

Analisi di Yelp Open Dataset

Lorenzo Vainigli

Corso di Intelligenza Artificiale a.a. 2019/20

Laurea Magistrale in Informatica

Università di Bologna

1 Introduzione

I file di questo progetto sono disponibili nel repository dell'autore su GitHub [1].

2 Dati

Yelp Open Dataset [2] è una base di dati che raccoglie informazioni su esercizi commerciali di varie categorie. I dati sono utilizzabili per uso personale, educativo o accademico, sono disponibili in formato JSON e sono divisi in alcuni file:

- *business.json*: contiene le informazioni relative agli esercizi commerciali tra cui ubicazione e categoria.
- *review.json*: contiene i testi delle recensioni includendo l'identificativo dell'utente che ha scritto la recensione e l'esercizio commerciale oggetto della recensione.
- *user.json*: contiene i dati associati ai singoli utenti, inclusi gli identificativi degli amici.
- *tip.json*: contiene dei suggerimenti che gli utenti scrivono a proposito degli esercizi commerciali. Possono essere visti come delle brevissime recensioni.

Il database contiene anche i file *checkin.json* e *photos.json*, ma non sono stati presi in considerazione per lo sviluppo di questo progetto.

3 Obiettivi

Lo scopo del progetto prevede l'analisi dei dati al fine di studiare la loro struttura e il loro contenuto, al fine di estrapolare osservazioni interessanti su di essi. Non si tratta solo di aggregare record o trovare valori minimi, massimi o medi, ma di applicare anche tecniche di NLP e machine learning. In particolare, le finalità del progetto richiedono:

- T1) il riconoscimento automatico di una review positiva o negativa;
- T2) il raggruppamento degli utenti in base alle loro preferenze o comportamento sulla piattaforma;
- T3) il raggruppamento automatico dei locali in base a criteri di similitudine data una certa località.

A queste, ne sono state aggiunte altre:

- T4) analisi dei singoli file JSON;
- T5) classificazione dei locali migliori e peggiori per ogni categoria;
- T6) locali aperti nelle vicinanze dell'utente;
- T7) utenti con le recensioni più affidabili (comparando il loro voto alla media dei voti di un determinato locale);
- T8) migliori recensioni e suggerimenti (tips) per un locale.

4 Strumenti

Per conseguire gli obiettivi sopra citati i dati sono stati elaborati in Python con l'utilizzo di Jupyter Notebook [3].

5 Sviluppo

Per ogni obiettivo (o target) T^* è stato creato un notebook presente nella cartella `notebooks`:

- T1) `reviews_classification.ipynb`;

T2) `users_grouping.ipynb`;
T3) `businesses_grouping.ipynb`;
T4) `business.ipynb`, `review.ipynb`, `tip.ipynb`, `user.ipynb`;
T5) `best_and_worst_businesses.ipynb`;
T6) `closest_opened_businesses.ipynb`;
T7) `best_reviewers.ipynb`;
T8) `best_business_tips.ipynb`.

6 Risultati

6.1 Esercizi commerciali

6.1.1 Migliori e peggiori

Per questa classificazione sono state prese in considerazione il numero di stelle assegnate a ogni esercizio commerciale e il numero di recensioni ricevuto. Si presume che, a parità di stelle, più il numero di recensioni è alto, più questo valore sia affidabile.

Sono state analizzate quattro delle categorie più diffuse: *Restaurants*, *Shopping*, *Health & Medical* and *Automotive*.

Little Miss BBQ (Phoenix, AZ), **Brew Tea Bar** (Las Vegas, NV) e **Cocina Madrigal** (Phoenix, AZ) sono i migliori ristoranti secondo la media delle stelle e il numero di recensioni ricevute, mentre **McDonald's** (Las Vegas, NV), **KFC** (Avondale, AZ) e **McDonald's** (Fort Mill, SC) sono i peggiori. Tra i negozi catalogati come *Shopping*, i migliori sono **Eco-Tint** (Las Vegas, NV), **Studio 21 Tattoo Gallery** (Las Vegas, NV) e **FINO for MEN** (Las Vegas, NV). I peggiori sono **DIRECTV** (Phoenix, AZ), **Bank of America Store and Heritage Center** (Charlotte, NC) e **Teleflora Fresh Flowers** (Las Vegas, NV).

Bangkok Thai Spa Massage (Las Vegas, NV), **Simply Skin Las Vegas** (Las Vegas, NV) e **Richards Cosmetic Surgery, Med Spa & Laser Center** (Las Vegas, NV) sono i luoghi migliori per la categoria *Health & Medical*. Sempre per quanto riguarda questa categoria, i luoghi peggiori sono **SilverScript Medicare** (Phoenix, AZ), **Apria Healthcare** (Henderson,

NV) e **OptumCare Primary Care - Deer Valley** (Phoenix, AZ).

I migliori esercizi commerciali per *Automotive* sono **Eco-Tint** (Las Vegas, NV), **Precision Window Tint** (Henderson, NV) e **DC Auto Luxury Window Tinting** (Las Vegas, NV). I peggiori sono **Phoenix Car Rental** (Phoenix, AZ), **LendingTree** (Charlotte, NC) e **Seller Networks** (Las Vegas, NV).

Considerando le città, **Las Vegas** è quella dove si possono trovare gli esercizi commerciali migliori, considerando queste quattro categorie, seguita da **Phoenix**.

7 Conclusioni

Riferimenti

- [1] *Analysis of Yelp Open Dataset*. URL: <https://github.com/lorenzovngl/ai-project>.
- [2] *Yelp Open Dataset*. URL: <https://www.yelp.com/dataset/>.
- [3] *Jupyter Notebook*. URL: <https://jupyter.org/>.