da critici esperti nelle classifiche succitate, in modo da fornire un'informazione completa ed esaustiva, per poter scegliere al meglio ed in modo oculato.

2 Sorgenti dati

2.1 TripAdvisor

Al fine di affrontare efficacemente la tematica proposta nel presente elaborato, si è ritenuto opportuno considerare, dapprima, i dati presenti su TripAdvisor.

Quest'ultimo è un sito web internazionale di recensioni di ristoranti, di hotel e di altre tipologie di strutture ricettive, fondato nel 2000 ed accessibile in modo completamente gratuito. In altre parole, esso offre la possibilità di consultare in qualsiasi momento e a costo zero le opinioni rilasciate da altri utenti. Inoltre, esso registra più di 830 milioni di recensioni e una media di 460 milioni di visitatori ogni mese[2].

Inoltre, TripAdvisor si è ritenuto un'ottima fonte dati per gli scopi trattati, in quanto esamina ben nr. 181.951 attività di ristorazione italiane a fronte delle nr. 196.000 attività presenti sul nostro territorio, secondo quanto riportato nel Rapporto Federazione Italiana dei Pubblici Esercizi (FIPE) 2021 (elaborato con le informazioni disponibili al 4 marzo 2022 e disponibile all'URL <https://www.fipe.it/wp-content/uploads/2022/03/Rapporto-Ristorazione-2021.pdf>).

Per poter fruire dei dati di TripAdvisor, si è quindi provveduto ad effettuare lo **scraping** dei ristoranti, attraverso la piattaforma online specializzata "Apify" [3]. In particolare, quest'ultima si serve di "attori", ossia di programmi cloud serverless che automatizzano tutto ciò che una persona può fare con un browser [4], quindi anche il processo secondo il quale un'applicazione estrae informazioni di valore da un sito web (definizione di web scraping).

A questo punto, si è ritenuto necessario attuare operazioni di **Data Preprocessing** del dataset precedentemente ottenuto. Tali operazioni si sono concretizzate in vari tipi di elaborazione eseguiti sui dati grezzi, al fine di prepararli ad un successivo utilizzo. In particolare, il data frame è stato opportunamente pulito da attività di ristorazione con un numero di recensioni inferiore a 5 e conseguente rating non valido e sono stati elisi quei campi considerati irrilevanti ai fini della ricerca, quali:

- "hours/1/3/close","hours/1/3/open" e similari;
- "type", in quanto sempre pari a "ristoranti";
- "rankingDenominator";
- "rankingPosition";
- "rankingString";

Inoltre, sono state rimosse le variabili ritenute ridondanti:

- "category"
- "reviews count"

Successivamente, le *features* aventi lo stesso contenuto informativo sono state poste in un'unica colonna e, conseguentemente, si è provveduto all'eliminazione delle singole. Tale operazione è stata svolta per i campi relativi alle diverse tipologie di cucina ("cuisine" la nuova variabile generata), alle svariate tipologie di dieta ("dietaryRestrictions" il nuovo campo creato) e ai certificati di eccellenza ("Certificati_Eccellenza" la nuova features realizzata).

Inoltre, si è provveduto a ridefinire la variabile "priceLevel", in modo da renderla maggiormente comprensibile e si è ritenuto opportuno creare la variabile "Comuni", al fine di aumentare il contenuto informativo del presente dataset.

Infine, è stata attuata una ridefinizione sia dell'ordine dei campi del *data frame*, sia dell'indice di quest'ultimo.

Pertanto, si è ottenuto un dataset composto da nr. 181.951 record e nr. 16 colonne, a fronte delle nr. 120 colonne presenti nel dataset generato dal processo di web scraping.

Al fine di studiare in modo puntuale i dati precedentemente estratti e ripuliti, si è svolta **un'analisi esplorativa** degli stessi, servendosi di due tipologie di grafico: il Bar Plot e la Heat Map.

In particolare, con riferimento al primo grafico citato, si è analizzata la distribuzione delle attività di ristorazione in Italia, suddividendo le regioni della stessa nei tre gruppi convenzionali: Nord, Centro e Sud. Specificatamente, il Nord Italia comprende Valle d'Aosta, Liguria, Lombardia, Trentino-Alto Adige, Friuli-Venezia Giulia, Veneto e Emilia-Romagna, il Centro racchiude Toscana, Lazio, Umbria e Marche, mentre il Sud contiene Abruzzo, Molise, Campania, Basilicata, Puglia, Calabria, Sicilia e Sardegna [5]. Tale ripartizione geografica è osservabile nella *Figura 1*. In particolare, il grafico proposto consente di affermare che in Italia la maggior quantità di attività di ristoro è rinvenibile nel Mezzogiorno, come mostrato dalla barra rossa in figura, mentre la minor concentrazione è nel Centro Italia (barra verde).

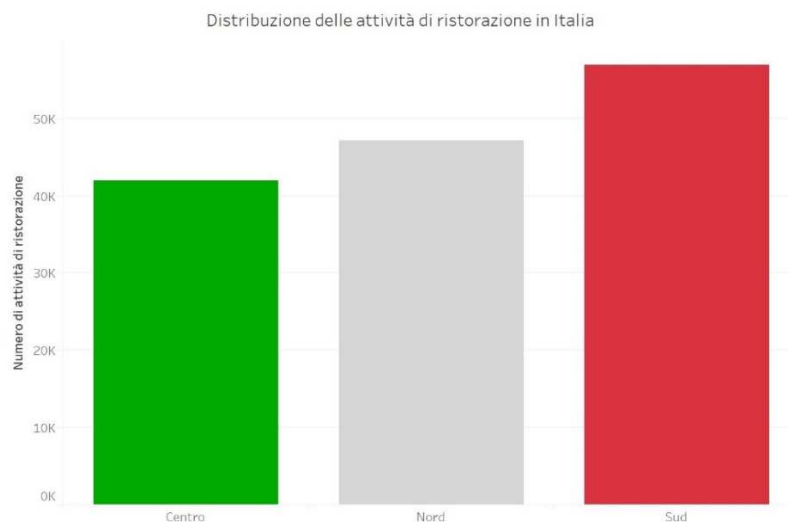


Figura 1: Distribuzione delle attività di ristoro, secondo la ripartizione delle regioni nei tre gruppi convenzionali (Nord, Centro, Sud).

La seconda tipologia di infografica consente anch'essa di geolocalizzare le attività di ristorazione in Italia, basandosi su quelle recensite da TripAdvisor, ma a differenza del primo grafico, permette di osservare agevolmente anche la concentrazione di esse in aree di piccole dimensioni, come le città e, in generale quindi, offre un maggior grado di dettaglio. Inoltre, al fine di aumentare l'informatività del grafico, si è posta in evidenza l'attività di ristorazione con il maggior numero di recensioni: "Tonnarello" a Roma.



Figura 2: Heat Map di tutte le attività di ristorazione presenti su TripAdvisor

Infine, si è ritenuto opportuno creare una seconda Heat Map interattiva che potesse mostrare la concentrazione di attività di ristoro di fascia di prezzo "Esclusiva". Dall'osservazione del grafico, è emerso che queste ultime si concentrano nei capoluoghi di regione, principalmente del Nord e del Centro Italia e in quelle località particolarmente dedite al turismo.

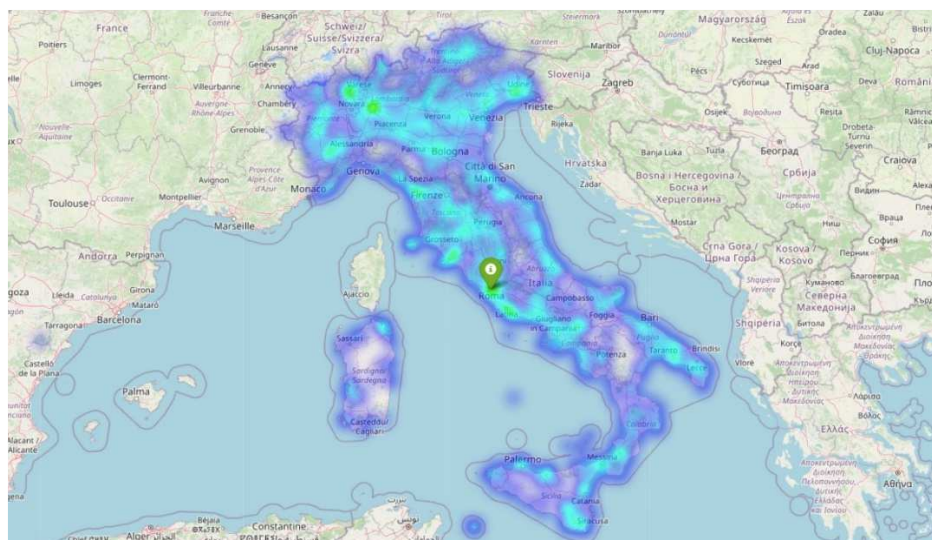


Figura 3: Heat Map di tutte le attività di ristorazione con fascia di prezzo esclusiva presenti su TripAdvisor.

2.2 Classifica 50 Top Italy - Ristoranti e trattorie top 50 (2020-2022)

Successivamente, si è ritenuto di fondamentale importanza, per gli scopi perseguiti, ottenere i dati relativi ai ristoranti e alle trattorie presenti nelle classifiche 50 Top Italy, stilate dal 2020 al 2022 [6][7][8]. Infatti, queste ultime sono delle guide pensate sia per gli italiani sia per i turisti amanti della buona tavola. In particolare, esse sono affidate a 150 esperti gastronomi, ai quali viene richiesto di effettuare visite in anonimato in ogni ristorante e trattoria potenzialmente includibile nelle suddette classifiche [9].

Pertanto, le informazioni scaturenti da tali dati si ritengono peculiari, in quanto consentono di analizzare la tematica della ristorazione dal punto di vista di ispettori esperti.

A tal proposito, a causa della mancata possibilità di scaricare i dati mediante pulsante di download e API, si è provveduto ad effettuare l'operazione di **web scraping** delle classifiche 50 Top Italy 2020, 2021 e 2022, rispettivamente per i ristoranti e per le trattorie italiani.

In particolare, il processo di estrazione di informazioni di valore dai siti web relativi alle diverse classifiche 50 Top Italy è stato effettuato utilizzando la *libreria Selenium* di Python, ossia una libreria di browser automation ("Chrome" nel caso in esame).

A titolo esemplificativo, le figure sottostanti mostrano il codice sorgente di uno dei siti web del quale è stato effettuato il web scraping (Figura 4) e il codice Python per l'effettiva raccolta dei dati desiderati (Figura 5).

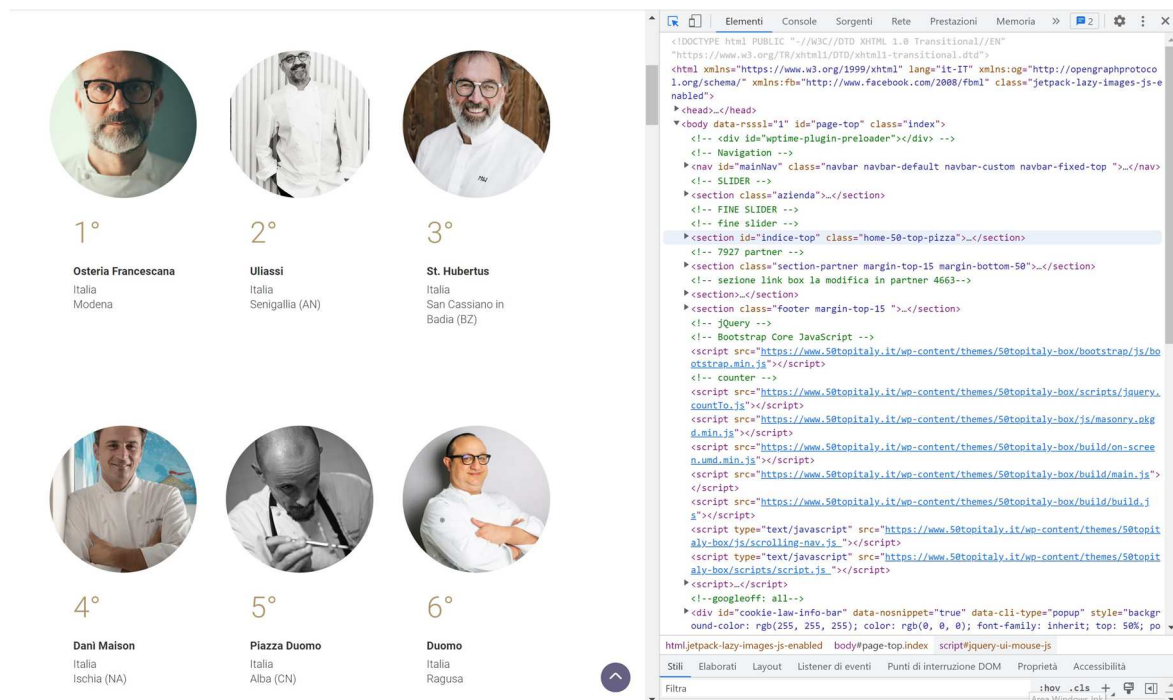


Figura 4: Codice sorgente di uno dei siti web utilizzati

▼ Scraping *RISTORANTI*

▼ Top 50 *ristoranti* anno 2020

```
[ ] # Scraping ristoranti top 50, anno 2020
search_url= 'https://www.50topitaly.it/it/50-top-italy-ristoranti-oltre-120-2020/'
driver.get(search_url)

tutti_ristoranti_2020 = driver.find_elements(By.XPATH, '//*[@id="indice-top"]/div/div[2]/div')

for value in tutti_ristoranti_2020:
    print(value.text)
```

Figura 5: Codice Python di web scraping.

In seguito all’ottenimento dei dati richiesti, si è reso necessario attuare un’attività di **Data Preprocessing**. In particolare, con riferimento ai due file .csv relativi ai ristoranti e alle trattorie presenti nella classifica 50 Top Italy del 2020, si è ridefinito il campo “città”, separando le città dalle relative province e creando opportunamente un nuovo campo, denominato “Provincia”. Successivamente, si è provveduto alla realizzazione del campo “Regione”, in modo da avere una certa similarità con i data frame relativi alle classifiche 50 Top Italy del 2021 e del 2022 e una corrispondenza con il campo “Regione” presente nel file .csv analizzato nella sezione 2.1 del presente elaborato e riguardante i dati ricavati dal sito di TripAdvisor.

Inoltre, per quanto concerne i *data frame* inerenti alle classifiche di ristoranti e trattorie Top50 Italy 2021 e 2022, oltre alla riorganizzazione dei campi “Città” e “Regione”, si è creato il campo “Nazione”. Infine, si sono ridefiniti l’ordine dei campi e l’indice dei vari dataset.

In seguito, i dataset così ripuliti sono stati concatenati, in modo da formare un unico dataset relativo a tutti i ristoranti e trattorie presenti nelle classifiche 50 Top Italy, anno 2020,2021 e 2022.

2.3 Stella Verde Michelin

Lo studio e l’analisi dei temi trattati non poteva che proseguire ponendo l’attenzione sui ristoranti premiati dalla Guida Michelin, che rappresenta uno dei maggiori riferimenti mondiali per la valutazione della qualità dei ristoranti e alberghi. In particolare, si è ritenuto interessante focalizzarsi su un riconoscimento assegnato dalla stessa ai ristoranti e alle trattorie presenti nelle proprie selezioni, che si sono distinti per il loro impegno in fatto di sostenibilità ambientale: la **“Stella Verde Michelin”**. Infatti, si ritiene che l’analisi di una tematica come la ristorazione, non possa prescindere dall’adottare un occhio di riguardo all’ambiente.

In particolare, la Stella Verde Michelin è destinata a quei ristoranti che prestano maggior attenzione all'utilizzo di prodotti biologici, locali e stagionali, allo smaltimento dei rifiuti a basso impatto per il pianeta e alla riduzione al minimo di sprechi e di utilizzo di materiali non riciclabili, oltre alla partecipazione attiva ad iniziative in campo ambientale [10].

Pertanto, si è scelto di selezionare i ristoranti e le trattorie italiane che hanno conseguito questo importante riconoscimento nell'anno in corso, acquisendo i dati desiderati mediante il processo di **web scraping**, di cui alla *Figura 6*.

Il risultato di tale operazione ha prodotto un dataset contenente informazioni riguardanti i 31 ristoranti premiati, con il nome di ogni ristorante e rispettivo comune di ubicazione.

Al fine di aumentare il contenuto informativo del data frame e di utilizzare lo stesso per scopi futuri, si è ritenuto necessario aggiungere il campo "Regione". Quest'ultimo è stato ottenuto mediante tecniche di web scraping, associando al comune in cui sono collocati i ristoranti premiati, la rispettiva regione [11].

2.4 Recensioni TripAdvisor di ristoranti e trattorie 50 Top Italy e Stella Verde Michelin.

Conformemente agli obiettivi opportunamente esplicitati nell'introduzione della presente relazione, si è deciso di integrare le informazioni fornite dal sito di TripAdvisor con quelle ottenute dall'acquisizione dei dati relativi alla classifica 50 Top Italy e inerenti alla Stella Verde Michelin. A tale scopo, si sono selezionate le recensioni redatte dagli utenti sul sito di TripAdvisor, sia per i ristoranti e le trattorie presenti nelle rispettive classifiche 50 Top Italy (2022), sia per quelli che hanno ricevuto l'assegnazione della Stella Verde Michelin nell'anno corrente.

La metodologia utilizzata per tale integrazione è stata la stessa per entrambe le fonti dato, 50 Top Italy e Stella Verde Michelin. Pertanto, al fine di evitare inutili ripetizioni e rendere quindi maggiormente piacevole e fluida la lettura, nella presente trattazione viene spiegato il procedimento attuato soltanto per i ristoranti presenti nella classifica 50 Top Italy (2022).

Innanzitutto, si è effettuata la ricerca dei ristoranti presenti nella classifica top50 (2022) all'interno del data frame relativo a tutta la ristorazione italiana indicata da TripAdvisor (tot_regioni), mediante l'associazione del link TripAdvisor di ciascun ristorante in classifica. In particolare, la corrispondenza si è effettuata per "regione" e per "nome" del ristorante, avendo preliminarmente eseguito le opportune operazioni di Data Preprocessing.

Tuttavia, in alcuni casi è accaduto che il nome del ristorante in classifica fosse indicato con una denominazione diversa nel dataset "tot_regioni" e questo ha comportato che la ricerca della corrispondenza venisse effettuata utilizzando un metodo alternativo: la *distanza di Levenshtein*.

Quest'ultima è stata scelta in quanto è una misura che esprime la differenza tra due stringhe, in base alle differenze rinvenute nei rispettivi caratteri. A questo punto, si è creato un dizionario avente come chiave, il nome del ristorante e come valore, il link TripAdvisor associato a tale ristorante, adoperando un controllo supervisionato dello stesso, in modo da scongiurare un'errata associazione del ristorante e del rispettivo link TripAdvisor.

In seguito alla ricerca di cui sopra, è divenuto possibile effettuare il web scraping di tutte le recensioni TripAdvisor associate ai ristoranti presenti nella classifica 50 Top Italy dell'anno 2022.

Tale processo ha reso possibile l'ottenimento di informazioni, quali:

- ◆ Data: giorno in cui è stata fatta la recensione
- ◆ Rating: variabile categorica indicante il voto dato all'esperienza nel ristorante, la quale assume valori 10,20,30,40 e 50.
- ◆ Titolo: variabile testuale che indica il titolo della recensione
- ◆ Testo: variabile testuale che contiene il testo della recensione

Il web scraping che ha condotto all'ottenimento del data frame di cui alla *Figura 6* è avvenuto tenendo presente che ogni pagina TripAdvisor contiene circa dieci recensioni e che, quindi, una volta terminate, si rendeva necessaria una variazione automatica dell'url del sito, per cercare informazioni nella pagina successiva. Benché TripAdvisor indichi il numero di recensioni relativo ad ogni ristorante, si è potuto constatare che non sempre quest'ultimo coincide con quelle realmente presenti. Pertanto, al fine di arginare il più possibile tale criticità, si è deciso di aggiungere alcuni parametri nel codice Python che eliminassero di volta in volta un numero di pagine proporzionale al numero totale previsto di esse.

	Date	Ratings	Title	Text	Name	Position
0	2022-08-25	10	Delusione totale	Onestamente se questo è stato il miglior rist...	Osteria Francescana	1
1	2022-08-01	50	Eccellente	Finalmente siamo riuscite a prenotare presso l...	Osteria Francescana	1
2	2022-08-01	50	Un viaggio straordinario	Era da tempo che desideravo provare questa esp...	Osteria Francescana	1
3	2022-07-12	50	Perfezione	Perfezione - L'Osteria Francescana non ha biso...	Osteria Francescana	1
4	2022-06-28	20	Grandissima delusione	Abbiamo atteso così tanti mesi per poter cenar...	Osteria Francescana	1
...
38508	2015-08-13	50	Superbo ristorante situato vicino al Castello	Questo ristorante è un must se visitate Bari. ...	Pashà Ristorante	50
38509	2015-08-05	50	QUESTO È da asporto	Non avrei mai trovato questo posto se i miei a...	Pashà Ristorante	50
38510	2015-05-10	40	Trattoria - non il ristorante	Pasha dispone di un ristorante elegante e una ...	Pashà Ristorante	50
38511	2015-04-25	40	Eccellente	Siamo stati accolti calorosamente il ristorante...	Pashà Ristorante	50
38512	2015-04-24	50	PERFEZIONE!!!!	Non c'è niente di meglio in TUTTE LE CATEGORIE...	Pashà Ristorante	50

38513 rows × 6 columns

Figura 6: Data frame ottenuto dalle operazioni di web scraping.


```

print("Nome: "+name+" Recensioni: "+str(recensioni)+ " Num pagine previsto: "+str(num_page))

for j in range(0, num_page):
    h=url.find("Reviews")
    h=h+len("Reviews")

    # Cambiamento dell'url, corrispondente al cambiamento di pagina
    url=url[:h] + "-or"+str(j)+"0"+ url[h:]
    wd.get(url)

    # Apertura del file per salvare le recensioni
    csvFile = open(path_to_file, 'a', encoding="utf-8")
    csvWriter = csv.writer(csvFile)

    time.sleep(2)

    wd.find_element("xpath", "//span[@class='taLnk ulBlueLinks']")#.click()

    # Definizione del blocco recensioni per poi prendere ogni recensione singolarmente
    container=wd.find_elements("xpath", ".//div[@class='review-container']")

    for t in range(len(container)):
        title = container[t].find_element("xpath", ".//span[@class='noQuotes']").text
        date = container[t].find_element("xpath", ".//span[contains(@class, 'ratingDate')]").get_attribute("title")
        rating = container[t].find_element("xpath", ".//span[contains(@class, 'ui_bubble_rating bubble_')]").get_attribute("class").split("_")[3]
        review = container[t].find_element("xpath", ".//p[@class='partial_entry']").text.replace("\n", " ")

        csvWriter.writerow([date, rating, title, review,name,i])
    # Cambiamento di pagina
    print("Mi fermo a pag "+str(j+1))

    #wd.close()

```

Nome: Irina Trattoria Recensioni: 96 Num pagine previsto: 9
Mi fermo a pag 9
Nome: Botteghe Antiche Recensioni: 459 Num pagine previsto: 40

Figura 7: Codice Python di web scraping con utilizzo della funzione csv.Writer().

Successivamente all’annotazione delle recensioni in tempo reale su csv, mediante la funzione csv.Writer(), di cui alla *Figura 7*, si è ritenuto utile studiare come i dati ottenuti si presentavano e l’esistenza o meno di determinati legami tra le variabili prese in considerazione. A tal proposito, concordemente agli obiettivi prefissati, si è deciso di effettuare un’analisi esplorativa di tali data frame, servendosi del linguaggio Python affiancato alla piattaforma Tableau.

Innanzitutto, per quanto attiene alla classifica 50 Top Italy 2022 di ristoranti e trattorie sono stati realizzati rispettivamente due Scatter Plot, con l’obiettivo di indagare il legame tra il numero di recensioni di ogni ristorante/trattoria presente in classifica e la votazione media attribuita agli stessi. Infatti, si è ritenuto interessante domandarsi se il numero di recensioni influisse sulla media del voto, dato che esso è ordinariamente sinonimo di luogo conosciuto e attrattivo, nonché se l’opinione delle persone comuni corrispondesse al parere espresso dai critici, in base alle posizioni da essi assegnate. In riferimento alle suddette visualizzazioni, si ritiene doveroso precisare che sono state prese in considerazione le opinioni espresse dagli utenti di Trip Advisor dal 2020 all’anno corrente, in modo che l’indagine non fosse affetta da bias legato al diverso anno di apertura dei ristoranti.

Osservando i grafici sottostanti è possibile constatare che non è rinvenibile un legame lineare tra il rating medio e il numero delle recensioni e si dimostra come la classifica stilata dagli esperti gastronomi sia ininfluyente per gli utenti di recensioni di TripAdvisor.



Figura 8: Scatter Plot relativi alla relazione tra il numero di recensioni e la media del ranking per ristoranti e trattorie in 50 Top Italy.

In seguito, la Data Exploration dei due dataset relativi alla classifica 50 Top Italy 2022 è proseguita con la realizzazione di due Line Plot, che consentissero di indagare se le prime tre posizioni in classifica avessero un qualche tipo di legame con l'opinione delle persone comuni e come il parere delle stesse si fosse evoluto nel tempo, fino al 2022, anno di riferimento della graduatoria.

Al fine di migliorare l'interpretabilità e la fruibilità dei grafici realizzati, i colori delle linee rappresentate non sono stati scelti casualmente: l'oro è associato al ristorante/osteria che occupa la prima posizione in classifica, l'argento affibbiato al secondo classificato e il bronzo al terzo.

Inoltre, si è ritenuto opportuno scegliere come unità temporale l'anno, data la mancata garanzia di recensioni mensili.

Ponendo l'attenzione sui grafici citati in *Figura 9*, è possibile asserire nuovamente che l'opinione dei critici non è pienamente condivisa dai clienti. Tale affermazione è particolarmente evidente in relazione al caso dell'Osteria Francescana di Modena, la quale è posizionata nella prima posizione della classifica 50 Top Italy 2022, nonostante il calo considerevole del suo rating annuale dal 2020 ad oggi.

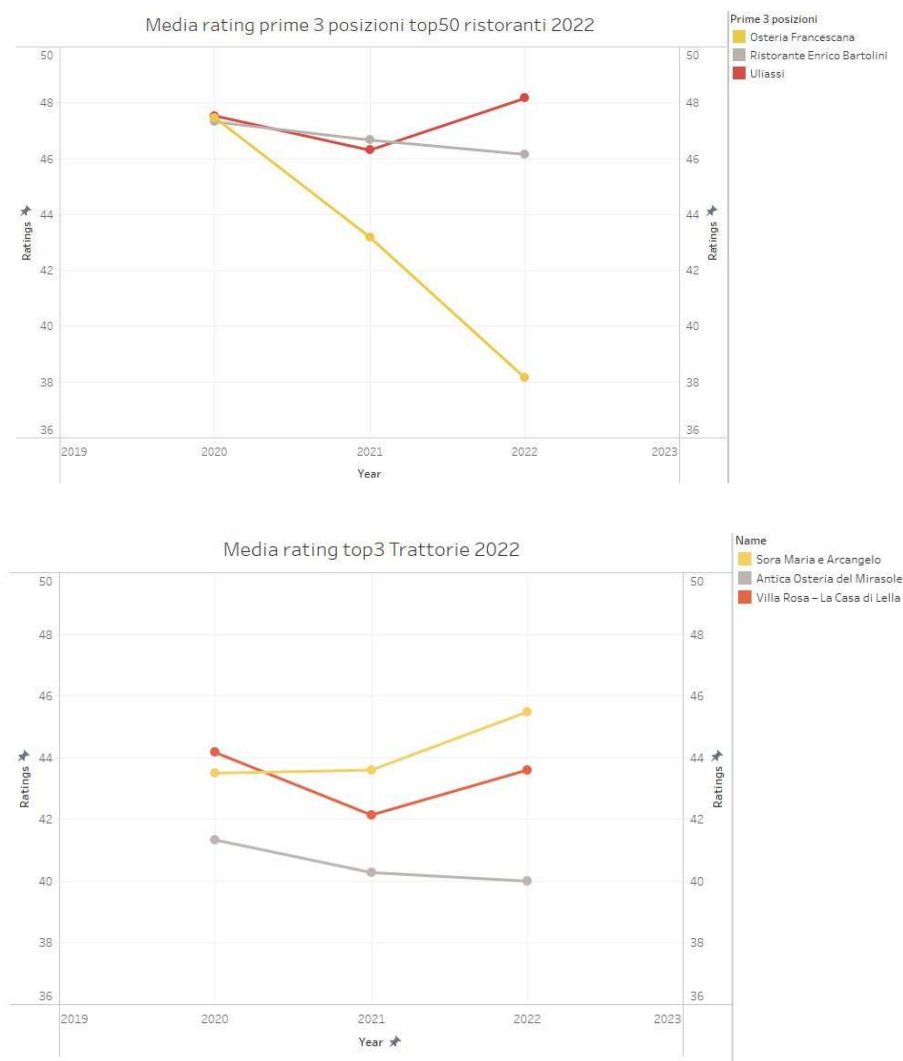


Figura 9: Line Plot sull'evoluzione del rating delle recensioni (dal 2020 al 2022) dei primi 3 classificati TOP50(2022).

Infine, si ritiene importante sottolineare che è stata svolta la medesima analisi esplorativa anche per i ristoranti e le trattorie ai quali è stata assegnata la Stella Verde Michelin 2022, eccetto nella realizzazione del Line Plot, in quanto tale riconoscimento non pone i ristoranti e le trattorie in

nessuna graduatoria. In altre parole, si è ritenuto opportuno creare soltanto lo Scatter Plot, anche se con qualche variazione rispetto a quelli creati per ristoranti e trattorie in classifica 50 Top Italy. Infatti, sebbene anch'esso consenta di indagare la presenza o meno di un legame tra il numero di recensioni e la media dei voti, per la sua caratteristica intrinseca di non collocare i ristoranti e le osterie in una classifica, non si cura di analizzare l'eventuale discrepanza di opinione tra critica e utenti delle recensioni. Nonostante le differenze sottolineate rispetto ai grafici in *Figura 8*, vale anche per la presente infografica quanto espresso in relazione all'esistenza o meno di un legame tra le variabili messe in relazione.

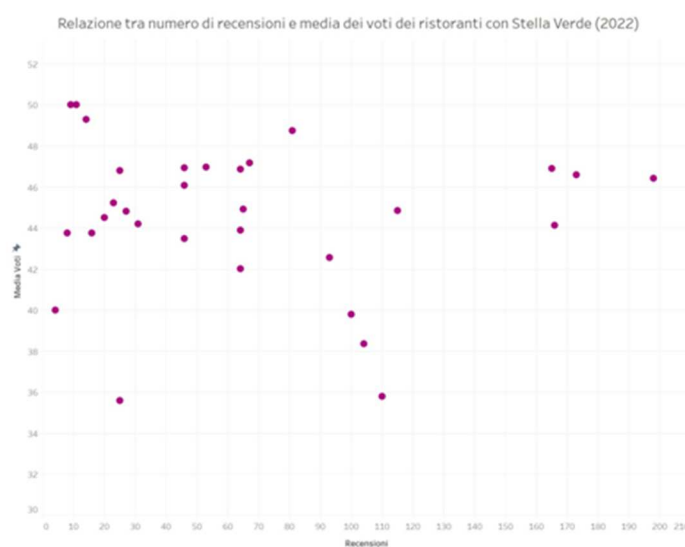


Figura 10: *Scatter Plot relativo alla relazione tra il numero di recensioni e la media dei voti per ristoranti e trattorie omaggiate di Stella Verde Michelin.*

Data Quality

In seguito all'ottenimento di tutti i dati desiderati per gli scopi prefissati nella presente ricerca, si è ritenuto fondamentale analizzare la qualità degli stessi. In particolare, il termine **Data Quality** identifica genericamente attività e processi volti all'analisi e all'eventuale miglioramento della qualità dei dati.

Inoltre, la Data Quality è comunemente considerata un requisito fondamentale per l'ottimizzazione della raccolta, dell'interpretazione e dell'analisi dei dati stessi.

Pertanto, avendo oltretutto già eseguito le fasi preliminari di gestione e pulizia dei dati grezzi, si sono sottoposti a tale valutazione i dataset interessati da un successivo caricamento nel database prescelto. Specificatamente, il dataset relativo alle attività di ristorazione italiane presenti su

TripAdvisor, quelli relativi ai ristoranti e alle trattorie presenti nelle classifiche 50 Top Italy 2022 e relative recensioni e, infine, quello inerente ai ristoranti che nel 2022 hanno ricevuto la Stella Verde Michelin e relative recensioni.

Ponendo l'attenzione su quanto svolto a livello qualitativo in riferimento ai dataset precedentemente citati, è possibile asserire che le misure di qualità trattate sono prevalentemente la *completezza*, la *consistenza* e l'*accuratezza*.

In letteratura, la completezza è definita come la misura di corrispondenza tra il mondo reale e il dataset specifico. Pertanto, essa si sostanzia nell'individuazione di quanti e quali dati mancano nel dataset, in modo da offrire una rappresentazione completa del contesto reale. Invece, la misura di consistenza attiene alla presenza o meno di duplicati, quindi, in generale, alla presenza di dati strutturati nello stesso modo. Infine, l'accuratezza fa riferimento al grado in cui i valori dei dati rappresentano i fatti del mondo reale (lontananza o vicinanza del valore stimato dai dati e il valore vero) o al grado in cui i valori dei dati rappresentano i valori di dominio [12].

La valutazione delle misure di qualità citate si è sostanziata nel controllo e nell'eliminazione dei duplicati, nonché nella verifica e nell'eventuale eliminazione di valori nulli, in ognuno dei dataset soggetti alla presente analisi di qualità. Per esempio, dalle analisi effettuate, è emerso che il dataset relativo alle attività di ristorazione italiane presenti su TripAdvisor presenta circa il 44.37% di valori mancanti nel campo "email", il 20.61% nella variabile "website", il 10.49% per il campo "PriceLevel" e il 6% per la *feature* denominata "phone". Tuttavia, al momento si è ritenuto opportuno non eliderli, procrastinando tale decisione in relazione a futuri sviluppi.

```
id          0.000000
name        0.000000
address     0.000000
latitude    0.574904
longitude    0.574904
Regione     0.000000
cuisine     0.000000
dietaryRestrictions 0.000000
rating      0.000000
Certificati_Eccellenza 0.000000
PriceLevel  10.490589
email       44.250112
numberOfReviews 0.000000
phone       6.009386
webUrl      0.000000
website     20.607885
dtype: float64
```

Figura 11: Percentuali di valori nulli presenti nel data frame relativo alle attività di ristorazione italiane presenti su TripAdvisor.

Inoltre, per il suddetto dataset, si è indagata la presenza o meno di tutti i possibili casi di dominio, per alcune variabili. Per esempio, si è notato che i dati relativi al campo “Price Level” non assumono tutti i possibili valori: soltanto [“Bassa”, “Medio-Alta”, “Esclusiva”], a fronte dei valori di dominio pari a [“Bassa”, “Medio- Bassa”, “Media”, “Medio-Alta”, “Alta”, “Esclusiva”].

Spostando il focus sui data frame relativi alle recensioni di ristoranti e osterie presenti nella classifica 50 Top Italy 2022 e quello caratterizzato dall’ottenimento della Stella Verde Michelin si è investigata la presenza di tutti i possibili voti nel campo “Rating”, l’effettiva buona riuscita della precedente acquisizione del testo delle recensioni e l’esistenza di queste ultime per tutti i ristoranti e le trattorie presi in considerazione.

In particolare, per quanto attiene le recensioni, si sono formulate delle analisi specifiche, con l’obiettivo di esaminare la presenza di “buchi temporali”, ossia archi temporali in cui tali opinioni risultano mancanti. Concordemente alle aspettative, l’esempio mostrato nella figura sottostante evidenzia che nel periodo interessato dal Covid 19 non è stata elaborata alcuna recensione per il ristorante analizzato.

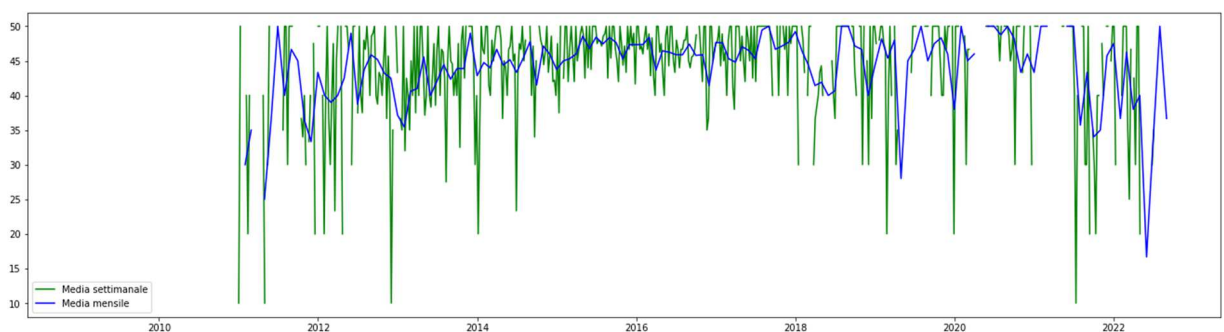


Figura 12: Andamento dei ratings del ristorante Osteria Francescana, effettuando dei raggruppamenti per settimana (verde) e per mese (blu).

Infine, si ritiene importante sottolineare che problemi di inconsistenza si sono osservati anche in fase di Preprocessing, in quanto, per esempio, il nome di alcuni ristoranti e trattorie nei csv relativi a 50 Top Italy e Stella Verde Michelin non risultava combaciare perfettamente con quello presente nel data frame relativo alle attività di ristoro esaminate.

3 DBMS

In seguito alle operazioni svolte preliminarmente relative all’acquisizione, alla pulizia e al controllo di qualità dei dati, si è reso necessario memorizzare i dati originati dalle stesse in un database.

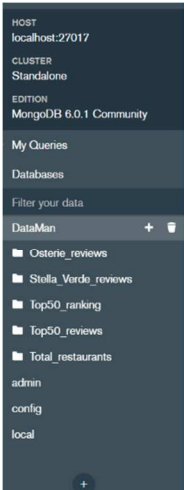
Nella sua definizione più generale, quest'ultimo è una raccolta organizzata di dati strutturati, in modo da renderli facilmente accessibili, gestibili e aggiornabili [13]. Inoltre, si ritiene opportuno precisare che, in base al modello dati più efficiente per gli scopi perseguiti, sia necessario scegliere il database che supporta al meglio tale modello.

Con riferimento al presente caso, si è deciso di affidarsi ad un particolare tipo di database: MongoDB. Tale scelta è stata influenzata dalla differente impostazione di dati ottenuti e dal fatto che MongoDB sia caratterizzato dall'essere né troppo complesso, né troppo semplice e quindi in grado di scalare abbastanza significativamente sul volume.

In particolare, MongoDB è un database non relazionale (NoSQL), che supporta modelli documentali, quindi csv convertiti in formato JSON e organizzati in diverse collezioni all'interno del database stesso. JSON è la notazione dalla quale si sviluppa il formato di salvataggio dei documenti in MongoDB (BSON) e le collezioni sono un insieme di documenti, i quali condividono informazioni simili tra loro.

Aumentando il grado di dettaglio, è possibile asserire che il database è stato utilizzato in locale (porta nr. 27017) e l'interazione con lo stesso è avvenuta servendosi della libreria Python denominata Pymongo.

Pertanto, i csv relativi alle recensioni e quello inerente alle attività di ristorazione indicate da TripAdvisor sono stati memorizzati nel database MongoDB denominato "DataMan", convertiti in formato JSON e i nomi degli stessi si sono utilizzati per attribuire una denominazione ad ogni collezione del database (Figura 13).



Collection Name	Storage size	Documents	Avg. document size	Indexes	Total index size
Total_restaurants	46.68 MB	182 K	640.00 B	1	1.81 MB
Top50_reviews	8.59 MB	39 K	403.00 B	1	409.60 kB
Osterie_reviews	7.00 MB	34 K	383.00 B	1	352.26 kB
Stella_Verde_reviews	3.37 MB	16 K	412.00 B	1	167.94 kB
Top50_ranking	28.67 kB	300	185.00 B	1	20.48 kB

Figura 13: Database MongoDB denominato DataMan e relative collezioni.

In seguito al caricamento in MongoDB, si è ritenuto di peculiare importanza integrare le informazioni possedute, attraverso l'aggiunta di una variabile dummy nella collezione relativa alle attività di

ristorazione italiane presenti su TripAdvisor. Infatti, concordemente agli obiettivi di analisi del presente elaborato, il nuovo campo consente di affiancare l'opinione della gente comune con quella degli esperti, in quanto indica l'assegnazione o meno di ogni attività di ristoro alla classifica 50 Top Italy 2022 e della Stella Verde Michelin. Osservando la *Figura 14*, ad esempio, è possibile notare che Osteria Francescana ha ottenuto sia la Stella Verde Michelin, sia una posizione in classifica 50 Top Italy 2022.

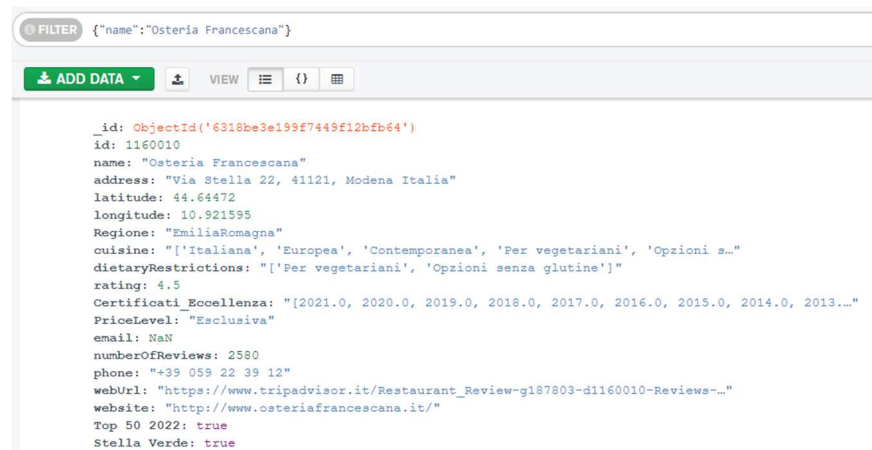


Figura 14: Esempio di integrazione delle fonti dato.

Infine, allo scopo di verificare il funzionamento del database creato e di sfruttare al meglio le opportunità offerte dallo stesso, si è proceduto all'effettuazione di alcune query. In particolare, si sono svolte due richieste al database:

- la prima relativa a quali ristoranti avessero ottenuto rispettivamente i singoli riconoscimenti nell'anno 2022 (ricoprire una posizione nella classifica 50 Top Italy o assegnazione della Stella Verde Michelin);
- la seconda inerente ai ristoranti ai quali avessero contemporaneamente attribuito sia un collocamento in classifica, sia il conferimento della Stella Verde Michelin.

La figura sottostante mostra il risultato della seconda query, opportunamente trasformato in un file csv, al fine di rendere maggiormente comprensibili i risultati ottenuti. L'interrogazione del database mostra che sono sette i ristoranti dotati di entrambi i riconoscimenti.

Stella Verde	OsterieTop 50 2022	Top 50 2022	name	address
true		true	D'O	Piazza della Chiesa, 14, 20010 Cornaredo
true	true		La crepa	Piazza Giacomo Matteotti 13, 26031 Isola Dovarese
true		true	Joia	Via Panfilo Castaldi 18, 20124 Milano
true		true	Ristorante Don Alfonso 1890	11 Corso Sant' Agata, 80061 Massa Lubrense
true		true	Osteria Francescana	Via Stella 22, 41121, Modena
true	true		Osteria di Suvereto da l'Ciocio	Piazza Dei Giudici 1, 57028 Suvereto
true		true	Ristorante Gourmet Virtuoso - Tenuta Le Tre Virtù	Vai di Lucigliano 13, 50038, Scarperia e San Piero

Figura 15: Csv del risultato di una query effettuata sul database.

4 Conclusioni

Alla luce dei risultati conseguiti con la presente indagine, è possibile affermare che gli obiettivi preposti in fase iniziale sono stati efficacemente soddisfatti.

Tuttavia, in un'ottica di miglioramento continuo, si ritiene importante suggerire alcuni sviluppi futuri. Per esempio, disponendo di dati testuali si potrebbero implementare svariati algoritmi di text analysis, come "text summarization", "sentiment analysis" o "topic modeling".

Inoltre, potrebbe essere particolarmente interessante raggruppare diversi locali con criteri più o meno oggettivi, creando svariate *wordcloud* o *network graph* interattivi, al fine di riconoscere le caratteristiche che più vengono scelte in relazione a un ristorante o a una certa categoria di ristoranti.

Sarebbe, altresì, interessante effettuare degli studi di approfondimento riguardanti il modo in cui le recensioni influenzano la pubblicità di un determinato locale.

Infine, si ritiene fondamentale sottolineare che qualsiasi siano gli sviluppi futuri implementati nel campo della ristorazione, sia fondamentale avere sempre un occhio di riguardo sull'ambiente!

5 Riferimenti bibliografici

- [1] https://www.corriere.it/opinioni/19_settembre_04/potere-d-influenza-della-cucina-italiana-a373a57a-cf2c-11e9-874e-4a9e2900aac3.shtml?refresh_ce
- [2] <https://it.wikipedia.org/wiki/Tripadvisor>
- [3] <https://docs.apify.com/about>
- [4] <https://docs.apify.com/actors#section-overview>
- [5] <https://www.tuttitalia.it/statistiche/nord-centro-mezzogiorno-italia/>
- [6] <https://www.50topitaly.it/it/50-top-italy-ristoranti-oltre-120-2020/>
- [7] <https://www.50topitaly.it/it/50-top-italy-ristoranti-oltre-120e-2021/>
- [8] <https://www.50topitaly.it/it/50-top-italy-grandi-ristoranti-2022/>
- [9] <https://www.50topitaly.it/it/i-nostri-principi/>
- [10] <https://guide.michelin.com/it/it/notizia/sustainable-gastronomy/la-stella-verde-michelin-per-un-pianeta-piu-sostenibile#:~:text=Introdotta%20nel%202020%2C%20la%20Stella,non%20riciclabili%20dalla%20Ioro%20filiera%2C>
- [11] <https://nominatim.openstreetmap.org/ui/search.html>
- [12] <https://vitolavecchia.altervista.org/data-quality-che-cose-e-come-si-misura-la-qualita-dei-dati/>
- [13] <https://www.studiosamo.it/glossario/database/>