Estrazione automatica di conoscenza da recensioni Trip Advisor di ristoranti italiani con R Elaborato Text Mining

Diego Pergolini

24 Gennaio 2019

1 Introduzione

Dato il sempre maggior utilizzo di piattaforme quali Trip Advisor per recensire hotel, ristoranti o attrazioni varie, i dati acquisiti da questi siti possono essere di vitale importanza per le stesse attività, potendo infatti ricavare preziose indicazioni sul grado di soddisfazione della clientela. Al contempo, dato il variegato panorama di opzioni disponibili, un utente può far difficoltà a scegliere il miglior luogo a seconda delle sue esigenze. Emerge quindi l'esigenza di un ausilio automatizzato per questo tipo di ricerche, senza dover leggere una a una le miriadi di recensioni presenti. Il progetto qui presentato si propone di esplorare alcune delle possibili opzioni atte a soddisfare questo bisogno.
In questo caso si è deciso di prendere in considerazione le recensioni presente su Trip Advisor, sia per comodità di utilizzo sia per confrontare più facilmente i risultati, ma le tecniche applicate potrebbero essere applicate a qualsiasi fonte di testo libero che contenga opinioni di utenti. Come buona norma, durante il progetto verrà seguita la metodologia CRISP-DM, che indica gli step fondamentali per condurre un buon progetto di Data/Text Mining.

2 Requisiti

Obbiettivo dell'elaborato è realizzare degli script per estrarre conoscenza dalle recensioni, in italiano, lasciate dagli utenti relativi ai ristoranti di una certa zona d'interesse. Nello specifico:

1. Le recensioni di tutti i locali della zona di interesse saranno ottenute automaticamente con degli script R.

2. Per ogni ristorante verrà eseguita la **sentiment analysis a livello di frasi e feature** (Cucina, Ambiente, Qualità/Prezzo, Servizio). Per farlo

verrà utilizzato l'**opinion lexicon** italiano ottenuto da kaggle [1]. Si potrà quindi fare un confronto tra i risultati ottenuti e quelli reali postati su trip advisor.

3. Sarà possibile, data una certa parola, ad esempio 'piadina', ottenere una classifica dei ristoranti ordinati in base al giudizio degli utenti su quella specifica feature.

4. Considerato uno specifico locale, si potranno visualizzare le k recensioni più vicine semanticamente ad un termine. Questa funzionalità verrà implementata utilizzando **LSA**. L'idea è che,ad esempio, un ipotetico utente una volta trovato il ristorante con la piadina migliore, possa visualizzare un certo numero di recensioni che la descrivano nello specifico.

5. Verrano creati dei grafici di tipo **'WordCloud'** che mostreranno i termini più frequenti nelle recensioni di un ristorante specifico e di tutti i ristoranti della zona.

3 Processo

Come già introdotto è stata adottata la metodologia CRISP-DM, le quali fasi verranno esplicitate in questo capitolo.

3.1 Capire il domino applicativo

Obbiettivo del progetto è di fornire all'utente uno strumento attraverso il quale, selezionata un'area di interesse, ordinare i ristoranti secondo:

- **Le feature standard di Trip Advisor** (Cucina, Servizio, Atmosfera, Qualità/Prezzo). Le valutazioni relative ad esse, vengono già fornite dagli utenti in fase di redazione della recensione, ma potrebbe essere comunque utile desumerle puramente dal testo, infatti i valori dati dagli utenti potrebbero non essere coerenti con quanto espresso nella recensione testuale.

- **Attraverso una specifica feature da esso introdotta**. Questa modalità potrebbe essere particolarmente utile per ordinare i ristoranti in base all'opinione espressa dagli utenti su di una specifica pietanza.

Nel secondo caso l'utente potrebbe poi volere leggere di persona le recensioni del ristorante migliore per la feature scelta (o di un altro ristorante a sua scelta), l'idea è quella di visualizzare solo le recensioni più semanticamente vicini al termine inserito, in modo da fornire il massimo contenuto informativo.
L'utente potrebbe poi desiderare qualche riferimento grafico ed immediato delle

[1] `https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages#sentiment-lexicons.zip`

parole che meglio descrivono un certo locale o una certa zona di interesse, andranno quindi mostrati dei grafici appropriati a questo scopo.

Fatte queste considerazioni si è definito il seguente piano di progetto:

- **Creazione di uno script R per fare "Scraping" su Trip Advisor**, dato che il sito non mette direttamente a disposizione tutte le recensioni in formato fruibile. Specificata una zona d'interesse lo script dovrà scaricare in file csv le recensioni di ogni ristorante, complete di titolo, testo completo e valutazione.
- **Scelta della tecnica per compiere Sentiment Analysis a livello di frasi e feature**. Per semplicità d'uso e buon grado di efficacia si è deciso fin da subito di adottare la funzione di scoring di Hu e Liu per stimare il sentiment di una frase riguardo ad una feature. Dovrà essere predisposta anche una tecnica per rilevare le frasi che parlano di una certa feature.
- **Individuare file a supporto della Sentiment Analysis a livello di feature**. Per compiere l'analisi sarà infatti necessario un Opinion Lexicon per la lingua italiana ed una lista di stopwords.
- **Individuare le feature words adatte**. Considerata la tecnica adottata, è necessario definire un insieme di feature word per ogni feature, al fine di individuare tutte le frasi che la riguardano. Sarebbe preferibile utilizzare un approccio automatico, soprattutto in riferimento al requisito 3
- **Creazione di uno script per aggregare le recensioni di più ristoranti**, in modo da poterle utilizzare tutte insieme per creare Word Embeddings.
- **Creazione di uno script per portare le recensioni in uno spazio LSA**. Per soddisfare il requisito 4 si dovranno preparare adeguatamente le recensioni relative al locale scelto in modo da portarle in uno spazio LSA, da utilizzare poi per recuperare i k documenti più vicini al termine inserito.
- **Creazione di grafici Word Cloud**. In merito al requisito 5 verranno creati grafici per mostrare al meglio i termini caratterizzanti il locale o l'area d'interesse.
- **Valutare i risultati ottenuti**. Saranno prese in considerazione quattro aree di interesse al fine di valutare i risultati ottenuti: Cesena, Jesi, Firenze, Napoli.

3.2 Capire i dati

Grazie allo script R realizzato è stato possibile ottenere quasi 340.000 recensioni di svariati locali così distribuite:

- 23405 Recensioni distribuite in 263 locali per Cesena.
- 171420 Recensioni distribuite in 232 locali per Firenze.
- 10692 Recensioni distribuite in 125 locali per Jesi.
- 134429 Recensioni distribuite in 178 locali per Napoli.

Le recensioni in generale contengono una valutazione in scala 1-5 (non usata nell'analisi), il titolo (non usato) ed il testo completo, il quale è di solito composto da un minimo di una frase ad un massimo di una decina. Si è notato che le recensioni positive sono circa 8 volte quelle negative e quasi 6 volte le neutrali.

3.3 Preparazione dei dati

Di seguito verranno illustrati i passi compiuti per preparare i dati al fine del soddisfacimento dei requisiti individuati.

3.3.1 Sentiment analysis a livello di frasi e feature standard

Per prima cosa, per applicare la funzione di scoring di Hu e Liu, si è ottenuto un Opinion Lexicon per la lingua italiana, contenente 1648 parole di orientamento positivo e 2694 di orientamento negativo.
Altro ingrediente fondamentale per utilizzare la funzione di scoring già citata è necessario disporre di una lista di feature words collegata ad ogni feature, al fine di aumentare il numero di frasi che fanno match con la caratteristica scelta. Per creare queste liste di feature words si è deciso di adottare un approccio automatico, basandosi sull'ottimo supporto fornito dai **Word Embeddings** nell'incarnazione Glove. L'idea è stata quella di addestrare un modello di Word Embeddings a partire da tutte le recensioni raccolte, così da ottenere delle rappresentazione dei termini che fossero organizzate nello spazio in modo semantico. Così facendo, infatti, la maggior parte dei termini relativi alle varie feature saranno vicini tra loro, potendo quindi estrarre in automatico quali sono le feature words associate. I parametri utilizzati per il training sono i seguenti:

- **Skip-gram windows: 10** si è scelto di adottare una finestra abbastanza larga in modo che i termini venissero associati al macro-aspetto a cui si riferiscono, finestre più piccole non facevano emergere relazioni soddisfacenti.
- **Dimension: 50** dato che il testo fornito pur essendo abbastanza vasto, non è particolarmente eterogeneo (si parla comunque di ristoranti), si è ritenuto preferibile non usare troppe dimensioni per rappresentare i termini, pena la perdita di alcune associazioni fondamentali.

Prima del training del modello il corpus è stato adeguatamente tokenizzato, ripulito dalle stopwords e dalle parole apparse meno di 5 volte. La bontà del modello risultante può essere notata anche da alcuni esempi di operazioni fra

vettori di termini:

$$cameriere - uomo + donna \approx cameriera(0.8521)$$
$$pizze - pizza + piadina \approx piadine(0.8874)$$
$$pizzeria - pizza + piadina \approx chiosco, piadineria(0.75)$$

Nello specifico, tramite questo metodo sono state estratte 141 parole per la feature **Atmosfera**, 190 per **Servizio**, 160 per **Qualità/Prezzo** e 472 per **Cucina**. A queste parole ne sono state aggiunte una minima parte manualmente.

Il modello creato viene utilizzato anche per soddisfare il requisito 3, infatti, al termine inserito dall'utente vengono associate come feature words quelle parole che hanno similarità maggiore a 0.7 nello spazio di rappresentazione, così facendo verranno considerati (sperabilmente) anche i sinonimi, senza l'ausilio di WordNet.

3.3.2 Individuazione recensioni simili a termine

Per applicare LSA alle recensioni di un certo locale dovremo preliminarmente creare una matrice termini-documenti. Questa matrice è stata creata con un apposito script, il quale converte tutte le parole in minuscolo, rimuove i segni di punteggiatura e le stopword specificate, per poi creare una TermDocument-Matrix che andrà poi ripulita dei termini che compaiono in meno dell'1% delle recensioni. A questo punto si dispone del necessario per applicare **LSA**.

4 Creazione del modello

In questa sezione verranno esplicate le tecniche adottate per risolvere le esigenze emerse dai requisiti.

4.1 Sentiment analysis a livello di frasi e feature

Come già introdotto, il desiderio è quello di calcolare il gradimento degli utenti riguardo varie caratteristiche dei locali esaminati. Per farlo si è utilizzato il seguente approccio:

- Ogni recensione viene divisa in frasi in corrispondenza di un punto.
- Ogni frase viene divisa in parole, viene valutato se contiene una delle feature words e per ogni match viene contato quante parole positive e negative sono presenti nella frase. Si otterrà quindi uno score della frase, che verrà quindi aggregato con le altre frasi della recensione per formare lo score finale.
- Tutti gli score delle recensioni relative ad un locale verranno aggregati e scalati in un punteggio da 0 a 50.

Il procedimento appena illustrato viene quindi ripetuto per ogni feature standard (Cucina, Servizio, Atmosfera, Qualità/Prezzo) o soltanto per la feature desiderata. Gli score ottenuti da tutti i ristoranti vengono poi messi in una tabella,ordinati e graficati.

4.2 Recensioni in uno spazio LSA

Al fine di visualizzare le k recensioni più semanticamente simili alla query inserita sono stati adottati questi passi:

- A partire dal TermDocumenMatrix si produce esegue una scomposizione SVD con un numero di dimensioni determinato automaticamente. A questo punto si dispone della matrice con i vettori dei termine, la matrice dei valori singolari, e la matrice delle recensioni.
- Viene individuato il numero di dimensioni da mantenere, ciò viene fatto scegliendo il punto di knee nella curvatura della sequenza dei valori singolari. Il punto con maggiore differenza di ordinata tra un numero di dimensioni ed il successivo verrà scelto.
- Vengono create le matrici di similarità semantica per termini e per documenti.
- La query specificata viene portata nello spazio LSA, moltiplicandola trasposta del suo vettore di termini pesato per la matrice termini,variabili latenti.
- Viene infine calcolata la similarità coseno fra query nello spazio LSA e documenti presenti, restituendo i prima k con similaritù più alta.

4.3 Creazione grafici Word Cloud

Per creare i grafici di tipo Word Cloud è necessario creare una matrice termini documenti, per poi visualizzare i termini in base alla loro frequenza nei documenti. Prima di creare la TDM vanno sempre eseguiti i passi fondamentali di pre-processing, quali conversione in minuscolo, rimozione di punteggiatura e stopwords. La matrice Termini-Documenti viene pesata attraverso Tf-IDf per dare il giusto peso ai termini.

4.4 Implementazione

In questa sotto-sezione verrà descritta brevemente l'implementazione del progetto, distinguendo per requisiti i vari file coinvolti. Tutto il materiale è disponibile all'indirizzo `https://github.com/DiegoPergolini/TextMiningProject`

4.4.1 Ottenere automaticamente recensioni di Trip Advisor

Nel file **TripAdvisorScraper.R** è contenuta tutta la logica per scaricare tutte le recensioni di tutti i ristoranti della zona specificata. E' sufficiente specificare

l'url iniziale, ad esempio `https://www.tripadvisor.it/Restaurants-g187785-Naples_Province_of_Naples_Campania.html` e verranno automaticamente scaricate le recensioni nella cartella di lavoro. Il tutto è stato realizzato tramite l'ottima libreria per il web scraping. **rvest**

4.4.2 Scoring delle opinioni per feature standard

Per creare la lista delle feature words per le feature standard di trip advisor si è utilizzato il file **FeaturesWordWithGlove.R**, in cui viene addestrato il modello di Word Embeddings e formate le liste di parole precedentemente descritte.

Tutta la logica per l'attribuzione degli score ad ogni ristorante della zona desiderata è nel file **StandardFeaturesRanking.R**. Questo file contiene quanto descritto precedentemente per calcolare gli score dei locali e per graficare i risultati aggregati.
Per calcolare gli score di un singolo ristorante si può utilizzare il file **StandardFeaturesScoreByRestaurant.R**.

4.4.3 Scoring delle opinioni per feature scelte dall'utente

Per ordinare i locali in base alle opinioni su di una singola feature scelta dall'utente si fa uso del file **RankingBySelectedFeature.R**, in cui, specificato il path dove risiedono le recensioni della zona d'interesse e la feature, crea automaticamente una classifica e ne grafica il risultato.

4.4.4 Visualizzare recensioni più simili a query

La tecnica descritta nella sezione 4.2 viene implementata nel file **GetSimilarReviewByTerm.R**, in cui, specificato il file contenente le recensioni del locale scelto e la query di interesse, viene prima creata la TermDocumentMatrix con il file **CreateTdmTable.R** e poi vengono mostrate le k recensioni più rappresentative per la query specificata.

4.4.5 Creazione grafici WordCloud

Per generare i grafici di tipo WordCloud relativi ad un ristorante viene utilizzato il file **WordCloudBuilder.R** mentre per quelli relativi ad una intera zona il **WordCloudForArea.R**.

5 Valutazione dei Risultati

Si procede di seguito ad illustrare alcuni dei risultati ottenuti durante il progetto, corredando il tutto con grafici e figure prodotte dai vari script.

5.1 Scoring di ristoranti per Features Standard

Per mostrare la bontà dell'approccio scelto si considereranno un paio di ristoranti per città, ovviamente questo non può assolutamente rappresentare un test esaustivo, ma vuole soltanto mostrare una indicazione generale sull'approccio adottato. Per la città di Cesena si sono considerati i locali: Micamat Piadineria, Mastrobirraio e Qbio.

Ristorante	Fonte	Cucina	Servizio	Qualità/Prezzo	Atmosfera	Media
Micamat	Calcolato	42	40	41	43	41.5
Micamat	Reale	45	40	45	35	45
Mastrobirraio	Calcolato	40	39	41	41	40.25
Mastrobirraio	Reale	40	40	35	40	40
Qbio	Calcolato	40	38	40	42	40
Qbio	Reale	40	40	30	40	40

Per la città di Firenze si sono considerati i locali: Panini Toscani e Haveli Indian Restaurant.

Ristorante	Fonte	Cucina	Servizio	Qualità/Prezzo	Atmosfera	Media
Panini Toscani	Calcolato	42	40	41	41	41
Panini Toscani	Reale	50	45	45	50	50
Haveli Indian	Calcolato	41	40	41	43	41.25
Haveli Indian	Reale	45	44	40	45	45

Per la città di Jesi si sono considerati i locali: Mare-Mare cucina di pesce e Pepito.

Ristorante	Fonte	Cucina	Servizio	Qualità/Prezzo	Atmosfera	Media
Mare-Mare	Calcolato	42	41	43	42	42
Mare-Mare	Reale	45	45	40	45	45
Pepito	Calcolato	34	32	35	36	34.25
Pepito	Reale	35	30	35	30	30

Per la città di Napoli si sono considerati i locali: Hachi Ristorante Giapponese e A'Cucina Ra Casa Mia .

Ristorante	Fonte	Cucina	Servizio	Qualità/Prezzo	Atmosfera	Media
Hachi	Calcolato	42	40	41	43	41.5
Hachi	Reale	40	40	40	45	45
A'Cucina Ra Casa Mia	Calcolato	40	40	42	42	41
A'Cucina Ra Casa Mia	Reale	45	45	45	40	45

Come si evince dai risultati, gli score calcolati sono abbastanza precisi, considerata anche la tecnica non raffinatissima utilizzata ed i valori di trip advisor che vanno di 5 in 5.

5.2 Ricerca di recensioni rilevanti per termine

Per mostrare la bontà dell'utilizzo di LSA per il compito descritto dal requisito 4, si mostreranno alcune delle recensioni più rilevanti ritornate dallo script, prendendo in considerazione il termine *Lampredotto* ed il ristorante *Trippaio del Porcellino*:

1. "Poco da dire, bisogna solo mangiare e godersi l'ambientazione e i rumori da street food. Ottimo lampredotto e il vinello da accompagnarci"

2. "Riscoprire un sapore estinto in alcuni posti penso che sia la cosa più buona che mi sia capitata riuscire a far mangiare la trippa ai miei figli e stato qualcosa di fantastico lo consiglio a tutti i Buon gustai"

3. "Indubbiamente il panino al lampredotto piú buono di Firenze, servito con il vero pane fiorentino e non con la rosetta. Orazio, il proprietario, é cortese e simpatico, il che non guasta mai.. "

4. "Il lampredotto a Firenze è un mito...ma se non ci si ferma al camioncino del Trippaio del Porcellino. ..non si capirà mai perché qui è un rituale...alla pizzaiola...con un po' di salsa verde al prezzemolo. ..diventa uno splendido intervallo..tra un visita d'arte ed un giro nei bellissimi negozi di Firenze.!!.e tornerete con un ricordo in più. ..ma che vi avrà saziato e soddisfatto...e che non si ripete da nessun altra parte...assolutamente da provare!!!!"

5. "Un sogno realizzato!!! Volevo proprio assaggiare un bel panino con il lampredotto e qui ho avverato il mio desiderio e non sono stata x niente delusa!!! Vera cucina fiorentina, in una delle piazze più vissute di Firenze, cibo ottimo a prezzi bassi, simpatia e cortesia dei proprietari, grazie di tutto!!! Provato anche il lampredotto all'anzimino ossia con sugo bietole e spinaci, che goduriaaaa"

6. "In vacanza a Firenze è quasi una tappa obbligatoria! Avevo sentito parlare con il miglior lampredotto lo servisse questo chioschetto in un noto programma tv. Consideriamo che il lampredotto cosi come la trippa non è un piatto di mio gradimento, ma andava cmq assaggiato! Mia moglie invece ha molto gradito ed apprezzato in pieno!"

7. "Ho mangiato qui il mio primo lampredotto ma non sarà di sicuro l ultimo, il pane era fresco e la salsina verde eccezionale! Il baracchino era pulito (cosa non scontata per nulla), il prezzo onesto, abbiamo pagato panino + coca cola 6 euro, si trova all interno di una piazzetta in mezzo al mercato, da provare!"

5.3 Classifiche dei primi 15 ristoranti per le feature standard

A titolo esemplificativo viene mostrato il grafico risultante dall'analisi dei ristoranti di Cesena per le Feature Standard.

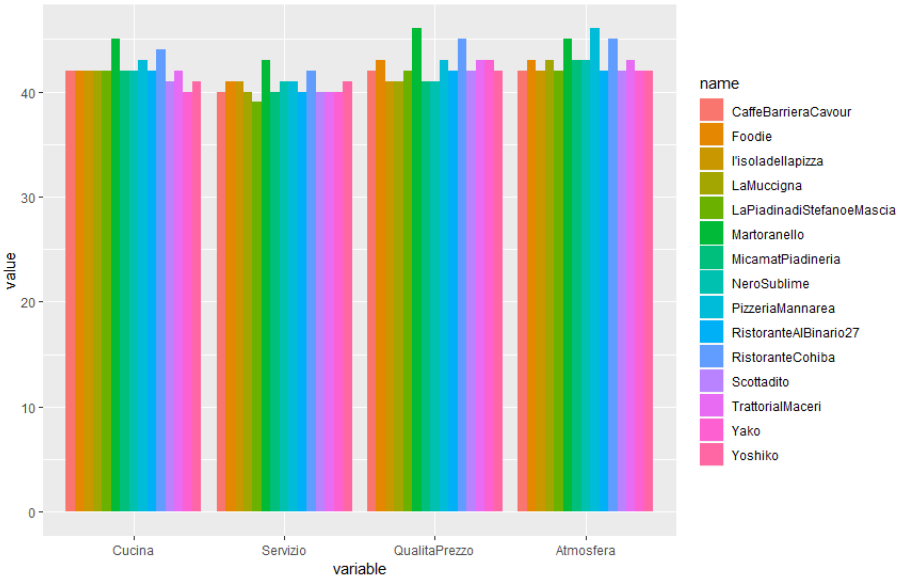

Figure 1

5.4 Visualizzazione WordCloud Località

(a) *Jesi*

(b) *Cesena*

(c) *Napoli*

(d) *Firenze*

Figure 2: Esempio di grafici WordCloud per le località

5.5 Visualizzazione WordCloud Ristoranti

(a) *Mastrobirraio, Cesena*

(b) *Micamat Piadineria, Cesena*

(c) *1947 Pizza Fritta, Napoli*

(d) *Panini Toscani, Firenze*

Figure 3: Esempio di grafici WordCloud