Relazione PyTripAdvisor

A,B,C,D,E

Giugno 2022

Indice

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. Vogliamo studiare la distribuzione dei ristoranti sul territorio romano, tenendo conto di:

- *informazioni sui ristoranti*

  url,
  nome del ristorante,
  totale recensioni,
  indirizzo del ristorante.

- *informazioni sul recensore*

  rating,
  data di visita,
  data della recensione,
  helpful,
  nome della recensione,
  città di provenienza del recensore,
  device utilizzato per recensire.

2 Indicazioni per la realizzazione del progetto

- Scaricare le prime 200 recensioni in lingua italiana dei ristoranti presenti a Roma.
- Salvare i risultati su un DataBase in maniera corretta e esaustiva.
- Processare le recensioni tramite tokenizzazione e applicazione delle stop words delle recensioni e salvare il risultato sul DB (Sempre in maniera completa ed esaustiva).
- Creare una Tag Cloud delle 25 e delle 50 parole significative pi'u ricorrenti (Tag Cloud per categoria di ristorante).
- Costruire un grafico di occorrenza delle 50 parole pi'u usate per ciascun ristorante e un grafico cumulativo
- (Eventualmente) Produrre un post creativo sui risultati ottenuti.
- Produrre un elaborato in LaTeX che descriva ogni aspetto dello sviluppo, delle scelte effettuate, dei problemi riscontrati, delle soluzioni avanzate e delle idee proposte.
- Produrre una heat-map delle zone di Roma più frequentate in base al sentiment di ogni ristorante.

3 Strumenti Utilizzati

- Python3
- SQLite3
- Regex
- Datetime
- Webdriver-manager
- Selenium
- Beautiful Soup

4 Scelte Effettuate

Per prima cosa abbiamo analizzato nel dettaglio il sito web in tutta la sua struttura. Abbiamo, poi, cominciato la raccolta di tutti i dati necessari per procedere con la nostra analisi. Abbiamo... Raccolte le informazioni abbiamo realizzato grafici e heat-map per mostrare i risultati ottenuti. In conclusione abbiamo progettato un poster creativo.

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 (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:

```
1             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;

```
1             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 Discussione sullo sviluppo

```
INSERT INTO `Utente` (`ID_Utente`, `Nome`, `Cognome`, `Codice_Fiscale`, `Email`,
↪  `Numero_di_Carta`, `Scadenza`, `Titolare`, `CAP`, `Via`, `Numero`, `Città`,
↪  `Feedback`, `Username`, `Pwd`, `Data_iscrizione`, `Data_ultimo_accesso`)
↪  VALUES (NULL, 'Janina', 'Engel', 'JNNNGL44R48A952M',
↪  'JaninaEngel@jourrapide.com', '4532497863930373', '2024-11-30', 'Janina
↪  Engel', '12081', 'Via Francesco Saverio Correra', '147', 'Beinette', NULL,
↪  'Hemed1979', SHA1('passwordEsempio2022!'), '2021-12-11', '2021-12-21
↪  10:34:10');
```