

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

DIPARTIMENTO DI INFORMATICA SISTEMISTICA E COMUNICAZIONE

Progetto Modelli Probabilistici per le Decisioni

- RELAZIONE FINALE -

Autori:

Cimmino Fabio 807070

Lotterio Roberto 807500

Ventura Samuele 793060

Anno accademico 2018/2019

Indice

1	Obiettivi	5
2	Preprocessing	7
3	Esplorazione del Dataset	9
3.1	Selezione termini per la creazione della rete bayesiana	11
4	Sviluppo Modello	13
5	Interfaccia Grafica	15
5.1	Review Classification	15
5.2	Hotel's sentiment words	16
6	Conclusioni	19

Capitolo 1

Obiettivi

Il progetto intitolato *TripAdvisor Review Classification* propone lo sviluppo di un modello basato su una rete bayesiana in grado di classificare le recensioni fornite dagli utenti relative ai soggiorni presso un hotel. I dati presi in analisi sono i commenti e i relativi metadati reperiti dal sito di TripAdvisor.

Al fine di dare una dimostrazione pratica del modello implementato è stata sviluppata un'applicazione. La demo consiste in due interfacce principali: la prima assegna una valutazione ad una recensione basata sul commento e sui metadati inseriti dall'utente; la seconda permette di visualizzare un istogramma ed un wordcloud delle parole più rilevanti di un hotel.

Capitolo 2

Preprocessing

Questo è un passo importante in qualsiasi processo di data mining, implica fondamentalmente la trasformazione dei dati grezzi in un formato comprensibile per i modelli NLP (Natural Language Processing). I dati del mondo reale sono spesso incompleti, incoerenti e possono contenere errori. La pre-elaborazione dei dati è un metodo utilizzato per risolvere tali problemi. Ciò contribuirà a far ottenere risultati migliori attraverso gli algoritmi di classificazione.

Per poter ottenere un dataset "pulito" è stato necessario eseguire i seguenti passaggi di preprocessing e cleaning del commento:

- La prima fase consiste nella **tokenizzazione** di ogni commento presente nel dataset, si tratta di un processo di scomposizione di un testo in parole, frasi, simboli o altri elementi significativi denominati token; si utilizza questa tecnica per dividere ogni commento in singole parole ed effettuare una pulizia più efficace.
- La seconda fase consiste nella rimozione delle **stopwords**, ovvero le parole che vengono ripetute spesso nel documento ma che non hanno alcuna rilevanza, come articoli e verbi; per questa fase abbiamo utilizzato un dataset di stopwords inglesi già esistente.
- Successivamente sono stati rimossi tutti i token che consistevano in elementi di punteggiatura, numeri, whitespace e caratteri non codificabili come simboli ASCII.
- La quarta fase prevede di trasformare tutti i token in caratteri minuscoli per uniformare ulteriormente il testo.
- La quinta ed ultima fase del preprocessing consiste nell'effettuare la **lemmatizzazione** delle parole, cioè la riduzione di una forma flessa di una parola alla sua forma canonica, detta lemma.

Capitolo 3

Esplorazione del Dataset

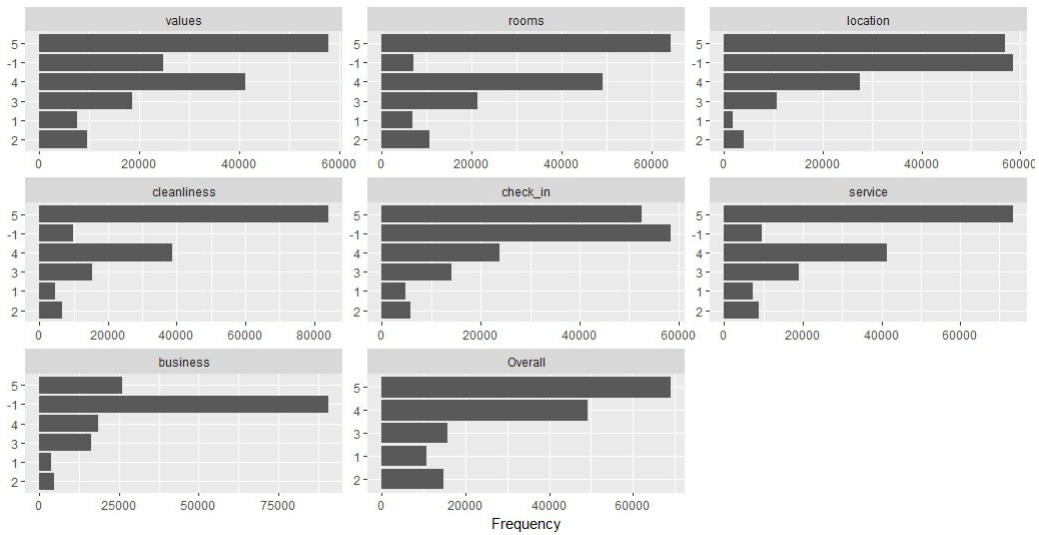
Per eseguire l'analisi esplorativa del dataset è stato utilizzato il package per R *DataExplorer*, il dataset iniziale risultava composto dalle seguenti feature:

- id hotel: il codice univoco con cui si rappresenta un hotel;
- id review: il codice univoco di ogni recensione per un solo hotel;
- content: il commento scritto dall'utente;
- no reader: il numero di persone che hanno letto la recensione;
- no helpful: il numero di persone che hanno trovato la recensione utile;
- value: rappresenta il rapporto qualità/prezzo dell'hotel;
- rooms: rappresenta la qualità delle camere, dalla grandezza alla pulizia;
- location: questo valore va assegnato tenendo conto della qualità della struttura alberghiera;
- service: rappresenta la qualità dei servizi dell'hotel, comprendono quindi la cucina, lo staff, i servizi in camera e i restanti servizi offerti dall'hotel;
- cleanliness: questa valutazione deve prendere in considerazione soltanto la pulizia dell'hotel in generale, tralasciando la pulizia delle stanze;
- business: rappresenta la valutazione del servizio di business;
- check-in: riguarda l'accoglienza e la prenotazione nella struttura;
- overall: rappresenta la valutazione complessiva della recensione.

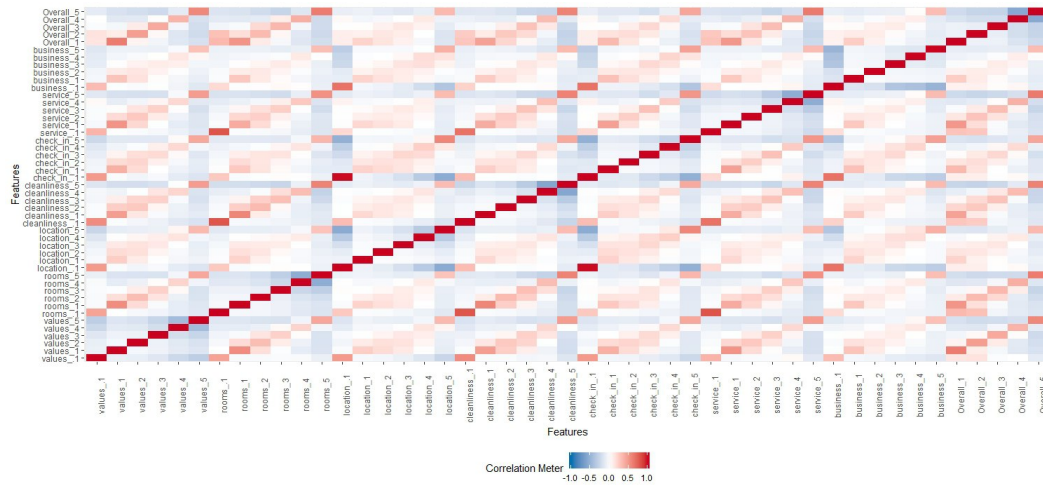
Analizzando il dataset abbiamo riscontrato la presenza di colonne non rilevanti per lo scopo del nostro progetto, le abbiamo quindi eliminate. Le feature restanti sono le seguenti, con il rispettivo tipo di dato e il range di valori assumibili.

Feature	Type	Range
Value	INT	0 - 5
Rooms	INT	0 - 5
Location	INT	0 - 5
Cleanliness	INT	0 - 5
Check in	INT	0 - 5
Service	INT	0 - 5
Business	INT	0 - 5
Overall	INT	0 - 5

Inizialmente abbiamo eseguito il metodo `plot_bar` per visualizzare i diagrammi a barre delle variabili discrete.



Come si evince dal grafico alcune variabili hanno una distribuzione dei valori non ottimale, ad esempio abbiamo circa il 60% dei valutatori ha assegnato un valore pari a -1 nella categoria business. Inoltre si può notare come in tutte le feature i valori meno frequenti sono 1 e 2, mentre il valore 5 risulta essere quasi sempre il più frequente.



Dall'immagine sopra riportata si può notare che ogni valore di overall è correlato con lo stesso valore di ogni metadato. Quindi a valori bassi di overall corrisponderanno valori bassi dei relativi metadati e viceversa.

3.1 Selezione termini per la creazione della rete bayesiana

Una volta ottenuto il dataset pulito su cui lavorare si è passati alla selezione dei termini più rilevanti, la selezione è stata effettuata tramite la funzione di peso **tf-idf** (term frequency–inverse document frequency), una funzione utilizzata in information retrieval per misurare l'importanza di un termine rispetto ad un documento. Tale funzione aumenta proporzionalmente al numero di volte che il termine è contenuto nel documento, ma cresce in maniera inversamente proporzionale con la frequenza del termine nella collezione. L'idea alla base di questo comportamento è di dare più importanza ai termini che compaiono nel documento, ma che in generale sono poco frequenti.

Dopo aver assegnato ad ogni termine il relativo peso ed aver pulito i termini si è passati all'eliminazione delle parole considerate "neutre", ovvero quelle che non esprimono emozioni particolari, per far ciò abbiamo utilizzato il dataset "Affin", contenuto nella libreria tidyverse, nel quale sono presenti circa 2000 termini; attraverso questo dataset abbiamo assegnato uno "score" compreso tra -5 e +5 ad ogni termine, -5 sono le parole che esprimono più negatività mentre +5 sono le parole che esprimono più positività. Avendo il dataset con tutte le parole tokenizzate e con il relativo "score" lo abbiamo filtrato mantenendo soltanto le parole con una rilevanza emotiva non neutra, sono state eliminate quindi tutte le parole con uno score compreso tra: -2 e +2.

Dopo questa prima fase di filtraggio dove abbiamo ottenuto il dataset delle parole che esprimono un'emozione ordinate per la loro rilevanza nel documento, abbiamo quindi preso i primi 58 termini che comparivano del dataset; questi sono i termini che verranno utilizzati successivamente per creare la rete bayesiana.

Lo scopo dello scegliere questi termini è quello di creare una rappresentazione vettoriale, di lunghezza fissa, di testi di varia lunghezza.

	Term_1	Term_2	Term_3	...	Term_58
Review_1	0	1	1	1	0
Review_2	1	1	0	0	0
Review_3	1	0	0	0	0
...	0	0	1	0	0
Review_n	1	0	0	1	1

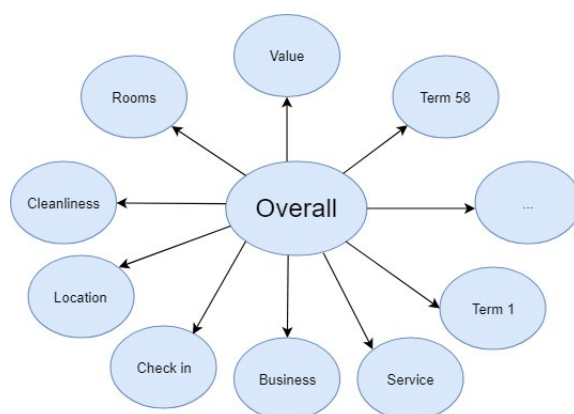
Alla fine della selezione il dataset finale risulta essere composto da un vettore di 66 elementi, di cui 8 sono i metadati e 58 sono i terms che assumeranno un valore compreso tra 0 e 1 a seconda che sia presente o meno il termine nel commento.

Feature	Type	Range
Value	INT	0 - 5
Rooms	INT	0 - 5
Location	INT	0 - 5
Cleanliness	INT	0 - 5
Check in	INT	0 - 5
Service	INT	0 - 5
Business	INT	0 - 5
Overall	INT	0 - 5
Term 1	BOOLEAN	1-0
...
Term 58	BOOLEAN	1-0

Capitolo 4

Sviluppo Modello

Per creare la rete bayesiana è stata utilizzata la funzione *model2network*, del pacchetto *bnlearn*, come nodi della rete sono abbiamo utilizzato i metadati e i 58 termini presi tramite la sentiment analysis condizionati dal nodo **Overall** che utilizziamo come variabile target essendo il voto finale del commento.



Partendo da questi 58 termini e i relativi metadati contenuti nelle varie recensioni abbiamo effettuato diversi esperimenti per studiare la performance del nostro modello con l'obiettivo di migliorarla, nello specifico si sono implementati quattro diversi modelli predittivi:

1. Per quanto riguarda il primo modello abbiamo scelto di testare le performance usando i 58 termini selezionati e l'intero dataset di training, ottenendo un'accuracy del 63%.
2. Analizzando meglio il dataset di training ci siamo accorti della presenza di valutazioni non coerenti, nello specifico casi in cui tutti i valori dei metadati corrispondevano a -1, mentre l'overall risultava essere un valore alto (da 3 a 5); oppure casi in cui il valore della variabile target corrispondeva a 0. Per questo si è deciso di eliminare

queste righe dal training set, passando da 192.000 righe a 160.000 circa. Dopo questa fase di data quality le performance del modello sono migliorate raggiungendo un'accuracy del 68%.

Accuracy	Precision	Recall	F-measure
0.68030	0.63809	0.67913	0.65198

3. Nella terza variante del modello implementato si è deciso di eliminare ulteriori righe dal training set, nello specifico quelle che presentavano 5 o più valori -1 nei metadati, riducendo così il dataset da 160.000 a 130.000 circa. Le performance di questo modello si sono rilevate le migliori con un'accuracy del 70%.

Accuracy	Precision	Recall	F-measure
0.70304	0.67658	0.67012	0.66569

4. Infine si è deciso di modificare il numero di termini usati nel modello, aumentando la soglia di filtraggio, e quindi mantenendo 74 termini. Le performance di questo modello si presentavano in linea con il terzo, probabilmente perché l'aumentare della lunghezza del vettore di termini rispetto a quello dei metadati ne diminuisce l'influenza sul modello finale.

Accuracy	Precision	Recall	F-measure
0.70173	0.67787	0.67083	0.66588

Per testare ogni modello sono state effettuate le stesse operazioni sia sul training set che sul test set.

Si è deciso di sviluppare una quinta variante del modello, tenendo in considerazione non solo il primo valore predetto ma anche il secondo, per dimostrare come a causa della grande soggettività del dominio in esame l'errore fosse al più rispetto ad una classe. Si sono ottenute le seguenti performance:

Accuracy	Precision	Recall	F-measure
0.94773	0.93721	0.94119	0.93887

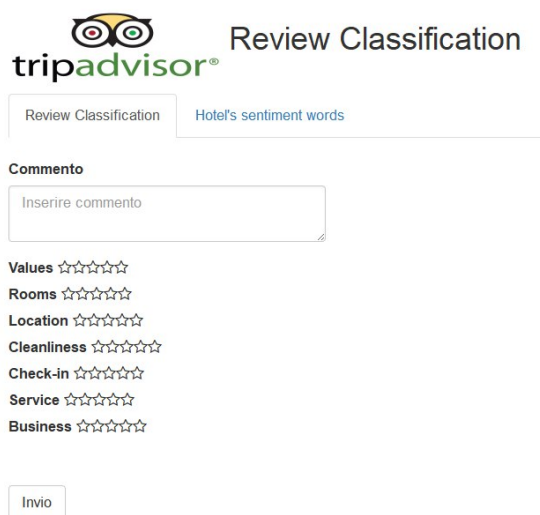
Capitolo 5

Interfaccia Grafica

L'interfaccia grafica è stata realizzata utilizzando il web application framework di R **Shiny** che permette la creazione di Web app interattive. L'applicazione consiste in due pannelli:

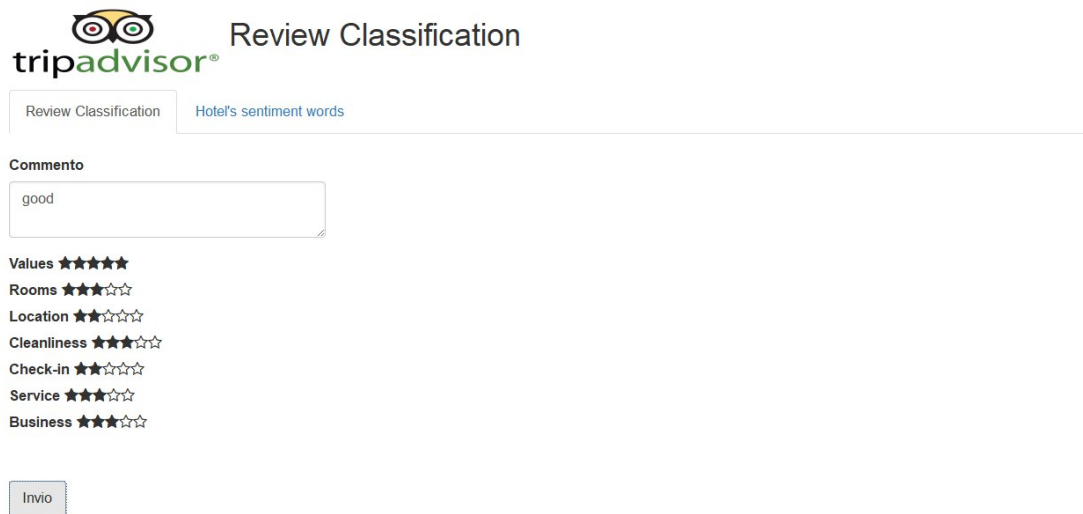
5.1 Review Classification

In questa sezione vi è una text area in cui poter scrivere il proprio commento senza limiti di caratteri. Inoltre vi sono le seguenti categorie a cui assegnare una valutazione compresa tra 0 e 5 rappresentati sotto forma di ratingstars, ovvero a 0 stelle corrisponde il valore -1 mentre ad una valutazione di 5 stelle corrisponde il valore 5.



The screenshot shows the TripAdvisor Review Classification web application interface. At the top, the TripAdvisor logo is on the left, and the title "Review Classification" is on the right. Below the logo, there are two tabs: "Review Classification" (active) and "Hotel's sentiment words". The main content area is titled "Commento" and contains a text input field with the placeholder "Inserire commento". Below the input field, there are seven categories for rating: "Values", "Rooms", "Location", "Cleanliness", "Check-in", "Service", and "Business". Each category is followed by five empty star icons for rating. At the bottom, there is an "Invio" button.

Una volta finita la recensione nel momento in cui si cliccherà su *invio* verrà visualizzato l'overall predetto dal modello.



tripadvisor® Review Classification

Review Classification Hotel's sentiment words

Commento

good

Values ★★★★★

Rooms ★★★★☆

Location ★★★★☆

Cleanliness ★★★★☆

Check-in ★★★★☆

Service ★★★★☆

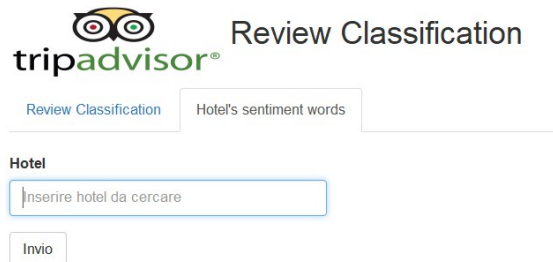
Business ★★★★☆

Invio

La valutazione del commento secondo i parametri selezionati risulta essere 3

5.2 Hotel's sentiment words

In questo pannello vi è la possibilità di poter cercare un hotel tra quelli nel dataset.



tripadvisor® Review Classification

Review Classification Hotel's sentiment words

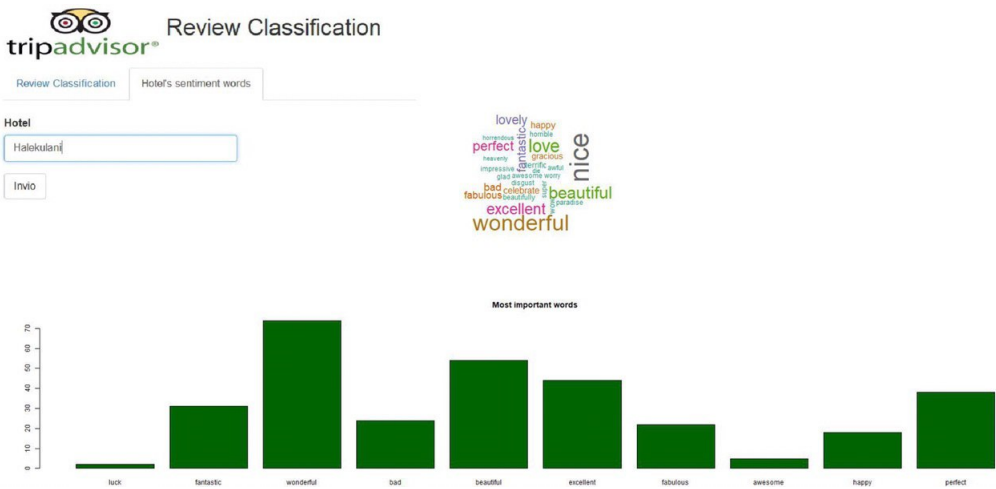
Hotel

Inserire hotel da cercare

Invio

Una volta premuto il tasto *invio* si visualizzerà a video:

- un wordcloud con le parole più frequenti tra tutti i commenti dell'hotel cercato;
- un istogramma con la frequenza delle parole più rilevanti dell'hotel.



Capitolo 6

Conclusioni

A seguito dei vari test e analisi che sono stati condotti in questo progetto possiamo concludere che abbiamo ottenuto risultati discreti in rapporto alla qualità del dataset iniziale a nostra disposizione.

Abbiamo osservato che il contributo dato dai termini contenuti nelle recensioni per la predizione del valore di Overall resta limitato, anche con l'aumentare dei termini scelti, principalmente perchè scegliendo i termini sulla base del valore di tf-idf, aumentando il range di tokens selezionati si scelgono termini con peso minore.

Data la soggettività dei commenti risulta difficile stabilire in modo univoco una classe di appartenenza, si è quindi deciso di mostrare le due classi più probabili nell'interfaccia grafica.

Nonostante la soggettività presente delle varie recensioni, il modello si comporta bene predicendo la valutazione della recensione con un'accuracy del 70%.

Al fine di migliorare il peso delle recensioni sul modello si potrebbe pensare di selezionare non token formati da singole parole ma da n-grammi.