NLP for aYelp Review Classification

CdLM Università di Bologna

Mario Sessa

Anno Accademico 2020/2021

Contents

Introduzione

1.1 Uso delle Recensioni

Oggigiorno tutte le persone si affidano a piattaforme di valutazione come **Yelp, The Fork, Trip Advisor** e altri per orientarsi sul ristorante da scegliere, su quale hotel prenotare o quale servizio sia il più affidabile per una consegna a domicilio.

Yelp è uno delle principali piattaforme in cui trovare recensioni su qualsiasi attività (*ristoranti, hotel, servizi di autonoleggio e altro*) in cui è possibile consultare recensioni e valutazioni al fine di condividere esperienze in grado di garantire un **feedback** che possa **orientare una propria decisione** nel visitare un particolare luogo, andare a mangiare in un ristorante o pernottare in un hotel per un viaggio.

Yelp ha semplificato queste decisioni andando a valutare a priori un servizio ancora prima di averne usufruito grazie alle valutazioni di utenti che ne hanno già usufruito.
Nello specifico, ognuno potrà consultare:

1. Un feedback numerico da **1** a **5 stelle** per una particolare attività
2. Un **testo** affiliata in cui si descrive la **propria esperienza** nell'utilizzo del servizio

In genere all'interno di una recensione vi possono essere errori lessicali o termini non propri della lingua che si sta utilizzando.

1.2 Obiettivi

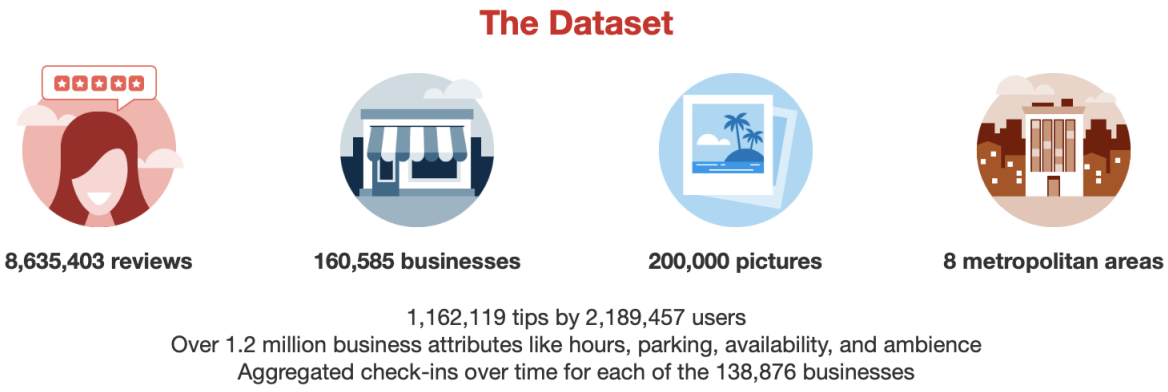
Gli obiettivi che ci andiamo a prefissare si basano principalmente sul natural language processing (NLP) di tali recensioni andando a riconoscere i pensieri e i sentimenti che un utente ha provato durante la creazione di una propria recensione e riuscire a classificarle come recensione negativa o positiva.

Tale obiettivo è molto difficile da raggiungere poichè **il metro di valutazione di ogni persona è differente**, due persone che abbiano scritto la stessa esperienza possono inserire un numero differente di stelle poichè potrebbero dare più **peso verso aspetti differenti** o semplicemente hanno un **diverso metro di giudizio**, sviluppato con esperienze pregresse e non prevedibili con solo l'uso di una sua semplice recensione. Tale ostacolo sarà considerato durante tutto lo studio e si cercherà di raggiungere un livello di predizione accettabile al fine di poter predire, in maniera oggettiva, il tipo di qualsiasi recensione.

Data Loading

2.1 Source del Dataset

Il dataset scelto è interamente scaricabile in **formato JSON** all'interno del link di download su Yelp Open Dataset.



Da come si può notare dall'immagine, vi è un totale di **8,635,403 reviews** rappresentante il dataset di input su cui lavorare. A questo vengono aggiunti **160,585 aziende** di vario tipo su cui si basano le reviews, **200,000 immagini** che illustrano i servizi recensiti su Yelp e **8 aree metropolitane** in cui sono situati. Il dataset ha una struttura relazionale che non sarà oggetto di studio, ma è importante sapere che, come nel nostro caso, utilizzando solamente i dati di un sottoinsieme di tabelle, si ha la presenza di colonne non utili al fine del progetto che, durante fasi apposite, verranno rimosse per allegerire il carico computazionale. Tutte le valutazioni in merito sono riportate nel capitolo dedicato al data pre-processing.

2.2 JSON Data Loading

Durante questa fase, sono stati caricati i file JSON per i business e per le reviews all'interno di dataframe appositi utilizzando pandas.

2.2.1 Load Business JSON

Il caricamento del file JSON dedicato ai business del dataset è stato relativamente **veloce** date le sue **scarse dimensioni**. In seguito al caricamento, si è presa la decisione di mantenere solamente quelle righe corrispondenti a **business ancora aperti**. Per poter selezionare tali righe, si è fatto uso della colonna *ís_open´* di tipo booleano. Inoltre, sono state **selezionate** solamente le colonne di interesse che siano legate direttamente con le **reviews**. Di seguito riportiamo la *head* del dataframe risultante:

	business_id	meanStars	reviewCount
0	6iYb2HFDywm3zjuRg0shjw	4.0	86
1	tCbdrRPZA0oiIYSmHG3J0w	4.0	126
2	bvN78flM8NLprQ1a1y5dRg	4.5	13
3	oaepsyvc0J17qwi8cfrOWg	3.0	8
4	PE9uqAjdw0E4-8mjGl3wVA	4.0	14

2.2.2 Load Reviews JSON

Uno dei principali **problemi affrontati** con il file JSON delle reviews è collegato alle **dimensioni del file stesso**. Per risolvere tale ostacolo, si è deciso di caricare il file secondo una **segmentazione di 100,000 chunks**, ossia leggendo 100,000 righe a step. Durante il caricamento delle righe all'interno del dataframe delle reviews, abbiamo **rimosso** le colonne con**parametri emotivi** (funny, cool e useful) poichè si intende lavorare sulle **correlazioni** tra il **testo** e i parametri dati alle **stelle** piuttosto che su altri valori numerici e, in aggiunta, abbiamo rimosso la colonna di *user_id e review_id* poichè non si ha nessuna correlazione con il dataframe di business caricato precedentemente.
Successivamente, durante il caricamento abbiamo effettuato un **merging** con il dataframe di business andando a mantenere solamente le reviews che fanno parte di un business ancora aperto, ossia di una attività che sia all'interno del dataframe di business risultante dal data cleaning precedente. Tutto il procedimento di review data loading ha impiegato in media **1 minuto e 56 secondi**. Infine, in seguito alla ridenominazione di alcune colonne, si è ottenuto il seguente **dataframe**:

	businessId	meanStars	reviewCount	reviewStars	text	date
0	6iYb2HFDywm3zjuRg0shjw	4.0	86	5.0	Stopped in on a busy Friday night. Despite the...	2018-03-04 00:59:21
1	tCbdrRPZA0oiIYSmHG3J0w	4.0	126	4.0	Elephant's contacted me the same day I posted ...	2012-07-16 05:04:05
2	tCbdrRPZA0oiIYSmHG3J0w	4.0	126	5.0	I'm not usually a fan of airport food. I usual...	2015-04-28 21:11:10
3	tCbdrRPZA0oiIYSmHG3J0w	4.0	126	4.0	If one must have breakfast at the airport, per...	2015-11-18 18:50:05
4	tCbdrRPZA0oiIYSmHG3J0w	4.0	126	5.0	Reasonably priced, tasty local joint. Lots of ...	2011-11-30 20:15:41

2.2.3 Data Converting

Per le fasi successive, abbiamo caricato il dataframe risultante in un **file .csv** in modo da poter lavorare nelle fasi successive direttamente sui dati già pre-elaborati dalla fase di data loading.