Università degli Studi dell'Insubria

Department of Theoretical and Applied Science DiSTA

Master's Degree in Computer Science

Data Mining Project on hotel reviews

Author:

Dario Bertolino

724118

Data Mining lectures by Prof. Ignazio Gallo

Academic year 2018 - 2019

Indice

Introduzione

L'idea alla base del progetto è quella di effettuare un'analisi per capire quali siano i migliori hotel europei. A tal proposito è stato scelto un dataset che raccoglie 515K recensioni rilasciate per 1493 hotel di lusso europei attraverso booking.com. I testi delle recensioni (in lingua inglese) sono accompagnati da numerosi metadati come il numero di giorni passati dalla data del checkout, la posizione geografica dell'hotel, la nazionalità dei reviewers e la data di rilascio della recensione.

Link del Dataset:

https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

Obiettivi in termini di Pulizia e Preparazione del dato:

- Estrarre dal testo delle recensioni uno score di "positività" attraverso sentiment analysis.
- Estrarre da latitudine e longitudine la città e il paese di appartenenza dell'hotel.
- Gestione dei valori nulli.
- Visualizzazione features al fine di poter meglio osservarne la distribuzione.

Principali obiettivi di analisi:

- Indagare su una possibile correlazione tra nazionalità dei reviewer e grado di positività delle recensioni.
- Indagare su una possibile correlazione tra il numero di giorni dal checkout e grado di positività delle recensioni.
- Individuare sottogruppi frequenti di parole per recensioni positive, negative.
- Addestrare un modello per la classificazione degli hotel basato sulle reviews.
- Addestrare un modello di regressione per mettere in relazione posizione geografica e periodo dell'anno con la positività delle recensioni.

Il progetto è stato interamente sviluppato con Jupyter notebook e Python, i dettagli relativi al progetto sono riportati direttamente nei notebook.

La seguente relazione ha lo scopo di evidenziare il ragionamento generale relativo all'implementazione.

Notebooks overview

Il progetto è compost da 4 notebook.

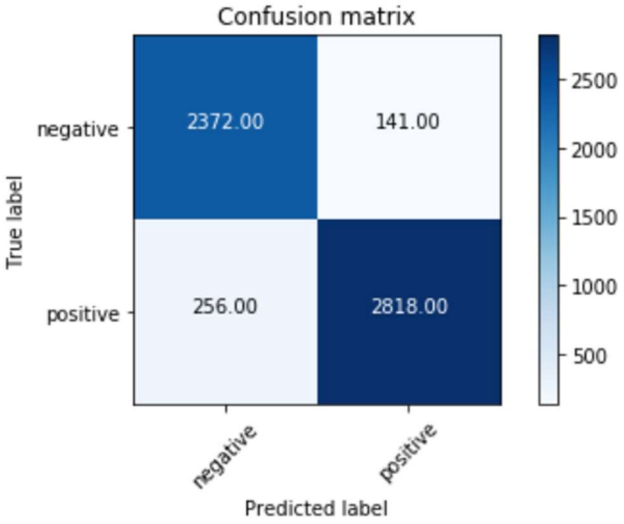
0_clean_explore_data

Dedicato ad una esplorazione preliminare dei dati, contiene gestione dei valori nulli, l'estrazione di paesi di appartenenza degli hotel da latitudine e longitudine, pulizia del dato e visualizzazione di alcune informazioni base come il numero di hotel e paesi, la distribuzione dei mesi in cui sono state rilasciate le reviews e la lunghezza massima/media delle reviews.

1_sentiment_analysis

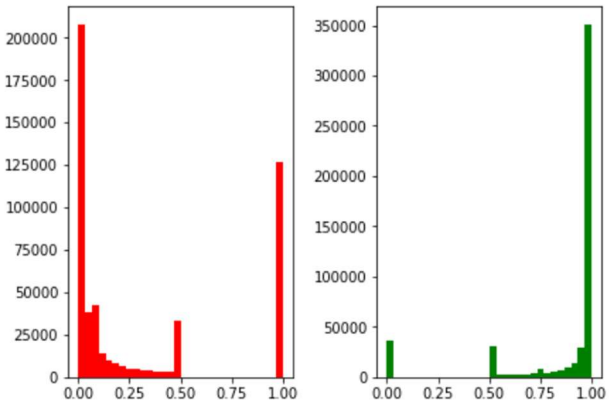
Dedicato alla definizione e l'addestramento di una rete neurale per la sentiment analysis, in particolare sono stati estratti gli embeddings delle parole usate nelle reviews con skip-gram model di Word2Vec. Le rappresentazioni vettoriali sono state usate per inizializzare i pesi del layer di embedding nella rete neurale.

Dopo un training di sole 5 epoche su un campione di 10000 reviews, la rete ha raggiunto i seguenti risultati nella classificazione di review positive o negative:



	precision	recall	f1-score	support
0	0.90	0.94	0.92	2513
1	0.95	0.92	0.93	3074
micro avg	0.93	0.93	0.93	5587
macro avg	0.93	0.93	0.93	5587
weighted avg	0.93	0.93	0.93	5587

Inserendo la funzione sigmoide come funzione di attivazione del neurone di output, la rete è stata successivamente usata per estratte uno score di positività tra 0 e 1, per ogni review.

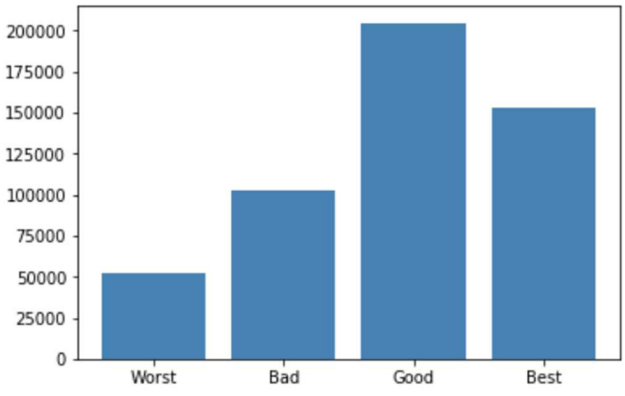


Essendo ogni review completa composta da una negativa e una positiva, la media tra i due score estratti dalla rete ha definito un nuovo score finale del sentimento. Questo score è stato usato per sostituire il punteggio da 0 a 10 presente nel dataset e per dividere tutte le recensioni in 4 diverse classi (Best, Good, Bad e Worst). Riporto alcuni esempi che evidenziano le differenze tra lo score di analisi del sentimento e il punteggio lasciato dal reviewer.

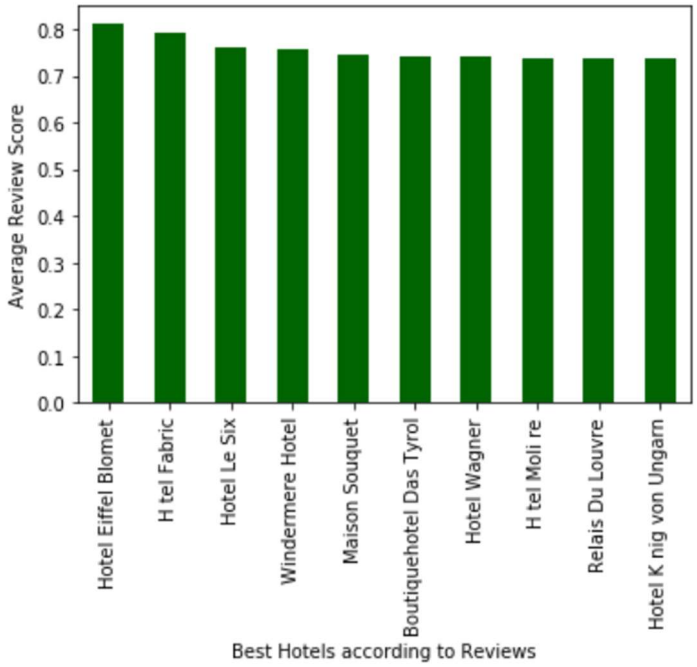
Positive review: the staff was so helpful the location is great you are just near champs leysee avenue
Negative review: the price for breakfast could be cheaper
Sentiment score: 0.5050491094589233
Reviewer score: 9.6

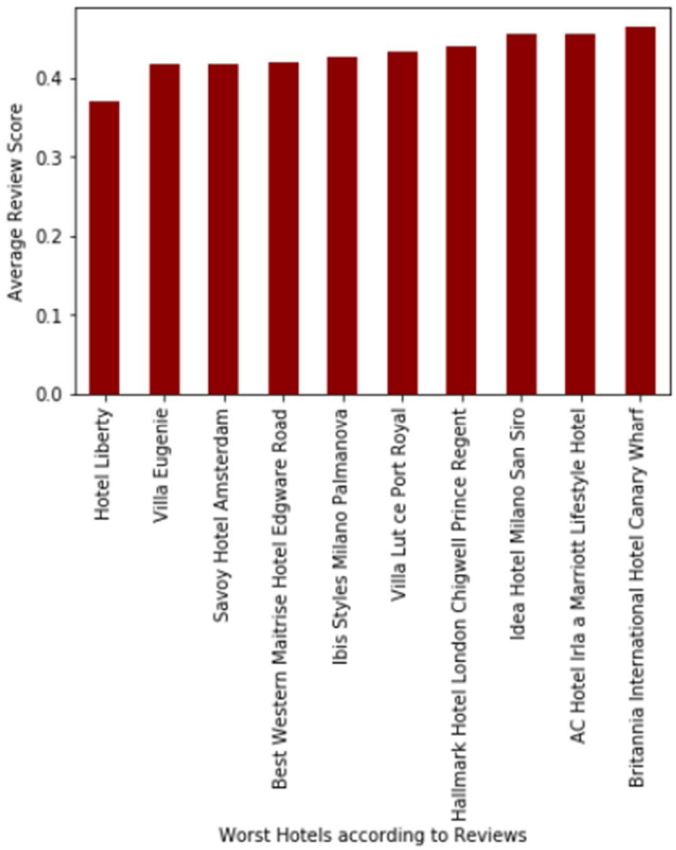
Positive review: no positive
Negative review: a bit far from city centre
Sentiment score: 0.0030985779594630003
Reviewer score: 9.2

Positive review: nice place to stay next to a great food court staff were very helpful and friendly the room was comfortable and had everything we needed nice bathroom also it was a little way to walk to the museumplein and center of town but worth the quiet of the west side really nice stay
Negative review: no negative
Sentiment score: 0.9999998807907104
Reviewer score: 9.6



Per concludere le classi best e worst sono state usate per scovare i migliori e i peggiori hotel secondo l'analisi del sentimento:





2_words_analysis

Dedicate all'estrazione della frequenza delle parole usate nelle varie classi di reviews scovate dall'analisi del sentimento.

Per prima cosa sono state eliminate le stopwords dal testo delle reviews.

Successivamente è stato implementato l'algoritmo aPriori, con il quale sono stati individuati i subset di parole (con supporto minimo pari 0.1) per ogni classe. Riporto le osservazioni riguardanti i risultati ottenuti:

Observations

Observing the most subsequent subsets of words with a minimum support of 0.1 it seems clear that:

- Most frequent **best reviews** contains great opinionis about the *staff*, that could be helpful, friendly or in general appreciated with the location or the hotel itself.
- Most frequent **good reviews** are very similar to best ones and again the staff play a key role. However I find really interesting that while [*location great*] is a most frequent subset in best reviews, [*location good*] is a most frequent one in good reviews. This could mean that the sentiment analysis done by the neural network is right.
- Most frequent **bad reviews** contains gripes about breakfast and small rooms.
- Most frequent **worst reviews** are equals to the bad reviews but the word hotel compare too, this could be because the worst reviews gripes about the entire hotel and not only a particular service.

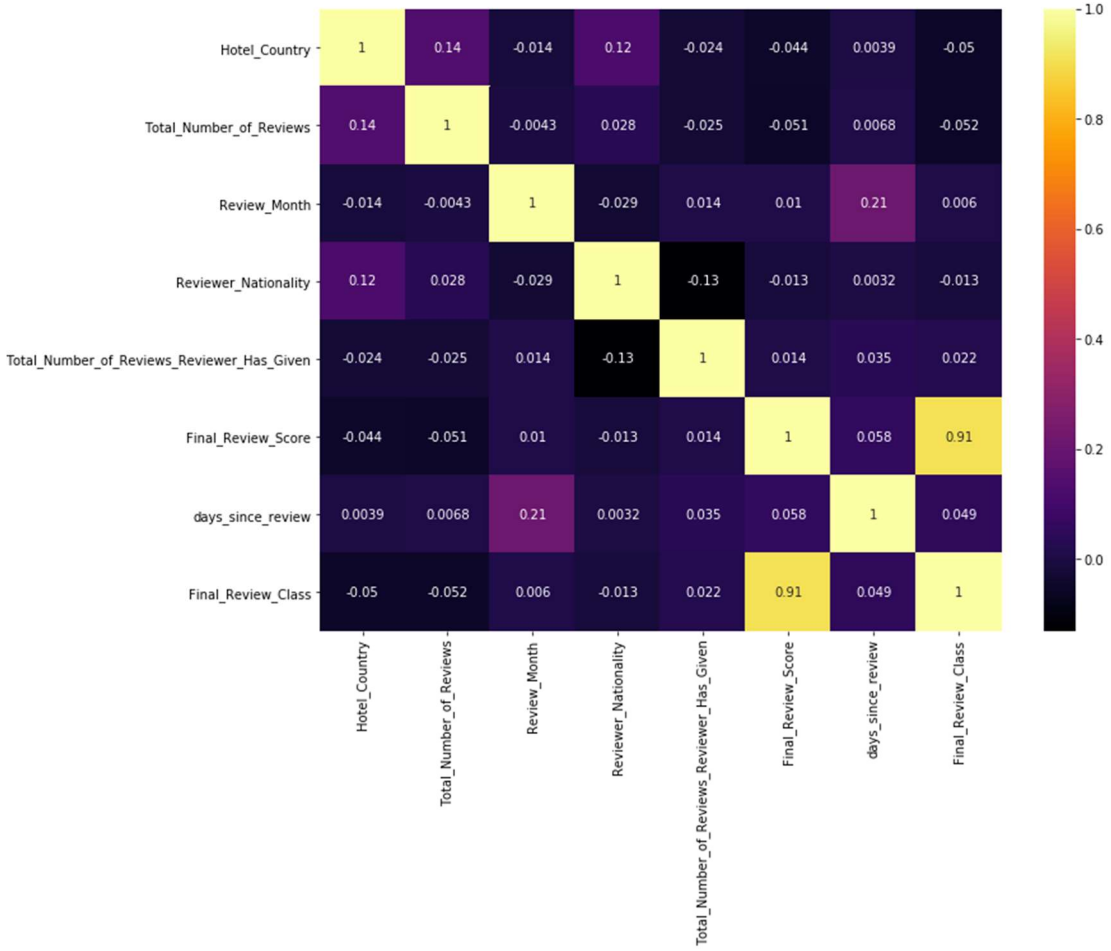
Infine, tramite l'utilizzo di un CountVectorizer, è stato creato un piccolo script che permette di estrarre la top ten di parole più frequenti (con relativo numero di apparizioni) per classe e paese di appartenenza degli hotel:

Observations

1. The word **staff** is the most frequent for each country in best reviews.
2. The word **location** is the most frequent for Italy, España and France in good reviews.
3. The word **room** is the most frequent for each country in bad reviews.
4. The word **room** is again the most frequent for each country in worst reviews, however also **would** and **could** are often present in the top ten. This means that a negative review often contains suggestions to the hotel.

3_classification_regression

Riporto una matrice di correlazione relativa alle le features dichiarate negli obiettivi di analisi:



Considerando che **Final_Review_Score** e **Final_Review_Class** sono rispettivamente lo score del analisi del sentimento e la classe della review, è facile notare che non esiste alcuna correlazione tra le singole features. La correlazione è una misura di associazione lineare, è quindi possibile che esistano anche forti associazioni non-lineari tra le features. Per questo motivo, oltre ad un semplice **Decisional Tree Classifier**, il quale non è riuscito ad effettuare la classificazione, è stata testata anche una **Rete Neurale** con **fully connected layers**. L'addestramento ha evidenziato la mancanza di ogni possibile forma di associazione:

```
Epoch 1/5
461223/461223 [==============================] - 19s 42us/step - loss: 0.6289 - acc: 0.0026
Epoch 2/5
461223/461223 [==============================] - 19s 40us/step - loss: 0.6230 - acc: 0.0026
Epoch 3/5
461223/461223 [==============================] - 18s 40us/step - loss: 0.6228 - acc: 0.0026
Epoch 4/5
461223/461223 [==============================] - 18s 40us/step - loss: 0.6227 - acc: 0.0026
Epoch 5/5
461223/461223 [==============================] - 19s 41us/step - loss: 0.6227 - acc: 0.0026
51247/51247 [==============================] - 1s 17us/step
Test accuracy: 0.002575760533341543
```