

DATA ANALYTICS

Amazon

Ceccon Martina 817613
Costantino Chantal 860621
Maggio Vittorio 817034

Indice

Elenco delle tabelle	2
Elenco delle figure	3
1 Sentiment Analysis	4
1.1 Richiami teorici	4
1.2 Textual Preprocessing	5
Bibliografia	6

Elenco delle tabelle

Elenco delle figure

1.1	<i>Distribuzione delle recensioni per ogni categoria.</i>	5
-----	---	---

Capitolo 1

Sentiment Analysis

???????????

1.1 Richiami teorici

L'analisi del *sentiment* o *sentiment analysis* (nota anche come *opinion mining*) è un campo dell'elaborazione del linguaggio naturale che si occupa di costruire sistemi per l'identificazione ed estrazione di opinioni dal testo, basandosi sui principali metodi di linguistica computazionale e di analisi testuale. Per identificare le informazioni soggettive che denotano opinioni, è necessario determinare la polarità insita in esse. Questa può essere di tre tipi: positiva, negativa o neutra.

polarità positiva: l'informazione che ricaviamo è positiva.

polarità negativa: l'informazione che ricaviamo è negativa.

polarità neutra: non ricaviamo alcuna informazione sulla polarità. Questo può avvenire per due diversi motivi:

- le opinioni positive sono bilanciate da quelle negative;
- non si manifesta alcun sentimento.

Per il calcolo della polarità esistono diversi tipi di approccio: basato sul lessico, supervisionato, semi-supervisionato e non supervisionato.

Approccio basato sul lessico: la polarità viene calcolata analizzando la frase.

Approccio supervisionato: la polarità viene calcolata utilizzando le tecniche del *Machine Learning*. Inserisco prima le parole in cui la polarità è esplicita, poi aggiungo le altre per fare inferenza.

Approccio semi-supervisionato: ???????

Approccio non supervisionato: ???????

Vedremo che per il nostro non è stato necessario andare a calcolare il *sentiment*; in quanto un attributo si prestava bene per questo tipo di analisi.

1.2 Textual Preprocessing

Il nostro primo obiettivo è quello di studiare l'intrattenimento, in particolare quali siano i prodotti più recensiti all'interno delle nostre quattro categorie. Abbiamo scelto di effettuare lo studio su un numero limitato di prodotti, riducendoli a duecento, e di analizzare la loro distribuzione in questo sottoinsieme. I risultati ottenuti sono presentati nella Figura 1.1

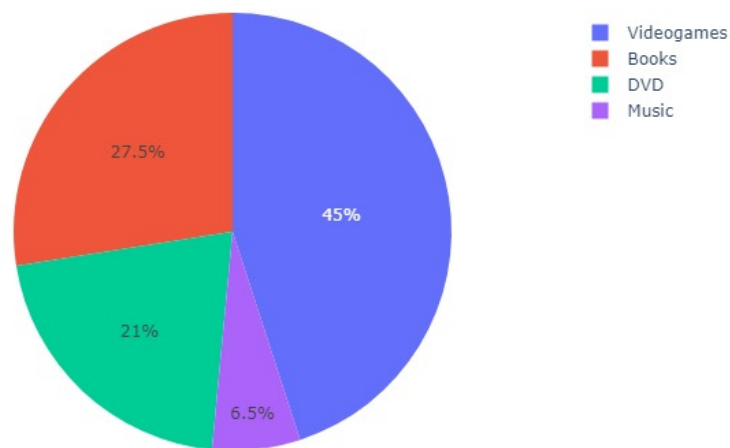


Figura 1.1: *Distribuzione delle recensioni per ogni categoria.*

Per poter eseguire delle analisi consistenti sono indispensabili una serie di operazioni atti a uniformare il *dataset*. Con il termine "uniformare" intendiamo riscrivere i dati seguendo delle specifiche regole comuni. Nel seguito ogni procedura verrà descritta nel dettaglio.

Bibliografia