TripAdvisor Restaurants Analysis

Matteo, Federica, Matteo, Alessia, Simone

20 Giugno 2022

1 Obiettivo del progetto

L'obiettivo del progetto è quello di proporre un'analisi sulle caratteristiche principali dei ristoranti di Roma, servendosi delle recensioni lasciate dai clienti sul sito web di TripAdvisor. Quest'ultimo presenta i nomi e le informazioni di 30 ristoranti su ogni pagina contenenti a loro volta 10 recensioni per foglio. Abbiamo raccolto i dati in una database SQL composto da tre tabelle:

- Informazioni sui **ristoranti**:
  - link ristorante;
  - nome;
  - rating medio;
  - totale recensioni;
  - fascia di prezzo;
  - tipo di cucina;
  - diete particolari;
  - indirizzo.
- Informazioni sul **recensore**:
  - username;
  - link del profilo;
  - recensioni totali scritte;
  - livello del recensore;
  - data di registrazione;
  - località di residenza;
  - totale città recensite;
  - totale voti utili ricevuti.
- Informazioni sulle **recensioni**:
  - link recensione;
  - link ristorante;
  - username del recensore;
  - titolo;
  - data della recensione;
  - data di visita del ristorante;
  - voto della recensione;
  - voti utili ricevuti;
  - dispositivo con cui è stata scritta la recensione;
  - testo della recensione.

2 Indicazioni per la realizzazione del progetto

- Scaricare le prime 200 recensioni in lingua italiana dei ristoranti presenti a Roma.
- Salvare i dati su un DataBase in maniera corretta e esaustiva.
- Processare le recensioni tramite tokenizzazione e applicazione delle stop words delle recensioni e salvare il risultato.
- Creare una Tag Cloud delle parole significative più ricorrenti.
- Costruire un grafico di occorrenza delle 20 parole più usate.
- Produrre un post creativo sui risultati ottenuti.
- Produrre un elaborato in pdf che descriva ogni aspetto dello sviluppo, delle scelte effettuate, dei problemi riscontrati, delle soluzioni avanzate e delle idee proposte.

3 Strumenti Utilizzati

- Python3
- MySQL
- Selenium
- Word Cloud
- GitHub

4 Scelte Effettuate

Per prima cosa abbiamo analizzato nel dettaglio il sito web in tutta la sua struttura avvalendoci degli strumenti del browser di analisi pagina.
Abbiamo utilizzato la libreria Selenium per elaborare gli script in Javascript che **impediscono** di visualizzare il codice sorgente della pagina. Inoltre, non è stato possibile utilizzare BeautifulSoup, poiché il nostro "scraper" doveva recuperare informazioni sui recensori selezionando l'avatar presente in ogni box delle recensioni. Nonostante ciò, il sito di TripAdvisor implementa degli script che tengono conto del percorso degli indirizzi attraversati per arrivare alla pagina di destinazione e, per questo, non vi è la possibilità di navigare direttamente alla pagina di interesse attraverso l'URL.
Per evitare i problemi sopra descritti e migliorare le prestazioni dello "scraper", abbiamo disabilitato JavaScrip e parallelizzato l'esecuzione del codice attraverso **concurrent futures**. In questo modo è stato possibile utilizzare contemporaneamente più "Chromedriver", ognuno dei quali ha analizzato una diversa pagina web di ristoranti o di recensioni. Non è stato possibile utilizzare questo miglioramento anche sulla raccolta dei dati sui recensori a causa del necessario utilizzo di JavaScript. Sono state necessarie circa 24h di computazione parallela per arrivare ad un totale di:

- $\sim$ 10k ristoranti (circa la totalità dei ristoranti di Roma, esclusi quelli senza recensioni);
- $\sim$ 250k recensori, comprendenti i campi descritti in precedenza;
- $\sim$ 280k recensioni, comprendenti il testo di ogni recensione e i campi decritti in precedenza.

Durante la fase di "scraping", abbiamo notato che, con 30 processi paralleli su una macchina con 6 core/12 threads e 32 GB di RAM, la connessione di 100 Mbps rappresentava il collo di bottiglia; infatti in poco più di 6 ore avevamo consumato $\sim$ 300 GB. Consigliamo di disattivare il download delle immagini sempre nelle impostazioni del Chromedriver.
In conclusione abbiamo progettato un poster creativo sui risultati ottenuti.

5 Problemi Riscontrati

Il sito di TripAdvisor contiene una serie di script in JavaScript che limitano la maggior parte dei tentativi di scraping che non tengono conto di alcune accortezze (nel nostro caso). Infatti:

1. ogni pagina acceduta attraverso il link diretto non è correttamente caricata e "renderizzata";
2. qualsiasi elemento della pagina sul quale bisogna fare un click rischia di non riceverlo: un altro elemento della pagina intercetta il click;
3. le date che presentano il nome del mese in lingua italiana non possono essere direttamente trattate dalla libreria *datetime* di python;
4. molti ristoranti presentano meno di 50 recensioni (italiane);

6 Soluzioni Proposte

Le soluzioni che abbiamo adottato per risolvere i problemi appena descritti sono, nell'ordine:

1. per raccogliere le informazioni dei ristoranti di una qualsiasi città, senza riscontrare anomalie nel codice html, bisogna passare dalla pagina principale di TripAdvisor. Quando decidiamo di procedere alla pagina successiva dei ristoranti non possiamo usare Beautiful Soup per prendere il link di quest'ultima, ma dobbiamo necessariamente premere sul bottone, altrimenti nel codice html vengono erroneamente scambiate o duplicate alcune informazioni.

2. Utilizzare Selenium per selezionare i bottoni, però, ci porta a un problema di posizione dell'elemento poiché se il tasto da dover premere non è visibile nella schermata renderizzata dal driver, oppure non è abbastanza lontana ($\sim$5 px) da altri bottoni, questi ultimi intercetteranno il click causando l'interruzione della sessione di Selenium. La soluzione al problema è semplice: si deve, di volta in volta, spostare la "visuale" centrandola sull'elemento scelto:

```
actionChains.move_to_element(element).perform()
```

3. Per trattare le date che presentano il nome del mese in lingua italiana nel formato *datetime* richiede la modifica delle impostazioni *locale* di python;

```
locale.setlocale(locale.LC_TIME, "it_IT")
```

4. Abbiamo deciso di fare una selezione dei ristoranti recensiti, filtrando quelli che presentano più di 50 recensioni in lingua italiana;

7 Risultati dell'analisi

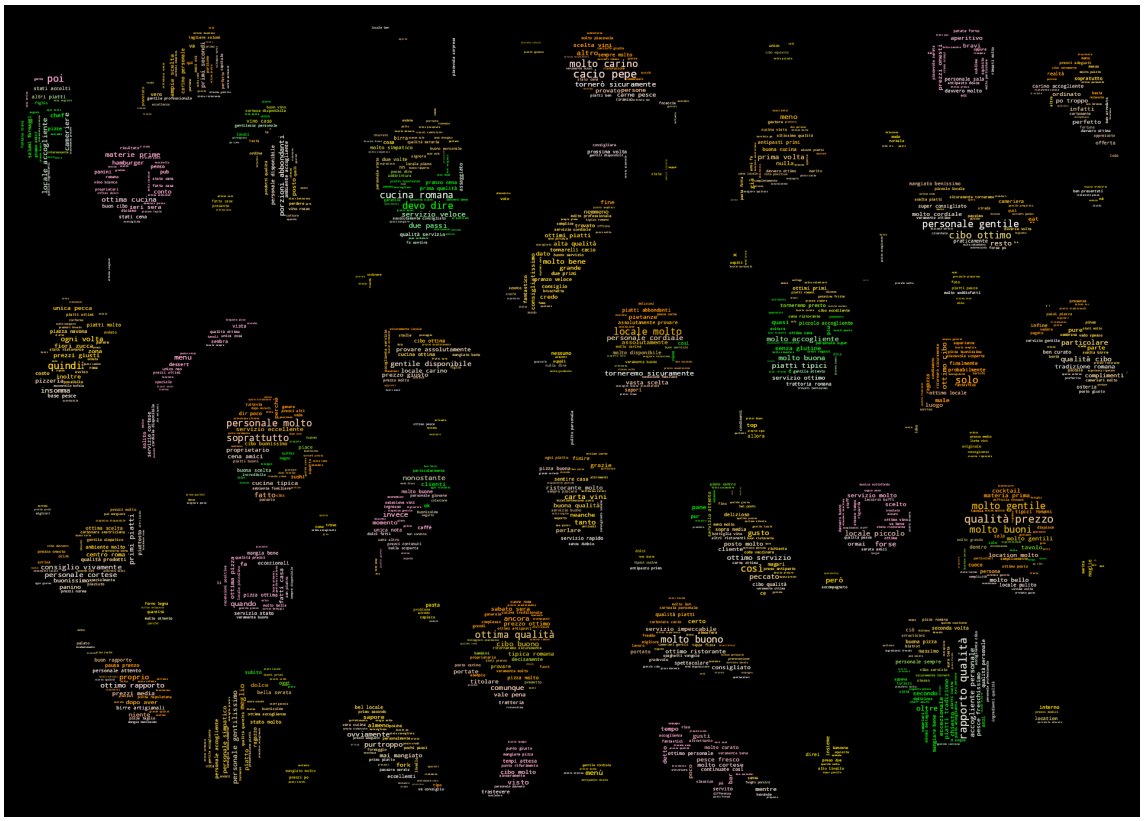

Figura 1: *WordCloud di tutte le recensioni*

Figura 2: *WordCloud dei ristoranti che hanno ricevuto più di 1000 recensioni*

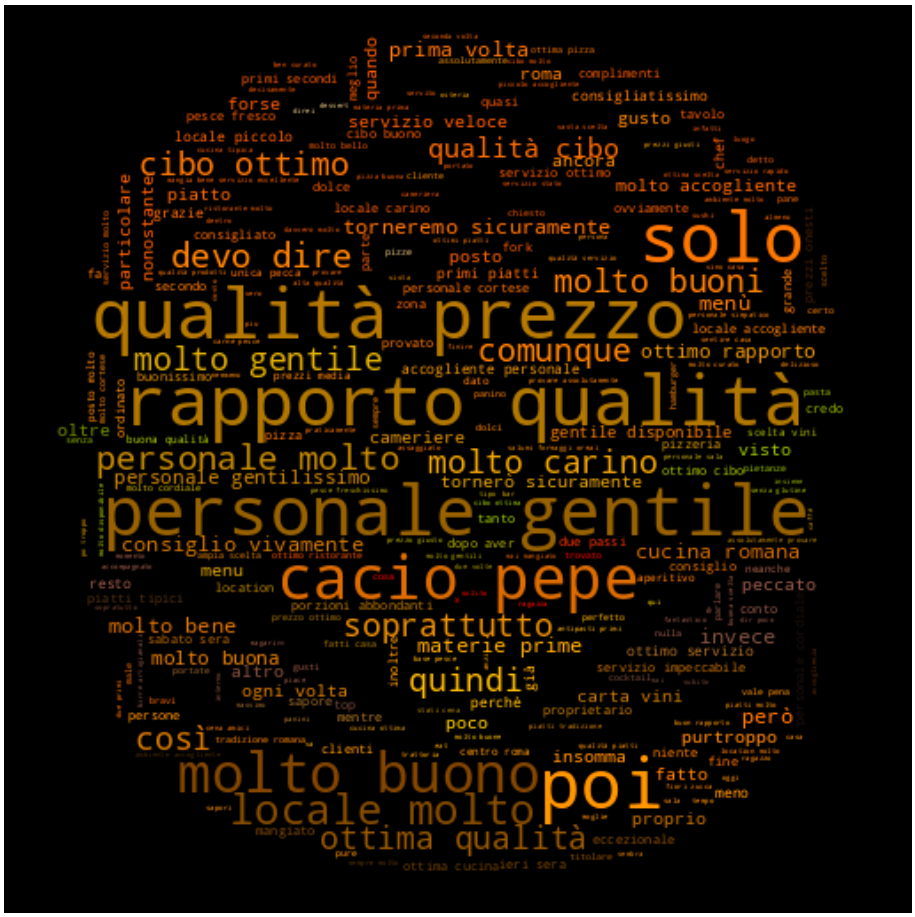

Figura 3: *WordCloud dei ristoranti che hanno ricevuto dalle 50 alle 1000 recensioni*

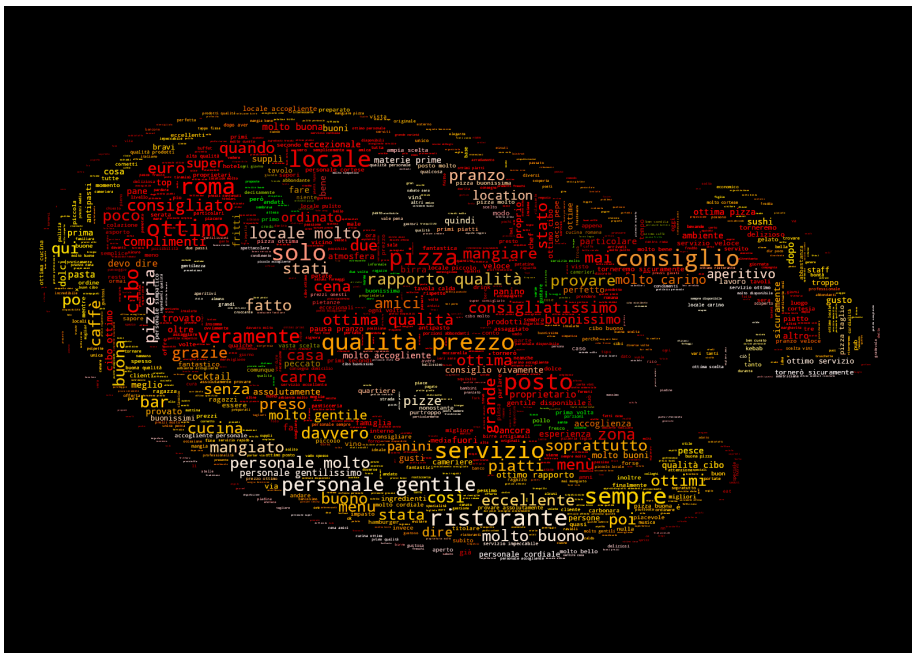

Figura 4: *WordCloud dei ristoranti con meno di 50 recensioni*