



UNIVERSITÀ DEGLI STUDI DI
MILANO-BICOCCA

F1801Q127

DATA ANALYTICS

Amazon Reviews Sentiment Analysis

Studenti:

Basso Matteo

Ferri Marco

Matricole:

807628

807130

Luglio 2019

Abstract

Lorem ipsum dolor sit amet.

Indice

1	Introduzione	5
1.1	Dominio di riferimento	5
1.2	Dataset	6
1.3	Strumenti	6
2	Basic Analysis	7
2.1	Schema	7
2.2	Dimensioni	8
2.3	Distribuzione di rating	8
2.4	Analisi temporale business-oriented	9
3	Network Analysis	12
3.1	Struttura della rete	12
3.2	Grado dei nodi	12
3.3	Misure di centralità ?	12
4	Sentiment Analysis	13
4.1	Assunzioni	13
4.2	Binarizzazione	13
4.3	Undersampling	13
4.4	Elaborazione del testo	13
4.5	Parole più usate	13
5	Sentiment Prediction	14
5.1	Pesatura dei termini (TF-IDF)	14
5.2	Termini più rilevanti	14
5.3	Modelli di predizione	14
5.3.1	Random Forest	14
5.3.2	Naive Bayes	14
5.3.3	SVM	14
5.4	Pipeline	14
6	Aspect Based Sentiment Analysis	15
6.1	Elaborazione del testo	15
6.2	Estrazione degli aspetti	15
6.3	Identificazione del sentiment	15
6.4	Risultati	15
7	Collaborative Filtering	16
7.1	Funzionamento	16

7.2	Risultati	16
8	Web Demo	17
8.1	Architettura	17
8.2	Sentiment Prediction	17
8.3	Aspect Based Sentiment Analysis	17
9	Conclusioni	18

Elenco delle figure

1	Distribuzione del campo rating	9
2	Distribuzione delle recensioni per data	9
3	Distribuzione delle recensioni per mese	10
4	Distribuzione delle recensioni per giorno della settimana	10

Elenco delle tabelle

1	Schema originale del dataset	7
2	Schema modificato del dataset	8

1 Introduzione

Lo studio ha lo scopo di condurre diversi tipi di analisi sulle recensioni del noto portale e-commerce Amazon (1). In questa sezione viene presentata una breve introduzione al problema, il dataset utilizzato e gli strumenti che sono stati impiegati per portare a termine gli obiettivi prefissati.

1.1 Dominio di riferimento

Sempre maggiore è il numero di siti web che fanno delle recensioni il proprio principale business. Si pensi ai portali dedicati alla recensione di località turistiche, film o ristoranti. Allo stesso modo, anche Amazon basa sulle recensioni parte della propria fidelizzazione clienti.

Trattandosi di dati testuali prodotti dagli utenti per valutare i prodotti acquistati, le recensioni esprimono attraverso il linguaggio naturale le impressioni dell'autore, le quali possono assumere un carattere di natura positiva o negativa. In questo contesto è inoltre pratica comune associare al proprio pensiero un punteggio che esprima una valutazione del prodotto su una scala numerica. Se questa informazione è fondamentale per i clienti della piattaforma, poiché permette di capire a colpo d'occhio quale possa essere la qualità dell'articolo che si sta considerando di acquistare, è anche vero che tale punteggio possa rappresentare un aiuto importante per riassumere in forma strutturata (e pertanto più facilmente comprensibile da un computer) l'opinione dell'autore riguardo un certo argomento. Pertanto, analizzare congiuntamente il testo di una recensione ed il punteggio ad essa associato è fondamentale per determinare una correlazione fra il linguaggio naturale e l'opinione dell'utente nei confronti del prodotto, anche detta **sentiment**. Inoltre, l'elaborazione del testo può considerare il piano morfologico del linguaggio per derivare l'opinione espressa riguardo le diverse caratteristiche del prodotto, dette **aspect**, espresse della recensione. Ciò è particolarmente utile per migliorare la qualità dei propri prodotti e ottenere quindi un vantaggio sul piano commerciale.

Infine, l'analisi delle recensioni può essere utile anche per ottenere un'approssimativa profilazione di un utente; questa può rivelarsi particolarmente rilevante dal punto di vista del marketing, ad esempio per dare suggerimenti ad altri utenti che dimostrano di avere le medesime preferenze del primo. Allo stesso modo, può essere costruito anche un sistema di suggerimenti basato su prodotti simili o solitamente venduti insieme. Tali tecniche vengono dette di **collaborative filtering**.

Il seguente elaborato si pone l'obiettivo di sperimentare con i concetti appena presentati per determinare quanto sia possibile ottenere attraverso l'analisi delle recensioni di un portale e-commerce come Amazon.

1.2 Dataset

Il dataset utilizzato per effettuare le analisi è fornito ufficialmente da Amazon e contiene recensioni redatte fra il 1996 e il 2014, per diverse categorie di prodotti (3). Poiché contenente un gran numero di recensioni che coinvolgono altrettanti utenti e prodotti, talvolta poco partecipativi all'interno del portale e-commerce, si è scelto di utilizzare ai fini del progetto una versione rielaborata del suddetto dataset, reperibile qui: <http://jmcauley.ucsd.edu/data/amazon>. Julian McAuley (6), professore dell'Università di San Diego, ha estrapolato dal dataset originale solamente i record delle recensioni riguardanti prodotti e utenti con almeno cinque recensioni ciascuno; ciò viene definito in teoria dei grafi con il termine **k-core**, cioè un sottografo in cui tutti i nodi hanno un grado almeno pari a k .

1.3 Strumenti

Per effettuare le analisi mostrate in questo elaborato si è utilizzato prevalentemente il linguaggio di programmazione open source Python, attraverso l'utilizzo di Google Colab (5) per la creazione di Jupyter Notebook interattivi. Le specifiche librerie di volta in volta utilizzate saranno presentate contestualmente alle singole analisi qualora lo si ritenesse necessario ai fini di una migliore comprensione del problema.

Per l'analisi della rete di prodotti e utenti ricavata dal dataset si è utilizzato il software Cytoscape (4), dedicato appositamente all'integrazione e visualizzazione di grafi anche molto complessi. Nell'ultima parte del documento verrà infine presentata una demo Web-based per testare con mano le potenzialità dei concetti studiati.

2 Basic Analysis

Il dataset è stato reperito in formato JSON e successivamente letto attraverso Python per la memorizzazione in una struttura adatta ad essere elaborata efficientemente dal linguaggio. A tal fine si è scelto di utilizzare la libreria **pandas** (?), dedicata all'analisi di dati anche molto voluminosi. È interessante notare come la lettura del dataset, ed in particolare la conversione di quest'ultimo dal formato JSON, sia stata una delle operazioni computativamente più *time-consuming* fra tutte quelle effettuate, a dimostrazione del fatto che **pandas** sia successivamente in grado di elaborare molto velocemente le informazioni memorizzate all'interno dei cosiddetti **DataFrame**.

2.1 Schema

Il dataset può essere descritto molto velocemente, poiché rappresentabile attraverso una singola struttura tabellare composta dai campi illustrati in tabella 2.

Tabella 1: Schema originale del dataset

Campo	Descrizione
reviewerID	ID utente
reviewerName	Nome utente
asin	ID prodotto
reviewText	Testo della recensione
summary	Titolo della recensione
helpful	Utilità della recensione
overall	Punteggio
reviewTime	Timestamp in formato string
unixReviewTime	Timestamp in formato unix

Ogni record è la rappresentazione di una singola recensione, svolta da parte di un utente per un certo prodotto nella data indicata. Mentre per quanto riguarda l'utente si è in possesso sia dell'ID che del nome (che non verrà utilizzato), il prodotto è rappresentato nel dataset solamente attraverso un ID (**asin**); ulteriori dettagli sul prodotto possono essere ricavati contattando il creatore del dataset attraverso un'apposita richiesta, che non è stata effettuata poiché tali informazioni si sono rivelate ininfluenti ai fini del progetto in esame.

Per quanto riguarda i campi relativi alla recensione, è possibile visualizzare sia il titolo che il testo del corpo, oltre al punteggio espresso su una scala numerica da 1 a 5. Poiché ciò rappresenta il sentiment associato alla recensione, questo attributo

costituirà anche la variabile target per l'analisi e l'addestramento dei modelli di machine learning.

Il campo `helpful` non è definito su un preciso dominio e pertanto si è scelto di ignorarlo. Per il timestamp è stato computato un campo `date` di tipo *datetime* e per semplificare il concetto associato ad ogni campo si è scelto di rinominarli. Al termine della trasformazione, il dataset risultante è il seguente:

Tabella 2: Schema modificato del dataset

Campo	Descrizione
<code>userID</code>	ID utente
<code>productID</code>	ID prodotto
<code>text</code>	Testo della recensione
<code>summary</code>	Titolo della recensione
<code>rating</code>	Punteggio
<code>date</code>	Timestamp in formato <i>datetime</i>

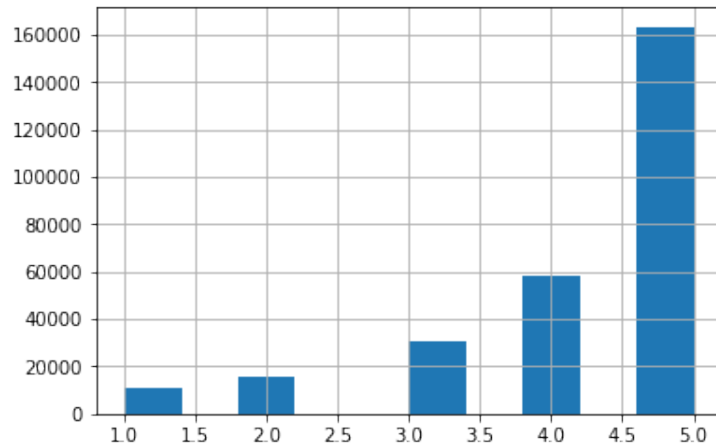
2.2 Dimensioni

Dalla fonte citata nell'introduzione, da cui è stato reperito il dataset, è possibile notare che siano presenti varie possibilità di download. Diverse categorie di prodotti sono state testate durante l'intero sviluppo dello studio, ma ai fini di redigere questo elaborato verrà considerato il dataset 5-core chiamato *Clothing, Shoes and Jewelry*, contenente recensioni relative al mercato dei vestiti, delle scarpe e dei gioielli.

Questo contiene un totale di **278.677 recensioni**, divise fra **39.387 utenti** e **23.033 prodotti**. Ciò significa una media di circa 7 recensioni ad utente e 12 recensioni per prodotto. Nel capitolo 3 verrà mostrata se tale media è anche effettivamente rappresentativa del dataset oppure diversi prodotti ed utenti presentano un numero di recensioni sbilanciato.

2.3 Distribuzione di rating

In figura 1 è possibile osservare la distribuzione del campo `rating`, che costituisce l'elemento fondamentale su cui costruire un modello supervisionato di sentiment analysis. Come è possibile osservare, questa variabile è fortemente sbilanciata sui valori 4 e 5, motivo che a fronte di alcune considerazioni future porterà il problema ad essere prima binarizzato e successivamente downsampled per ridurre il gap fra la classe positiva e quella negativa.

Figura 1: Distribuzione del campo `rating`

2.4 Analisi temporale business-oriented

Si era già accennato che la fonte fornisca recensioni per più di dieci anni di vendite. Più precisamente, il dataset qui considerato considera recensioni fra il marzo 2003 e giugno 2014 secondo la distribuzione indicata in figura 2. Pur essendo la maggior parte delle recensioni concentrata negli ultimi 3 anni, qualsiasi sia il filtro temporale applicato non c'è differenza nella distribuzione della variabile target.

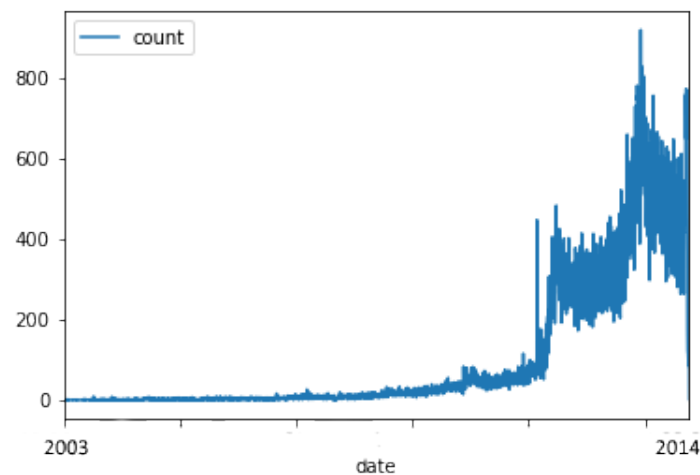


Figura 2: Distribuzione delle recensioni per data

Considerato il dominio in esame, potrebbe essere molto utile nell'ottica di prendere decisioni di business per i produttori analizzare come le recensioni si distribuiscono anche in un anno relativamente ai mesi e i giorni della settimana, per evidenziare eventuali periodi di maggiore attività su Amazon. A tal proposito sono pertanto stati prodotti i grafici in figura 3 e 4.

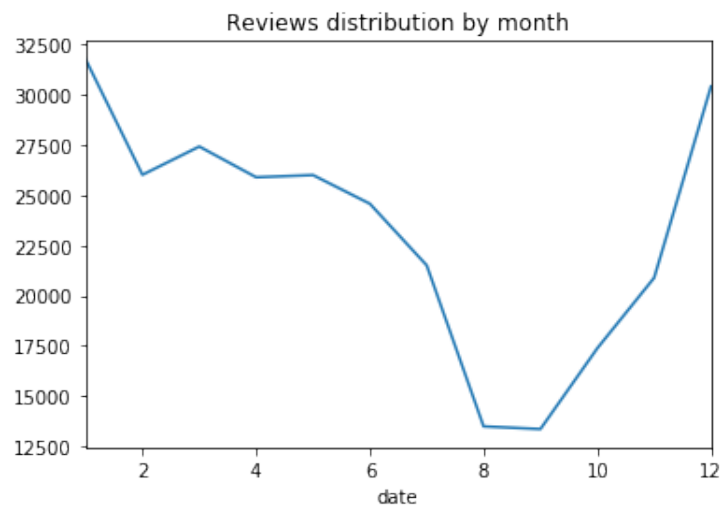


Figura 3: Distribuzione delle recensioni per mese

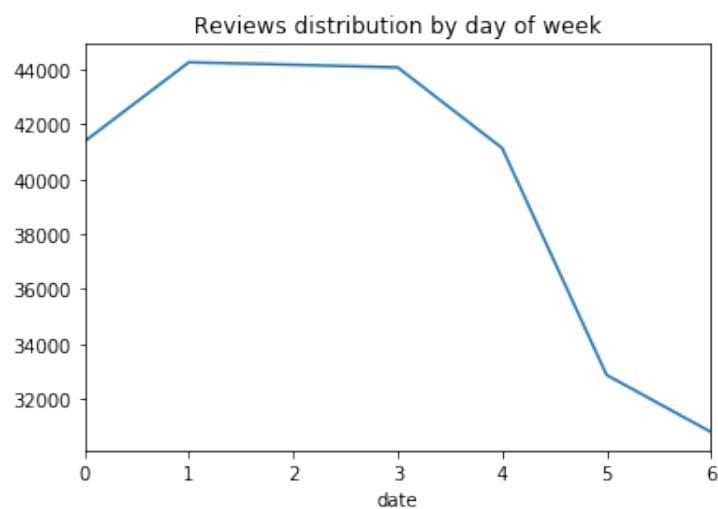


Figura 4: Distribuzione delle recensioni per giorno della settimana

Osservando la distribuzione per mese, è possibile notare come la maggiore concentrazione di recensioni si riscontri nel periodo compreso fra il black friday (Novembre) e tutte le vacanze natalizie, probabilmente proprio per via degli sconti e della necessità di comprare regali ai propri conoscenti. È durante questi mesi e quelli immediatamente precedenti che i produttori dovrebbero concentrare maggiormente le proprie campagne pubblicitarie. Tale attività decrementa gradualmente nei mesi successivi a Gennaio, per raggiungere livelli particolarmente bassi ad Agosto e Settembre.

Concentrandoci invece sui dati che riassumono l'andamento di acquisti durante la settimana, è chiaramente evidente come sabato e domenica costituiscano i giorni in cui gli utenti Amazon sono meno propensi a fare recensioni.

Analizzando quindi la distribuzione del **rating** suddiviso per mesi e giorni, emerge comunque che i risultati sono del tutto analoghi a quelli della figura 1. Da questa osservazione si potrebbe concludere che non vi sono particolari periodi temporali durante i quali i clienti sono più propensi a dare recensioni più positive o negative. Ciò su cui le aziende dovrebbero concentrarsi, da questo punto di vista, è solamente la necessità di raggiungere più persone possibili nei momenti di maggiore attività.

3 Network Analysis

3.1 Struttura della rete

3.2 Grado dei nodi

3.3 Misure di centralità ?

4 Sentiment Analysis

4.1 Assunzioni

4.2 Binarizzazione

4.3 Undersampling

4.4 Elaborazione del testo

4.5 Parole più usate

5 Sentiment Prediction

5.1 Pesatura dei termini (TF-IDF)

5.2 Termini più rilevanti

5.3 Modelli di predizione

5.3.1 Random Forest

5.3.2 Naive Bayes

5.3.3 SVM

5.4 Pipeline

6 Aspect Based Sentiment Analysis

6.1 Elaborazione del testo

6.2 Estrazione degli aspetti

6.3 Identificazione del sentiment

6.4 Risultati

7 Collaborative Filtering

7.1 Funzionamento

7.2 Risultati

8 Web Demo

8.1 Architettura

L'interfaccia web è stata sviluppata utilizzando l'architettura a 3 layer, con separazione di frontend, backend e database.

Il database utilizzato in fase di lettura è quello fornito inizialmente, senza alcuna modifica. Esso consiste quindi in un file SQLite interrogabile e modificabile semplicemente tramite un web server. Questo risulta particolarmente utile per fornire i dettagli dei giocatori ed eventualmente dei team così che l'utente possa visualizzarli e sceglierli attraverso l'opportuna interfaccia.

Per lo sviluppo del backend è stato deciso di utilizzare l'engine Javascript tramite il popolare progetto Node.js (7). Esso è in grado di agire come middleware tra il frontend e il database, separando al meglio le logiche di manipolazione del dato. É inoltre incaricato di chiamare adeguatamente lo script R per la predizione del vincitore della partita e per svolgere le inferenze richieste.

Il frontend è invece sviluppato utilizzando la libreria Javascript React.js (8)

8.2 Sentiment Prediction

8.3 Aspect Based Sentiment Analysis

9 Conclusioni

Riferimenti bibliografici

- [1] Amazon. URL: <https://www.amazon.it/>.
- [2] Amazon reviews 5-core dataset source. URL: <http://jmcauley.ucsd.edu/data/amazon/>.
- [3] Amazon reviews original dataset source. URL: <https://s3.amazonaws.com/amazon-reviews-pds/readme.html>.
- [4] Cytoscape for network integration, analysis and visualization. URL: <https://cytoscape.org/>.
- [5] Google colab. URL: <https://colab.research.google.com/>.
- [6] Julian mcauley. URL: <http://jmcauley.ucsd.edu/data/amazon/>.
- [7] Node.js. URL: <https://nodejs.org/it/>.
- [8] React. URL: <https://reactjs.org/>.