Amazon fine food reviews

Lorgna Lorenzo 829776
Marzorati Stefano 830272

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Contenuti

Introduzione

Introduzione al progetto, domande di ricerche e dataset di partenza

01

02

Pre-processing

Text pre-processing, Text representation

Text mining tasks

Text classification, Clustering, Topic modelling

03

04

Conclusioni

Conclusioni sul progetto

01

Introduzione

Introduzione

E-Commerce

Lo shopping online è ormai diventato una costante al giorno d'oggi, una comodità che permette di risparmiare sia dal punto di vista economico che dal punto di vista del tempo.

Reviews

Parte integrante dello shopping online sono le recensioni che guidano l'utente tra la moltitudine di offerte e che permettono una scelta più consapevole basata sull'esperienza di altri clienti.

Text classification

Clustering

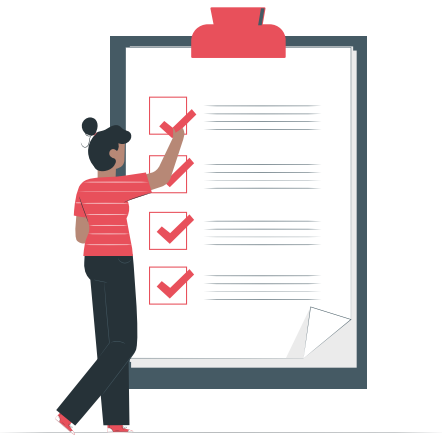
Topic modelling

Domande di ricerca

Classificazione

A partire dal testo di una recensione è possibile prevedere la valutazione positiva o negativa data dall'utente?

Andando più nello specifico è possibile determinare il punteggio in un intervallo da 1 a 5?



Clustering

Si possono individuare gruppi di recensioni con caratteristiche simili?

Topic modelling

È possibile estrarre le tematiche di maggior rilievo che emergono dalle recensioni e in caso positivo quali sono?

Dataset di partenza

Il dataset di partenza, «**Amazon Fine Foods**», contiene circa **500.000** recensioni che spaziano in un periodo di più di 10 anni, da Ottobre 1999 a Ottobre 2012.

568,454 → numero di recensioni
256,059 → numero di utenti
74,258 → numero di prodotti
10 → numero di features

- **Id**
- **ProductId**
- **UserId**
- **ProfileName**
- **HelpfulnessNumerator**
- **HelpfullnessDenominator**
- **Score**
- **Time**
- **Summary**
- **Text**

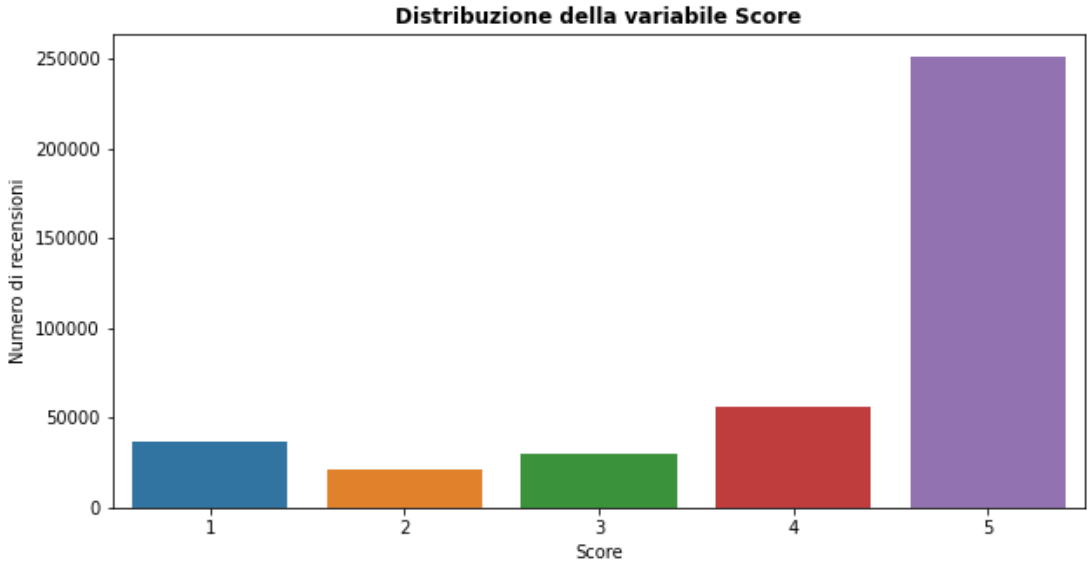
Analisi esplorativa

Data cleaning

Rimozione duplicati → eliminate circa il 30% delle osservazioni

Rimozione inconsistenze → eliminati 2 record

Gestione missing values → modifica di 3 record



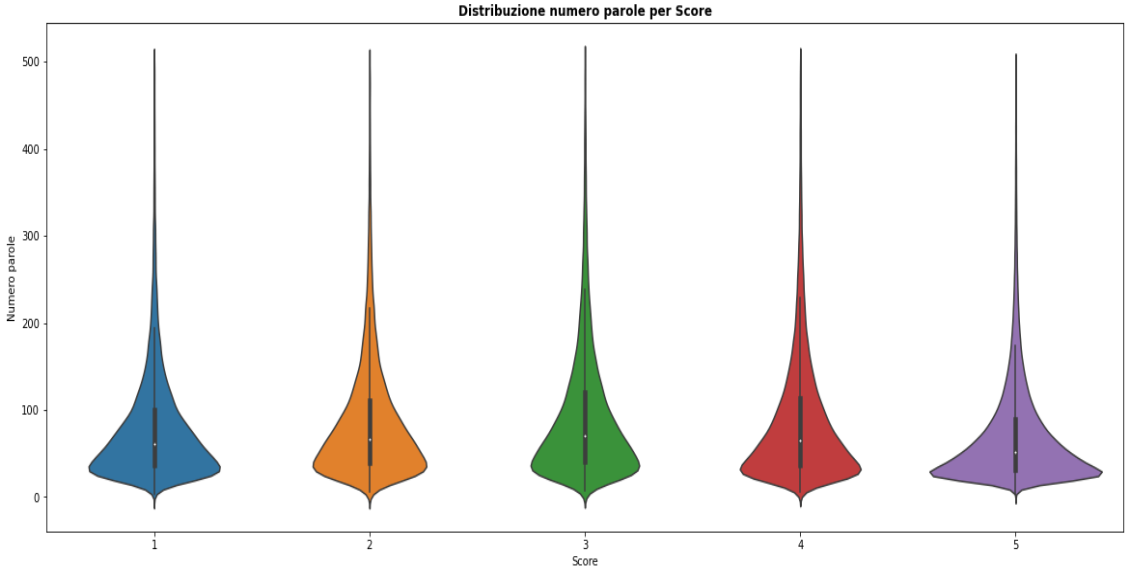
Analisi esplorativa

Data cleaning

Rimozione duplicati → eliminate circa il 30% delle osservazioni

Rimozione inconsistenze → eliminati 2 record

Gestione missing values → modifica di 3 record



02

Pre-processing

Pre-processing

Alcune operazioni che servono a migliorare l'efficienza degli algoritmi utilizzati successivamente e al tempo stesso uniformare e ripulire il testo.

Normalization

- Conversione in caratteri minuscoli
- Rimozione emoji
- Rimozione di numeri
- Espansione delle contrazioni
- Rimozione HTML
- Rimozione URL
- Rimozione di parole con 3 o più caratteri ripetuti

Tokenization

Stop words removal

Lemmatization

Text representation

Per l'applicazione di modelli è necessaria una **rappresentazione formale** del testo delle recensioni considerate.

Bag Of Words

TF-IDF

Entrambe le rappresentazioni sono composte da matrici molto grandi, dall'elevato numero di features. Per questo motivo nell'ottica di migliorare le performance in termini computazionali si è proceduto applicando una tecnica di **dimensionality reduction**, in particolare Singular value decomposition (SVD).

03

Text mining tasks

Classificazione binaria

I dati sono stati suddivisi in training set (67%) e test set (33%). Successivamente la variabile *Score* è stata portata in forma binaria e nel training set è stata gestita la class imbalance, tramite undersampling. Per rispondere alla prima domanda sono stati implementati i seguenti 3 modelli.

	Accuracy	Precision	Recall	F1-Measure	Representation
SVM	0,56	0,50	0,50	0,46	BOW
Logistic Regression	0,88	0,78	0,88	0,81	TF-IDF
Random Forest	0,76	0,66	0,77	0,67	TF-IDF

A partire dal testo di una recensione è possibile prevedere la valutazione positiva o negativa data dall'utente?

Classificazione Single-Label Multi-class

Considerando la seconda domanda di ricerca per provare a dare una risposta è stato implementato come algoritmo un modello Random Forest. Quest'ultimo è stato applicato sul testo rappresentato con TF-IDF, considerando in questo caso le 5 modalità che la feature *Score* può assumere.

	Accuracy	Precision	Recall	F1-Measure	Representation
Random Forest	0,42	0,31	0,35	0,30	TF-IDF

Andando più nello specifico è possibile determinare il punteggio in un intervallo da 1 a 5?

Clustering

Per rispondere alla terza domanda di ricerca sono stati implementati 2 algoritmi di clustering. Il numero di cluster è stato posto pari a 5. Prima dell'esecuzione degli algoritmi si è provveduto a risolvere il problema della class imbalance, text reprenentation (TF-IDF), dimensionality reduction (SVD) e infine è stata effettuata la normalizzazione dei dati.

	V-measure	ARI
Agglomerative clustering	0.006884	0.006316
K-means	0,001802	0,000394

Si possono individuare gruppi di recensioni con caratteristiche simili?

Topic modelling

Considerando l'ultima domanda di ricerca è stato utilizzato un approccio LDA per l'estrazione dei topic all'interno dei testi delle recensioni. Il numero di topic in questo caso è stato scelto pari a 5.

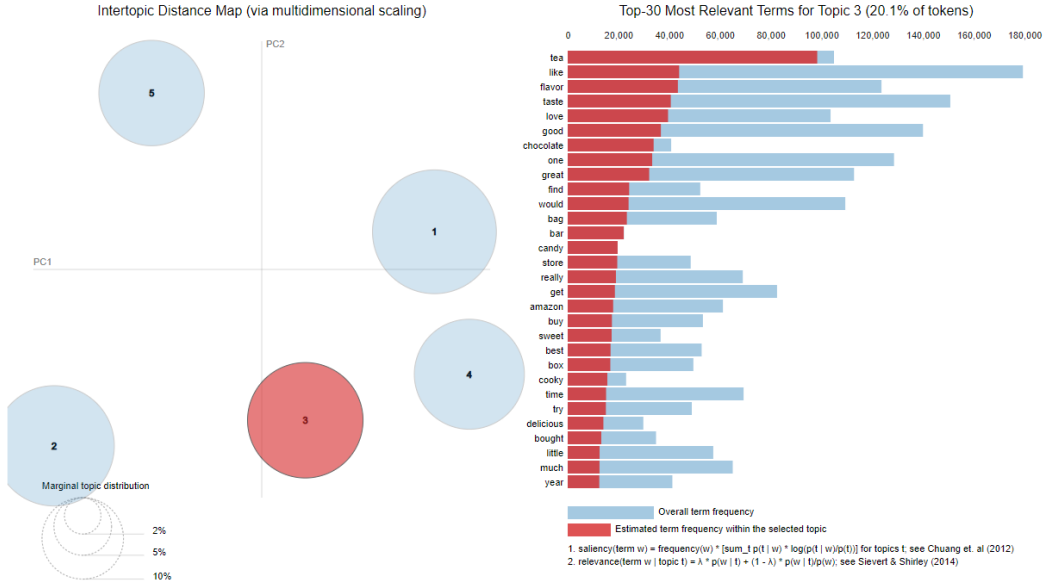
TOPIC 1: gusti/sapori

TOPIC 2: prodotti Amazon

TOPIC 3: tea

TOPIC 4: coffee

TOPIC 5: animali



È possibile estrarre le tematiche di maggior rilievo che emergono dalle recensioni e in caso positivo quali sono?

04

Conclusioni

Conclusioni

A partire dal testo di una recensione è possibile prevedere la valutazione positiva o negativa data dall'utente?

Utilizzando algoritmi di classificazione binaria si è in grado di raggiungere **performance soddisfacenti**, in particolare il miglior modello è Logistic Regression con l'utilizzo di una rappresentazione TF-IDF che presenta un valore di accuracy pari a 0.88.

Andando più nello specifico è possibile determinare il punteggio in un intervallo da 1 a 5?

Risulta **più complesso** ottenere una classificazione multiclasse. Seppur esista una relazione tra il contenuto e il voto di una recensione, gli algoritmi non sono in grado di classificarla in modo soddisfacente.

Si possono individuare gruppi di recensioni con caratteristiche simili?

É possibile identificare dei cluster ognuno **con recensioni simili** in termini di contenuto e differente dagli altri.

È possibile estrarre le tematiche di maggior rilievo che emergono dalle recensioni e in caso positivo quali sono?

Attraverso tecniche di topic modelling è possibile costatare che i topic sono direttamente legati alle **categorie di prodotti venduti oppure temi riguardanti gli E-Commerce**.

Riferimenti

[1] Stanford Network Analysis Project. Amazon Fine Food Reviews, https://www.kaggle.com/snap/amazon-fine-foodreviews, 2017.

[2] J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews, 2013.

[3] G. Pasi and M. Viviani. Text Mining and search course lecture notes and slides, a.a. 2021/2022.

Grazie per l’attenzione!