

DATA ANALYTICS

Amazon

Ceccon Martina 817613
Costantino Chantal 860621
Maggio Vittorio 817034

Indice

Elenco delle tabelle	3
Elenco delle figure	4
1 Richiami teorici alla sentiment analysis	6
2 Sentiment Analysis	8
2.1 Analisi esplorativa: distribuzione del sentiment nel corso degli anni	8
2.1.1 Distribuzione del sentiment negli anni: Libri	9
2.1.2 Sentiment negli anni: Dvd	9
2.1.3 Sentiment negli anni: Videogiochi	10
2.1.4 Sentiment negli anni: Musica	10
2.2 Studio sull'intrattenimento	11
2.2.1 Wordclouds libri	15
2.2.2 Wordclouds dvd	18
2.2.3 Wordclouds videogiochi	19
2.3 Correlazione tra prezzo e sentiment	20
2.3.1 Categoria libri	21
2.3.2 Categoria dvd	22
2.3.3 Categoria videogiochi	23
2.3.4 Categoria musica	23
2.4 Utilità delle recensioni per gli utenti	23
2.4.1 Analisi dei risultati tramite box plot	24
3 Analisi tramite modelli supervisionati	26
3.1 Logistic Regression	26
3.2 Support Vector Machine	28
3.3 Reti neurali	30
4 Network Analysis	33
4.1 Analisi sulla rete: videogiochi	33
4.1.1 Metodo Louvain	33
4.2 Identificazione comunità	34

5 Conclusioni	35
6 TODO	36
Bibliografia	37

Elenco delle tabelle

3.1	Logistic Regression: Matrice di confusione	27
3.2	Logistic Regression: Precision	27
3.3	Accuratezza	27
3.4	Recall	28
3.5	Logistic Regression: F1-mesaure	28
3.6	SVM: Matrice di confusione	29
3.7	SVM: Precision	29
3.8	SVM: Accuratezza	29
3.9	SVM: Recall	30
3.10	SVM: F1-mesaure	30
3.11	NN: Matrice di confusione	31
3.12	Precision	31
3.13	NN: Accuratezza	31
3.14	NN: Recall	32
3.15	NN: F1-mesaure	32

Elenco delle figure

2.1	Distribuzione sentiment negli anni - Libri.	9
2.2	Distribuzione sentiment negli anni - Dvd.	10
2.3	Distribuzione sentiment negli anni - Videogiochi.	11
2.4	Distribuzione sentiment negli anni - Musica.	12
2.5	Distribuzione delle recensioni per ogni categoria.	13
2.6	Primi cinque libri con più recensioni positive	13
2.7	Primi cinque libri con più recensioni negative	14
2.8	Primi cinque dvd con più recensioni positive	14
2.9	Primi cinque dvd con più recensioni negative	15
2.10	Primi cinque videogiochi con più recensioni positive	15
2.11	Primi cinque videogiochi con più recensioni negative	16
2.12	Wordclouds libri	16
2.13	Wordclouds dvd	18
2.14	Wordclouds videogiochi	19
2.15	Libri: studio sulle recensioni medie dei prodotti	22
2.16	Dvd: studio sulle recensioni medie dei prodotti	22
2.17	Videogiochi: studio sulle recensioni medie dei prodotti	23
2.18	Musica: studio sulle recensioni medie dei prodotti	24
2.19	Utilità delle recensioni negative e positive per ogni categoria.	25

Introduzione

???Informazioni sul dataset??????

Capitolo 1

Richiami teorici alla sentiment analysis

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.

Capitolo 2

Sentiment Analysis

Nel corso del capitolo verranno mostrate tutte le analisi effettuate relative al *sentiment*; in particolare quali siano le informazioni ricavate durante questo studio. La parte mostrata nel seguito sarà così strutturata; inizialmente verrà eseguita un'analisi esplorativa su uno dei due *dataset*, per comprendere la distribuzione del sentiment nel corso degli anni, si procederà poi con l'indicazione e la rappresentazione delle informazioni ricavate sempre sulla base del *sentiment* e infine il confronto dei valori di *sentiment* dati in *input* con quelli ottenuti dall'addestramento di alcuni modelli appartenenti al mondo del *Machine Learning*.

2.1 Analisi esplorativa: distribuzione del sentiment nel corso degli anni

Prima di procedere con un'analisi sui *dataset* per ricavare informazioni circa alcune precise domande, abbiamo scelto di svolgere un'analisi esplorativa atta a visualizzare l'andamento del *sentiment* degli utenti nel corso degli anni. In pratica quindi, partendo da tutte le recensioni assegnate a ogni prodotto, siamo andati a stimare come queste si distribuissero in un dato *range* temporale.

Come primo passaggio abbiamo creato un *dataset*, sottoinsieme di quello relativo a tutte le recensioni, e in questo abbiamo scelto di mantenere solamente quattro categorie:

- books
- dvd
- videogames
- music

A questo punto all'interno di ogni categoria abbiamo creato tre diversi gruppi formati da tutte le recensioni positive per il primo, tutte quelle negative per il secondo e infine neutre per il terzo. Per ogni gruppo di una stessa categoria si è studiata la distribuzione nel corso degli anni tramite un grafico, in cui sull'asse delle ordinate è rappresentato il numero delle recensioni; su quelle delle ascisse gli intervalli temporali. Si noti che i due valori posti agli estremi sull'asse delle ascisse, rispettivamente 2010 e 2019 registrano dei valori bassi di recensioni. Questo comportamento è dovuto al fatto che probabilmente la campionatura non è rappresentativa di tutto l'annata; pertanto è bene non considerare l'andamento del grafico nei pressi dei due valori.

2.1.1 Distribuzione del sentiment negli anni: Libri

Dalla Figura 2.1 è chiaro che il numero di recensioni positive negli anni è stato sempre molto più elevato rispetto a quello delle recensioni negative o neutre, fino a essere quasi il quintuplo più grande intorno al 2018.

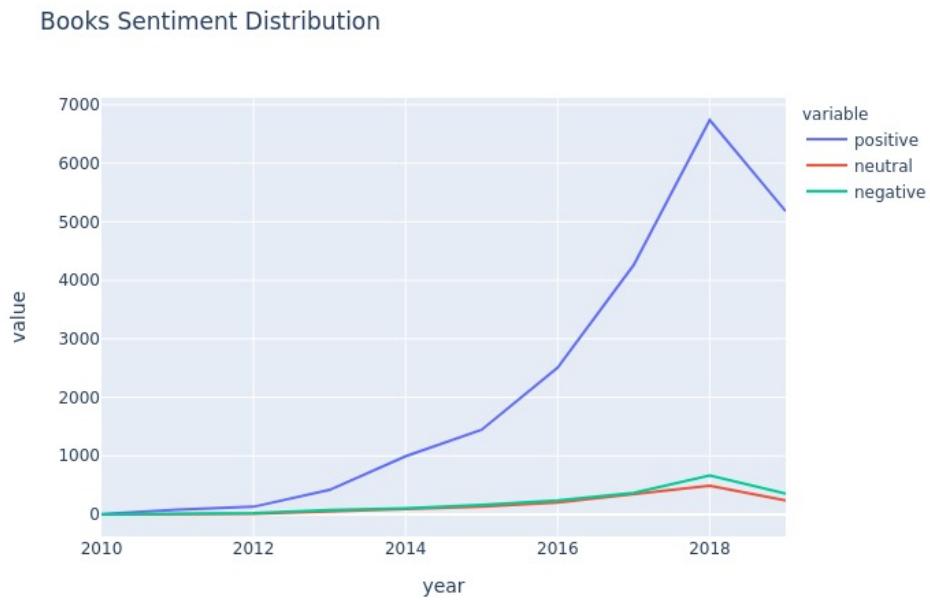


Figura 2.1: Distribuzione sentiment negli anni - Libri.

2.1.2 Sentiment negli anni: Dvd

A differenza del caso precedente, l'andamento delle recensioni positive ha un andamento irregolare (Figura 2.2), registrando, subito dopo un massimo

locale, una prima fase di decrescita e quasi immediata ricrescita. Le recensioni neutre e negative hanno invece valori bassissimi; in entrambe il massimo si registra però sempre nel 2018, anno di massimo globale per le recensioni positive.



Figura 2.2: Distribuzione sentiment negli anni - Dvd.

2.1.3 Sentiment negli anni: Videogiochi

La categoria in questione è quella con più esemplari in assoluto (Figura 2.3). Come per i libri, anche in questo caso si ha un andamento quasi esponenziale per le recensioni positive, che raggiungono il loro apice nel 2018, come del resto anche quelle negative e neutre. Un aspetto interessante è che si registra una forte crescita solamente a partire dal 2014.

2.1.4 Sentiment negli anni: Musica

Il gruppo in esame è quello che registra meno recensioni positive, solamente 2500. L'andamento inoltre presenta un'irregolarità; nel 2014, infatti, si registra un massimo locale, mentre il massimo globale si riscontra nel 2018, anticipato nel 2015 da un minimo locale. Per le altre tipologie di recensioni andiamo un andamento crescente fino al 2016, dopodiché raggiunge la stazionarietà nelle annate successive.

Videogames Sentiment Distribution

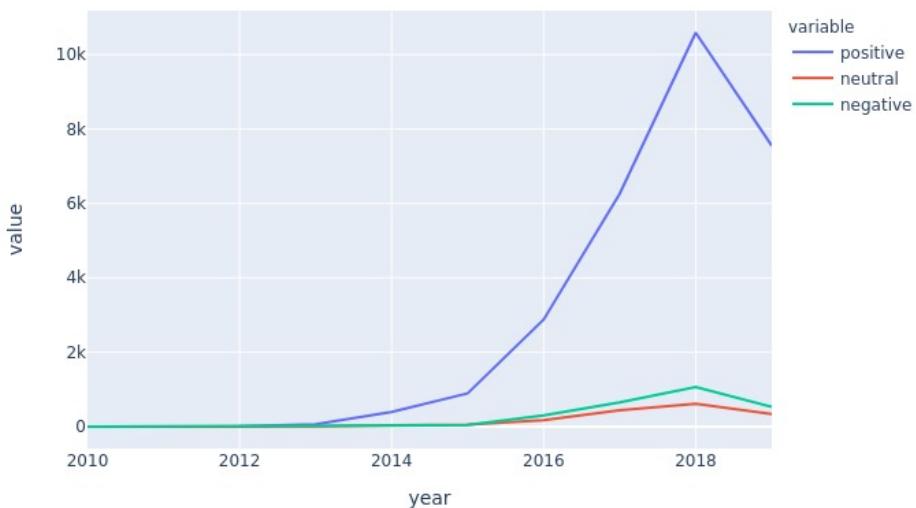


Figura 2.3: Distribuzione sentiment negli anni - Videogiochi.

2.2 Studio sull'intrattenimento

Il nostro primo obiettivo è quello di studiare l'intrattenimento, in particolare quali siano i prodotti più recensiti all'interno del *dataset* riguardante le recensioni. Come primo passo abbiamo quindi scelto di effettuare lo studio su un numero limitato di prodotti, riducendoli a duecento, e di analizzare la loro distribuzione. I risultati ottenuti sono presentati nella Figura 2.5.

Dalla figura è possibile notare che, in percentuale, i prodotti più recensiti appartengono alla categoria dei videogiochi (45%), seguiti dai libri (27.5%), dai dvd (21%) e dalla musica (6.5%). Quest'ultima categoria è davvero esigua perché delle analisi diano risultati consistenti e si possano ricavare informazioni utili; inoltre poiché lo studio su cui abbiamo posto l'attenzione riguardava i più popolari, non aveva senso esaminare dei prodotti quasi privi di recensioni. Da questa considerazione è derivata la scelta di escludere **music** e procedere ad analizzare l'intrattenimento solo sulle tre categorie più popolari, quindi **videogames**, **books** e **dvd**. Arrivati a questo punto il nostro studio è stato diviso in due diversi passaggi, durante i quali abbiamo cercato di rispondere a due domande ovvero quali siano i prodotti più recensiti e perché proprio loro.

Primo passaggio: distribuzione del sentiment

Come già accennato, per poter valutare la polarità di una recensione è

Music Sentiment Distribution

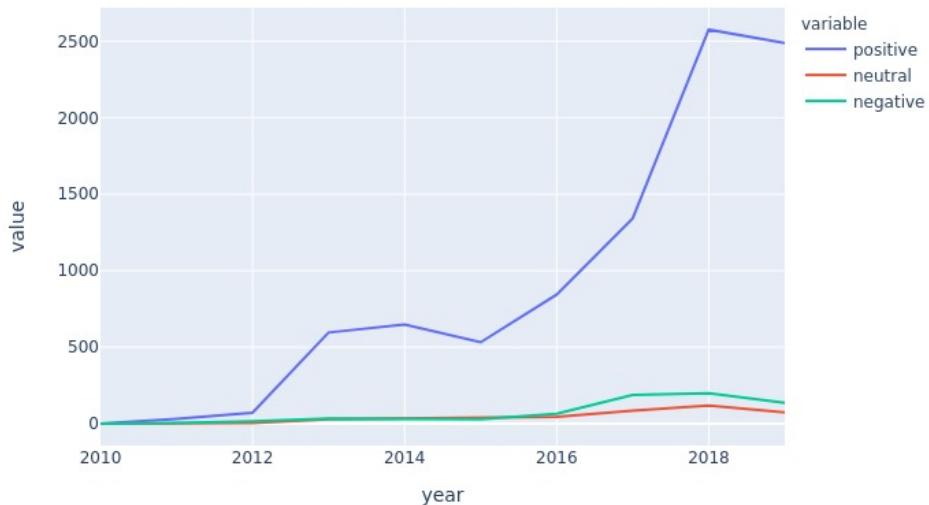


Figura 2.4: Distribuzione sentiment negli anni - Musica.

necessario calcolare il *sentiment*. Abbiamo visto che questo calcolo può essere eseguito secondo diverse metodologie, tuttavia il nostro *dataset* relativo ai prodotti, presentava al suo interno un campo che ben si prestava a questo tipo di analisi. L'attributo in questione è quello delle stelle. *Amazon* infatti per ogni recensione riferita a un determinato prodotto associa una valutazione in stelle da 0 a 5. Questo ha permesso di effettuare un conteggio per ogni prodotto di tutte le sue recensioni, dividendole in positive neutre e negative, secondo lo schema di seguito.

- **positive:** la recensione aveva un punteggio di maggiore di tre stelle.
- **neutre:** la recensione aveva un punteggio uguale a tre stelle.
- **negative:** la recensione aveva un punteggio minore di tre stelle.

Procedendo secondo queste modalità, preso per esempio il prodotto "Zelda", con 200 recensioni, si calcolano quante di queste sono risultate positive negative oppure neutre e un possibile risultato potrebbe essere costituito da 100 recensioni positive, 70 negative e 30 neutre. Ottenuto questo elenco abbiamo calcolato la distribuzione probabilistica associata alle tre diverse polarità. In particolare per ognuna associata a uno specifico prodotto, abbiamo effettuato un rapporto tra

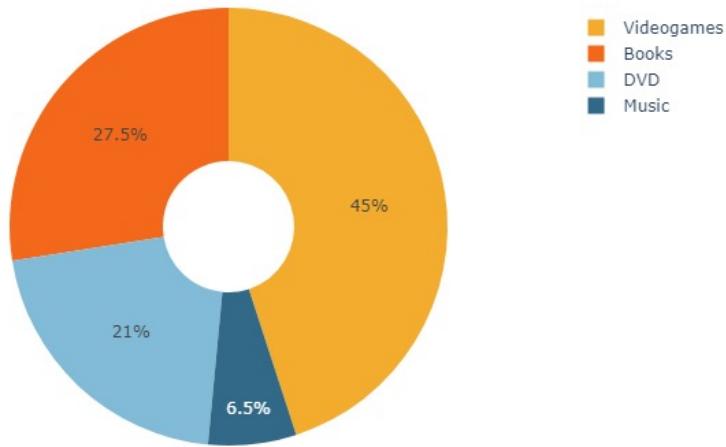


Figura 2.5: Distribuzione delle recensioni per ogni categoria.

il conteggio delle recensioni positive e negative, escludendo le neutre (dato non rilevante per la nostra analisi) e il numero di recensioni totali. La distribuzione probabilistica della polarità ottenuta è stata quindi combinata al sottoinsieme dei prodotti più recensiti/più popolari e il risultato di questo *match* è stata l'identificazione per ogni categoria dei primi cinque prodotti più popolari suddivisi in due elenchi. Il primo contenente i prodotti valutati più negativamente e nel secondo quelli valutati più positivamente.

I risultati ottenuti sono visibili nelle Figure 2.6 2.7, 2.8, 2.9, 2.10, 2.11.

ID	Numero di recensioni	Titolo del libro	Recensioni positive	Recensioni negative	Recensioni neutrali
8817055778	262	Per questo mi chiamo Giovanni. Da un padre a un figlio il racconto della vita di Giovanni Falcone	99.24%	0.38%	0.38%
8867143336	298	Il Gruffalò. Ediz. illustrata	97.99%	1.34%	0.67%
8858012534	174	I colori delle emozioni. Ediz. illustrata (pop-up)	97.70%	0.00%	2.30%
8868364026	150	25 grammi di felicità . Come un piccolo riccio può cambiarti la vita	97.33%	1.33%	1.33%
8817109444	122	Metodo universitario. Come studiare meglio in meno tempo e superare gli esami senza ansia	96.72%	2.46%	0.82%

Figura 2.6: Primi cinque libri con più recensioni positive

Secondo passaggio: : analisi delle parole più usate

ID	Numero di recensioni	Titolo del libro	Recensioni positive	Recensioni negative	Recensioni neutrali
8820067706	130	After	46.15%	40.77%	13.08%
8869876616	131	Il magico potere del riordino. Il metodo giapponese che trasforma i vostri spazi e la vostra vita	63.36%	29.01%	7.63%
8830100463	322	La verità sul caso Harry Quebert	58.70%	28.57%	12.73%
8891817597	235	Divertiti con Lui e Sofi. Il Fantaribro dei Me contro Te	62.55%	28.09%	9.36%
8890955503	556	Vivere 120 anni. Le verità che nessuno vuole raccontarti	71.40%	20.68%	7.91%

Figura 2.7: Primi cinque libri con più recensioni negative

ID	Numero di recensioni	Titolo del dvd	Recensioni positive	Recensioni negative	Recensioni neutrali
B004XKRO6Q	1166	The Lord of the Rings - The Motion Picture Trilogy, Extended Edition	98.00%	1.00%	1.00%
B0041KWD6Y	122	Pomi d'ottone e manici di scopo	97.00%	2.00%	2.00%
B078H2XDQ2	264	Coco	96.00%	2.00%	2.00%
B00AGD6MUS	196	Quasi amici	95.00%	2.00%	3.00%
B0041KY2Q8	146	Gli Aristogatti	95.00%	3.00%	2.00%

Figura 2.8: Primi cinque dvd con più recensioni positive

Avendo risposto al primo quesito ci siamo potuti soffermare sulla seconda interrogazione, ovvero perché fossero risultati proprio questi prodotti. Siamo così andati alla ricerca delle parole più utilizzate all'interno delle recensioni (negative e positive), cercando una corrispondenza tra queste e i prodotti migliori. Tuttavia per poter eseguire questo confronto sono state necessarie delle operazioni atte a uniformare il *dataset*; i dati infatti non sempre sono puliti, in essi sono presenti errori di battitura, intere frasi scritte in maiuscolo, eccessiva punteggiatura, etc. Ecco dunque il motivo del termine "uniformare", intendiamo con esso il processo di riscrittura delle frasi seguendo degli specifici passaggi, che saranno descritti nel seguito.

tokenizzazione: suddivide un testo in singole parole, i *tokens*, che saranno utilizzati per altri tipi di analisi o attività.

standardizzazione: riscrittura delle parole da *upper case* a *lower case*.

rimozione delle stopwords, parole comuni prive di significato, ma che ricorrono spesso all'interno della frasi.

rimozione delle cifre numeriche

rimozione della punteggiatura, tuttavia nella nostra implementazione questa fase è subentrata all'interno della **tokenizzazione**.

ID	Numero di recensioni	Titolo del dvd	Recensioni positive	Recensioni negative	Recensioni neutrali
B0781YXHRH	152	Star Wars: Gli Ultimi Jedi	64.00%	25.00%	11.00%
B01BMCRNFY	135	Deadpool	74.00%	18.00%	8.00%
B019C00JYA	246	Star Wars: Il Risveglio della Forza	74.00%	14.00%	12.00%
B07C6FCK6G	174	Harry Potter Collection (Standard Edition) (8 Dvd)	86.00%	12.00%	2.00%
B0071AO594	157	Rocky - La Collezione Completa	80.00%	11.00%	9.00%

Figura 2.9: Primi cinque dvd con più recensioni negative

ID	Numero di recensioni	Titolo del videogioco-accessorio	Recensioni positive	Recensioni negative	Recensioni neutrali
B072N597H5	453	Super Mario Odyssey - Nintendo Switch	98.68%	0.88%	0.44%
B01N7QUSQ3	439	The Legend of Zelda: Breath of the Wild - Nintendo Switch	98.41%	0.23%	1.37%
B01MS894C1	406	Mario Kart 8 Deluxe - Nintendo Switch	98.28%	1.23%	0.49%
B01M8MO8XS	125	Red Dead Redemption 2 - Xbox One	97.60%	0.80%	1.60%
B07BZV8R5R	151	Super Smash Bros Ultimate - Nintendo Switch	97.35%	1.99%	0.66%

Figura 2.10: Primi cinque videogiochi con più recensioni positive

stemming: processo di riduzione delle parole flesse (o talvolta derivate) alla loro forma di origine, base o radice.

Al termine del processo, l'*output* risultante era composto da un elenco di parole scritte in Italiano corretto, privo di segni di punteggiatura o numeri, scritto in forma minuscola, ridotto alla radice. A questo punto è stato quindi possibile effettuare un'operazione di visualizzazione delle parole più usate all'interno delle recensioni, per comprendere il motivo per cui proprio quei prodotti presenti nella lista rientrino nei migliori, per entrambe le polarità. La tecnica visiva utilizzata è quella dei *wordclouds*, che consiste nella raccolta delle parole più usate nelle recensioni (*word*), rappresentate per mezzo di *cloud*, ha permesso di confrontare le recensioni positive e negative per ogni categoria e di comprendere che cosa determinava l'appartenenza dei prodotti all'insieme dei "migliori" o dei peggiori.

2.2.1 Wordclouds libri

Nella Figura 2.12 sono mostrati i *wordclouds* ottenuti per la categoria libri. In particolare l'immagine di sinistra rappresenta le parole più usate nelle recensioni positive, mentre a destra sono rappresentate quelle negative.

ID	Numero di recensioni	Titolo del videogioco-accessorio	Recensioni positive	Recensioni negative	Recensioni neutrali
B0746PN4PZ	198	Nacon Revolution Pro Controller 2, Nero - PlayStation 4	59.60%	30.81%	9.60%
B06XWY4CQP	140	Destiny 2 + DLC Esclusivo Amazon - PlayStation 4	55.71%	27.86%	16.43%
B0746R8TKN	130	Nacon Compact Controller, Blu - PlayStation 4	62.31%	24.62%	13.08%
B01BVEEJG4	580	Gran Turismo Sport - PlayStation 4	68.45%	23.10%	8.45%
B00499DB7W	436	Two Dots Universal Pro Driving Simulator Supporto universale ripiegabile	57.34%	21.33%	21.33%

Figura 2.11: Primi cinque videogiochi con più recensioni negative

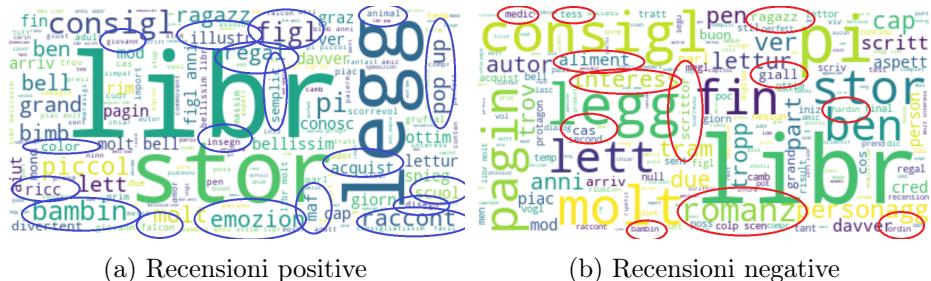


Figura 2.12: Wordclouds libri

Recensioni positive: nello schema sono rappresentate con dimensione maggiore le parole **libr**, **stori**, **legg**, a dimostrazione del fatto che queste sono in assoluto le parole più frequenti all'interno delle recensioni degli utenti. Non sorprende che i tre termini si riferiscano alla categoria selezionata, nel caso attuale quella dei libri; vedremo che questo comportamento ricorrerà sempre durante l'analisi tramite *wordclouds*. Concentriamoci invece sulle parole che racchiudono informazioni, ricordando che l'obiettivo primario è comprendere perché i libri di Figura 2.6 siano risultati i migliori.

- **ricc, figl, falcon, giovann, "emozion", pop up:**
sono tutte parole presenti in alcuni dei titoli della lista dei migliori, in ordine "*Per questo mi chiamo Giovanni. da un padre a un figlio il racconto della vita di Giovanni Falcone*", "*I colori delle emozioni (pop-up)*", "*25 grammi di felicità - Come un piccolo riccio può cambiarti la vita*".
 - **maf:**
termine che se esteso indica la mafia, riferito quindi al primo libro dell'elenco.
 - **illistr, bambin, color, semplic, disegn, raccont, animal :**
sono tutti termini riguardanti l'infanzia. Da queste informazioni

comprendiamo che la maggior parte delle recensioni commenta positivamente i libri per bambini, che sono semplici, colorati, ricchi di disegni e di racconti basati sugli animali. Quindi questi vocaboli sono riferiti per lo più al secondo e al terzo libro dell'elenco dei migliori; tuttavia l'ultimo termine può essere esteso anche al quarto prodotto.

- **insegn, scuol:**

sono vocaboli riferiti all'ambiente scolastico; quindi probabilmente riferito all'ultimo libro, criticato positivamente poiché insegna tematiche relative alla scuola.

- **acquist, regal:**

molto probabilmente alcuni libri, acquistati per un regalo, hanno suscitato delle impressioni positive nell'utente, che ha quindi recensito positivamente.

Recensioni negative: nello schema sono rappresentate con dimensione maggiore le parole **libr, pagin, legg**, a dimostrazione del fatto che queste sono in assoluto le parole più frequenti all'interno delle recensioni degli utenti. Notiamo che due su tre sono parole già trovate all'interno dell'elenco delle parole positive, poiché anche in questo caso sono termini utili a determinare la categoria in questione. Studiamo ora le altre parole in relazione all'elenco già stilato (Figura 2.7).

- **hardin, tess, romanz:**

sono i termini più esplicativi dell'insieme, indicano infatti i nomi dei protagonisti del romanzo "*After*"

- **ordin, cas:**

sono parole usate nel secondo prodotto dell'elenco

- **scrittor, giall:**

si riferiscono probabilmente a "*La verità del caso di Harry Quebert*", un giallo che vede protagonista uno scrittore. **bambin:**

fra tutti i prodotti dell'elenco quello che certamente si adatta alle linee giovanili è sicuramente il quarto libro dell'elenco, pieno di attività creative e non proposte ai più piccini. **aliment:**

vocabolo riferito a "*Vivere 120 anni. la verità che nessuno vuole raccontarti*". **medic, ragazz, interessant:**

tra i proposti sono quelli che più di tutti racchiudono le possibili cause di una critica. Il libro, presupponibilmente l'ultimo dell'elenco, probabilmente aveva un approccio troppo rivolto alla medicina o forse troppo poco. Il secondo è forse più rivolto ad "*After*", criticato spesso per rivolgersi esclusivamente a un pubblico femminile. Più generico, invece, è il vocabolo **interessante**, che ben si adatta a ognuno dei prodotti proposti.

2.2.2 Wordclouds dvd

Nella Figura 2.13 sono mostrati i *wordclouds* ottenuti per la categoria dvd. In particolare l'immagine di sinistra rappresenta le parole più usate nelle recensioni positive, mentre a destra sono rappresentate quelle negative.

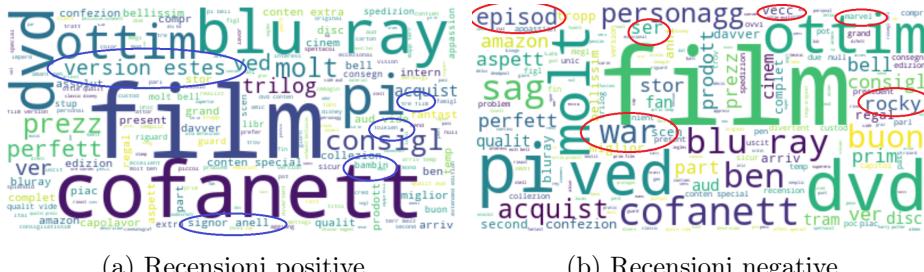


Figura 2.13: Wordclouds dvd

Recensioni positive: nello schema sono rappresentate con dimensione maggiore le parole **dvd**, **film**, **blu ray**, riferite genericamente alla categoria presa in esame.

- **signore anell, tolkien, tre film:**

I termini selezionati sono accomunati dalla tematica "*Signore degli Anelli*", trilogia ispirato all'omonimo romanzo scritto da Tolkien, che ha raccolto molti consensi, dato che più parole lo caratterizzano.

- **bambin:**

I dvd criticati positivamente sono stati quelli adatti a un giovane pubblico. Da notare che il termine risponde ancora alle caratteristiche del "*Signore degli Anelli*".

- **version estes:**

Tra i tanti comprati o visti, i dvd in versione estesa hanno riscontrato un consenso positivo.

Si noti che le parole trovate ben si accordano con i risultati trovati nei passaggi precedenti (Figura 2.8), dove il miglior prodotto era la trilogia del "*Signore degli anelli*", seguito poi da altri film adatti ai bambini ("*Coco*", terzo posto, e "*Gli Aristogatti*", ultimo posto, sono addirittura dei cartoni animati).

Recensioni negative: nello schema i termini generici riscontrabili con la categoria in questione sono **film**, **dvd**, **blu ray**. Poiché nella Figura 2.9, abbiamo trovato cinque prodotti di diversa natura come maggiormente criticati, ci aspettiamo di trovare termini che li identifichino e che argomentino le recensioni negative.

- war, marvel, rocky:

sono tutti termini utili a identificare il dvd in questione. Come ci aspettavamo, il primo di questi si riferisce al film di "Star Wars", il primo peggio criticato, il secondo invece riguarda "Deadpool", un film della marvel e il terzo del film *"Rocky"*.

- ser, episod, vecc:

In queste parole sono racchiuse le cause delle critiche. I primi due probabilmente rivolti a "Star Wars", una serie di nove film, ognuno dei quali denominato con il termine di "episodio". Il terzo vocabolo, che molto probabilmente si riferisce a "Rocky", criticato invece per l'età.

2.2.3 Wordclouds videogiochi

Nella Figura 2.14 sono mostrati i *wordclouds* ottenuti per la categoria videogiochi. In particolare l’immagine di sinistra rappresenta le parole più usate nelle recensioni positive, mentre a destra sono rappresentate quelle negative.

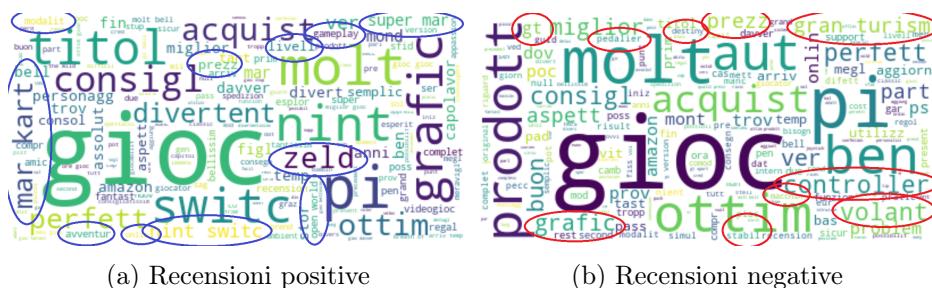


Figura 2.14: Wordclouds videogiochi

Recensioni positive: nello schema sono rappresentate con dimensione maggiore le parole **gioc**, **grafic**, riferite genericamente alla categoria presa in esame, di cui si apprezza la natura del gioco e la grafica. Per quanto riguarda invece il contenuto informativo, soffermiamoci sulla Figura 2.10; dalla quale apprendiamo quali sono i cinque prodotti migliori e cerchiamo di ricostruire cosa li caratterizza.

- super mar, mar kart, zeld:

I termini selezionati rappresentano infatti il podio dei migliori, rispettivamente "*Super Mario Odyssey*", "*The Legend of Zelda*" e *Mario Kart 8 Deluxe*".

- nint switch:

i videogiochi più apprezzati sono quelli per *nintendo switch*; non a caso infatti i tre migliori sono proprio accomunati da questa caratteristica.

- **avventur, online, modalit, open world, livell, gameplay:** tutti questi vocaboli sono caratteristiche positive dei giochi; probabilmente apprezzati quindi dalle avventure proposte dalla storia, dalla possibilità di giocare online in più modalità, compresa anche quella esplorativa, organizzate in livelli o dal *gameplay* intuitivo, etc.
- **prezz:** l'appartenenza dei giochi alla lista dei migliori può anche essere stata determinata dal prezzo, trovato forse dagli utenti accessibile o anche basso.

Recensioni negative: nello schema i termini generici più rilevanti per la categoria sono **gioc, prodott**. Concentriamoci ora sulla *top five* dei prodotti negativi (Figura 2.11), contrariamente ai casi precedenti abbiamo rappresentati tra i primi cinque prodotti ci sono sia videogiochi, sia accessori appartenenti alla categoria in questione.

- **gran turism, destiny, controller:** sono tutti termini appartenenti all'elenco dei primi cinque prodotti con più recensioni negative.
- **gt, pedalier, macchin, volant:** Sono parole riferite al gioco di corse automobilistiche "*Grand Turism*", apprezzato soprattutto quando si gioca con un volante o una pedaliera.
- **grafic, stabil:** i giochi sono probabilmente apprezzati per la grafica efficace e con il termine **stabile** ci si vuole forse riferire alla qualità da parte del gioco di mantenere la fluidità dell'immagine priva di scatti, o anche la capacità del gioco di non "*crashare*".
- **prezz:** anche in questo caso il prodotto è stato recensito rispetto al suo prezzo, contrariamente a prima però questa volta trovato troppo alto.

2.3 Correlazione tra prezzo e sentiment

Avendo trovato un riscontro positivo alle domande poste, siamo passati a una seconda analisi del *sentiment* sul *dataset* prodotti. In particolare ci siamo chiesti se il prezzo associato a ogni prodotto, fosse determinante per la polarità delle recensioni. Esiste infatti un'idea duale legato a un prezzo ridotto o eccessivo. Nel primo caso un prezzo ridotto può essere fonte di:

- **reazioni positive**, legate all'esigua quantità di denaro spesa per un prodotto funzionante.
- **reazioni negative**, legate alla qualità infima dei materiali.

Medesimo comportamento è legato ad alti costi. In questo caso abbiamo:

- **reazioni positive**, perché il prodotto oltre che essere funzionante è di ottima qualità.
- **reazioni negative**, dovute al costo eccessivo di alcuni prodotti, dove si paga non tanto la qualità, quanto piuttosto il marchio.

Attributo essenziale per la nostra analisi è il campo **avg-rating**, rappresentante la valutazione media di tutte le recensioni relative a ogni prodotto. Tuttavia contrariamente a quanto è stato compiuto per le prime analisi, ora non si è eseguita una campionatura sui prodotti migliori, ma si è stato calcolato il *sentiment* medio di ogni prodotto. Questo *sentiment* è stato poi raccolto in base alle categorie, sviluppando ancora una volta due gruppi di insiemi, per le recensioni positive e negative, questa volta però sull'intero *set* di prodotti. Ad operazione conclusa abbiamo raccolto tutti i possibili prezzi per ogni categoria e li abbiamo suddivisi in intervalli, così che per esempio, per la categoria libri ci fosse, tra i tanti, l'intervallo [1-10] che racchiudesse al suo interno tutti i prodotti con un prezzo compreso tra il valore 0 e il valore 10.

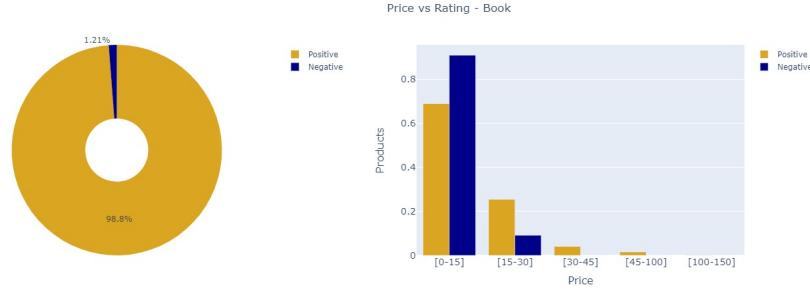
Una volta ottenuti i sotto-raggruppamenti dei prezzi per ogni categoria, siamo andati a visualizzare la tipologia delle recensioni per ognuno di essi. In particolare, tramite un istogramma, abbiamo rappresentato sull'asse delle ascisse tutti gli intervalli di prezzo per una certa categoria e sull'asse delle ordinate il conteggio dei prodotti in percentuale. Analizziamo i risultati per ogni categoria.

2.3.1 Categoria libri

Per la categoria "libri", abbiamo il 98.8% composto da recensioni positive, mentre le restanti sono negative, come mostrato in Figura 2.15a Studiamo come queste sono distribuite in base al prezzo, più precisamente all'intervallo di prezzi.

Dalla Figura 2.15b si evince che le recensioni positive sono maggiormente distribuite nei prodotti con un prezzo che oscilla tra 0 e 15 euro. Andando avanti la percentuale diminuisce, fino a non avere esemplari su un range dal valore compreso tra i 100 e 150 euro. Per quanto riguarda le recensioni negative, invece, si riscontra una presenza dominante tra quelle appartenenti al primo intervallo, cala drasticamente nel secondo fino a non avere prodotti per i tre *range* successivi. Una possibile chiave di lettura per questa rappresentazione, potrebbe essere che, per quanto riguarda i commenti positivi, con un prezzo basso, se l'utente è soddisfatto a maggior ragione lascerà una recensione positiva, in quanto non ha dovuto spendere molto per un prodotto funzionante. A prezzi elevati invece, è difficile che il prodotto non soddisfi le aspettative, in quanto un libro richiede pochi requisiti di qualità,

ne consegue che chi spende rimane soddisfatto, pertanto i commenti negativi si concentrano solo nel primo intervallo.



(a) Distribuzione delle recensioni (b) Correlazione prezzo e sentimento

Figura 2.15: Libri: studio sulle recensioni medie dei prodotti

2.3.2 Categoria dvd

Per la categoria "dvd", abbiamo una distribuzione probabilistica di recensioni medie pari a 97.7 per quelle positive e 2.33 per quelle negative (Figura 2.16a). Dalla Figura 2.16b si può notare che il 60% delle recensioni positive appartengono al primo intervallo, diminuiscono per il secondo e terzo *range*, mentre per il quarto si ha un impercettibile aumento delle stesse, probabilmente dovute al fattore qualità-prezzo. Come prima per l'ultimo intervallo abbiamo esclusivamente recensioni positive, seppur molto basse; in quanto difficilmente un utente spende grandi somme di denaro per comprare dvd. Dall'altra parte abbiamo invece quelle negative, il cui numero diminuisce all'aumentare del prezzo, ma che comunque si mantiene pressoché uguale per ogni intervallo.



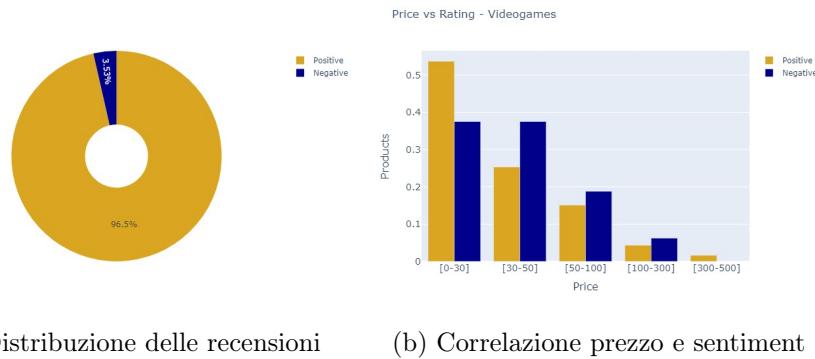
(a) Distribuzione delle recensioni (b) Correlazione prezzo e sentimento

Figura 2.16: Dvd: studio sulle recensioni medie dei prodotti

2.3.3 Categoria videogiochi

Per la categoria "videogiochi", abbiamo una distribuzione probabilistica formata da recensioni positive per il 96.5, mentre dal 3.55 per quelle negative (Figura 2.17a). Studiamo come queste sono distribuite in base al prezzo, più precisamente all'intervallo di prezzi

Dalla Figura 2.17b si può notare come anche in questo caso il comportamento per le recensioni positive sia a discesa, con più del 50% concentrate nel primo intervallo. Per quelle negative invece abbiamo una percentuale più o meno simile per i primi due intervalli, cala proseguendo, fino a scomparire del tutto nell'ultimo intervallo. Anche in questo caso la chiave di lettura può essere la stessa utilizzata per i libri.



(a) Distribuzione delle recensioni

(b) Correlazione prezzo e sentiment

Figura 2.17: Videogiochi: studio sulle recensioni medie dei prodotti

2.3.4 Categoria musica

Un dato importante è rappresentato dalla categoria "musica", che conta solo recensioni positive (Figura 2.18a). A riprova di questo, vi è anche l'istogramma in Figura 2.18b, dove non compare il colore blu delle recensioni negative.

2.4 Utilità delle recensioni per gli utenti

La nostra analisi del *sentiment* è proseguita nella direzioni delle recensioni; in particolare ci siamo chiesti se gli utenti abbiano trovato più utili le recensioni negative rispetto a quelle positive. Lo studio si è concentrato all'interno di ogni categoria, raggruppando per esse l'attributo **sentiment**, del *dataset* sulle recensioni in due diversi gruppi, uno per ogni polarità (esclusa quella neutra). Il risultato di questa operazione erano otto diversi gruppi, due per ogni categoria.

Definita questa divisione siamo andati a calcolare per ognuno degli otto insiemi il valore di utilità complessivo per ogni categoria. In particolare per



(a) Distribuzione delle recensioni (b) Correlazione prezzo e sentimento

Figura 2.18: Musica: studio sulle recensioni medie dei prodotti

il calcolo del valore abbiamo calcolato i valori su diverse metriche: utilità massima, minima e media.

Minimo: valore di utilità minimo raggiunto dalle recensioni.

Massimo: valore di utilità massimo raggiunto dalle recensioni.

Media: valore di utilità medio raggiunto dalle recensioni.

Di tutte le misure prese in considerazione, quella che più ci è sembrata utile per la rappresentazione è stata la media, in quanto esprime l'andamento generico di utilità. Calcolare, invece, l'utilità minima non portava alcun informazione alla nostra analisi; in quanto per tutte le categorie il valore minimo era sempre uguale a zero. Ragionamento più o meno simile è quello seguito per il valore massimo a livello informativo; non è infatti di alcun interesse conoscere il valore massimo, questo potrebbe essere l'unico a fronte di altri valori tutti minimi.

2.4.1 Analisi dei risultati tramite box plot

Avendo scelto la metrica per il calcolo dei risultati, siamo andati a visualizzarli tramite un *boxplot*. Il box plot è un metodo per rappresentare graficamente i gruppi di dati numerici attraverso i loro quartili. In particolare la rappresentazione prevede la visualizzazione di cinque valori: minimo, massimo, mediana, primo quantile e terzo quantile.

- **Minimo:** il punto dati più basso esclusi eventuali valori anomali.
- **Massimo:** il punto dati più alto esclusi eventuali valori anomali.
- **Mediana(Q2/50th percentile):** il valore medio del set di dati.

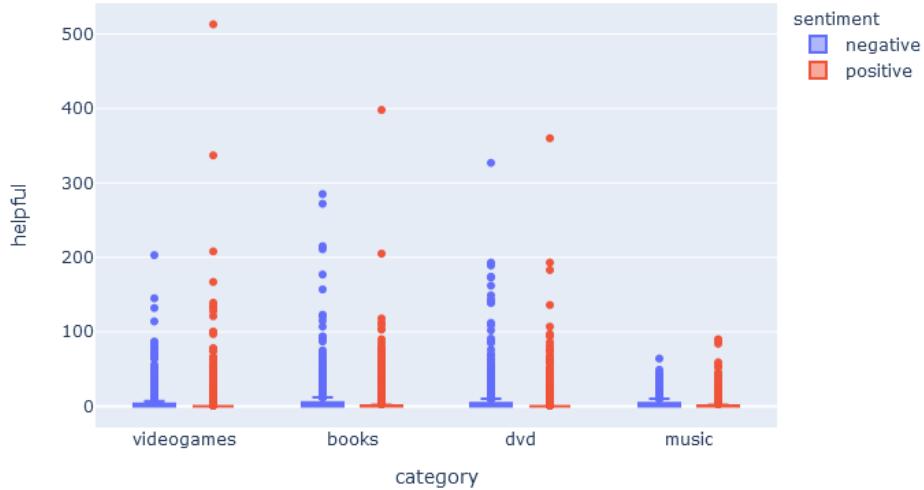


Figura 2.19: Utilità delle recensioni negative e positive per ogni categoria.

- **Primo Quartile(Q1/25th percentile):** è la mediana della metà inferiore del set di dati.
- **Terzo Quartile(Q1/75th percentile):** è la mediana della metà superiore del set di dati.

Importante specificare che in questo specifico diagramma esistono anche dei valori anomali, che possono essere tracciati come punti individuali.

Dalla Figura 2.19 si può osservare come per ogni categoria ci siano moltissimi valori sparsi; la "box" del *box plot* si nota a malapena. Tuttavia la considerazione più importante che si può sollevare è la superiorità di utilità delle recensioni negative su quelle positive; questo a significare che gli utenti trovano molto più utile le critiche negative mosse contro i prodotti rispetto che quelle positive.

Capitolo 3

Analisi tramite modelli supervisionati

Per terminare la nostra analisi, abbiamo scelto di allenare determinati modelli sui dati in *input* per fare delle previsioni sul *sentiment* e confrontare i valori ottenuti con quelli effettivi del *dataset*. Prima di entrare nei dettagli implementativi rimandiamo all'attenzione del lettore alcuni richiami teorici circa i modelli da noi utilizzati: *Logistic Regression*, *Support Vector Machine*, *Neural Network*. Per tutti questi saranno calcolate le misure associate a ogni modello.

3.1 Logistic Regression

La regressione logistica è un algoritmo di classificazione di *Machine Learning* che viene utilizzato per prevedere la probabilità di una variabile dipendente categoriale. Nella regressione logistica, la variabile dipendente è una variabile binaria che contiene dati codificati come 1 (sì, esito positivo) O 0 (no, esito negativo). La regressione logistica binaria richiede che: la variabile dipendente sia binaria; le variabili indipendenti siano indipendenti l'una dall'altra e linearmente correlate alle probabilità del registro; la regressione logistica richiede campioni di dimensioni piuttosto grandi.

- **Matrice di confusione:** restituisce una rappresentazione dell'accuratezza di una classificazione statistica. In particolare ogni colonna della matrice rappresenta i valori reali, mentre ogni riga rappresenta i valori predetti.

Tabella 3.1: Logistic Regression: Matrice di confusione

	Act Negative	Act Neuter	Act Positive
Pred Negative	850	90	809
Pred Neuter	210	122	1010
Pred Positive	226	148	20522

- **Precision:** capacità di un modello di classificazione di identificare solo i dati pertinenti). Essa è data dal rapporto tra veri positivi(TP) e la somma di veri positivi(TP) e falsi positivi(FP).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.1)$$

Tabella 3.2: Logistic Regression: Precision

Precision		
Negative	0.66096423	Questo dimostra che tra i dati predetti ben più del 60% risultano essere veri negativi.
Neuter	0.33888889	Anche in questo caso solamente il 33% dei dati predetti sono effettivamente neutri.
Positive	0.91858019	Abbiamo più del 90% di predetti veri positivi.

- **Accuratezza:** capacità di un modello di trovare tutti i casi veritieri, ovvero quante istanze vengano classificate correttamente. Questa è la somma dei veri positivi più i veri negativi diviso il totale dei dati.

$$\text{Accuratezza} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3.2)$$

Tabella 3.3: Accuratezza

Logistic Regression: Accuratezza		
LogisticR	0.896068703881269	Il modello ha classificato correttamente quasi il 90% delle istanze.

- **Recall/Sensitività/TPR:** capacità di un modello di trovare tutti i casi pertinenti all'interno di una serie di dati; quindi l'abilità nell'identificare i positivi. Matematicamente la sua definizione è data dal

rapporto tra veri positivi (TP) e la somma di veri positivi (TP) e falsi negativi (FN).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.3)$$

Tabella 3.4: Recall

Logistic Regression: Recall		
Negative	0.485992	Questo dato evidenzia che non si riesce a identificare correttamente nemmeno la metà dei negativi effettivi.
Neuter	0.09090909	In questo caso il valore è ancora peggio di quello dei negativi
Positive	0.98210184	Contrariamente agli altri, il modello identifica correttamente quasi la totalità dei positivi.

- **F1-mesaure:** media armonica di precisione e richiamo, prese entrambe le metriche. Un punteggio di F1 è considerato perfetto quando è pari a 1, mentre un fallimento se pari a 0.

Tabella 3.5: Logistic Regression: F1-mesaure

F1-mesaure		
Negative	0.5601318	Il valore è metà di 1; il modello sta classificando in maniera neutra i dati negativi.
Neuter	0.14336075	In questo caso il modello sta classificando in modo pessimo i dati neutri.
Positive	0.94927955	Il valore è vicino a 1, il modello nel complesso sta classificando bene i dati positivi.

3.2 Support Vector Machine

Nell'apprendimento automatico , le *Support Vector Machine*¹, sono modelli di apprendimento supervisionato con algoritmi di apprendimento associati che analizzano i dati utilizzati per l'analisi di classificazione e regressione. Dato un insieme di esempi di addestramento, ognuno segnato come

¹SVM:macchine di supporto vettoriale

appartenente all'una o all'altra di due categorie, un algoritmo di addestramento SVM costruisce un modello che assegna nuovi esempi a una categoria o all'altra, rendendolo un classificatore lineare binario non probabilistico. Un modello SVM è una rappresentazione degli esempi come punti nello spazio, mappati in modo tale che gli esempi delle categorie separate siano divisi da uno spazio vuoto il più ampio possibile. Nuovi esempi vengono quindi mappati nello stesso spazio e previsti per appartenere a una categoria in base al lato dello spazio su cui cadono.

- **Matrice di confusione**

Tabella 3.6: SVM: Matrice di confusione

	Act Negative	Act Neuter	Act Positive
Pred Negative	889	92	768
Pred Neuter	243	95	1004
Pred Positive	272	131	20493

- **Precision:**

Tabella 3.7: SVM: Precision

Precision		
Negative	0.63319088	Questo dimostra che tra i dati predetti ben più del 60% risultano essere veri negativi.
Neuter	0.29874214	Anche in questo caso solamente meno del 30% dei dati predetti sono effettivamente neutri.
Positive	0.9204132	Abbiamo più del 90% di predetti veri positivi.

- **Accuratezza:**

Tabella 3.8: SVM: Accuratezza

Accuratezza		
SVM	0.8953599866594405	Il modello ha classificato correttamente quasi il 90% delle istanze.

- **Recall/Sensitività/TP:**

Tabella 3.9: SVM: Recall

Recall		
Negative	0.50829045	Questo dato evidenzia che non si riesce a identificare poco più della metà dei negativi effettivi.
Neuter	0.07078987	In questo caso il valore è ancora peggio di quello dei negativi
Positive	0.98071401	Contrariamente agli altri, il modello identifica correttamente quasi la totalità dei positivi.

- **F1-mesaure:**

Tabella 3.10: SVM: F1-mesaure

F1-mesaure		
Negative	0.56390739	Il valore è metà di 1; il modello sta classificando in maniera neutra i dati negativi.
Neuter	0.11445783	In questo caso il modello sta classificando in modo pessimo i dati neutri.
Positive	0.94960728	Il valore è vicino a 1, il modello nel complesso sta classificando bene i dati positivi.

3.3 Reti neurali

Una rete neurale è un sistema computazionale che crea previsioni basate su dati esistenti. In altre parole esegue calcoli per rilevare le caratteristiche e decide se un input appartiene o meno ad una specifica classe.

Una rete neurale è composta da tre parti principali:

Livelli di input: accettano input in base a dati esistenti. Nel nostro caso 5000 neuroni per ogni *feature*.

Livelli nascosti: utilizzano la *backpropagation* per ottimizzare i pesi delle variabili di input al fine di migliorare la potenza predittiva del modello. Il nostro progetto ha previsto 3 livelli composti rispettivamente da 700, 400 e 100 neuroni.

Livelli di output: *output* di previsioni basate sui dati dell'input e dei livelli nascosti. Nel nostro caso 5 neuroni, uno per ogni punteggio delle stelle di *amazon* (da 1 a 5)

- **Matrice di confusione**

Tabella 3.11: NN: Matrice di confusione

	Act Negative	Act Neuter	Act Positive
Pred Negative	895	220	634
Pred Neuter	213	180	949
Pred Positive	286	301	20309

- **Precision:**

Tabella 3.12: Precision

NN: Precision		
Negative	0.6420373	Questo dimostra che tra i dati predetti ben più del 60% risultano essere veri negativi.
Neuter	0.25677603	Anche in questo caso solamente il 25% dei dati predetti sono effettivamente neutri.
Positive	0.92769048	Abbiamo più del 90% di predetti veri positivi.

- **Accuratezza:**

Tabella 3.13: NN: Accuratezza

Accuratezza		
NN	0.8914828865635552	Il modello ha classificato correttamente quasi il 90% delle istanze.

- **Recall/Sensitività/TP:**

Tabella 3.14: NN: Recall

Recall		
Negative	0.51172098	Questo dato evidenzia che non si riesce a identificare poco più della metà dei negativi effettivi.
Neuter	0.13412817	In questo caso il valore è ancora peggio di quello dei negativi
Positive	0.9719085	Contrariamente agli altri, il modello identifica correttamente quasi la totalità dei positivi.

- **F1-mesaure:**

Tabella 3.15: NN: F1-mesaure

F1-mesaure		
Negative	0.56951957	Il valore è metà di 1; il modello sta classificando in maniera neutra i dati negativi.
Neuter	0.17621145	In questo caso il modello sta classificando in modo pessimo i dati neutri.
Positive	0.94928485	Il valore è molto vicino a 1, il modello nel complesso sta classificando bene i dati positivi.

Capitolo 4

Network Analysis

?? In questo capitolo inizia la seconda parte di analisi, quella relativa alla rete. L'analisi delle reti sociali (o Network Analysis) rappresenta un insieme di strumenti finalizzati a descrivere le principali caratteristiche di una struttura di nodi e connessioni rifacendosi alla teoria dei grafi. Un grafo è definito come un insieme di coppie ordinate:

$$G = (V, A),$$

dove con V si indicano l'insieme di vertici e con A l'insieme di archi. Un grafo può essere orientato (directed) o non orientato (non-directed). Nel primo caso, i legami che connettono i nodi hanno una direzionalità (in uscita da un nodo e in entrata in un altro nodo), mentre nel secondo la relazione non ha un orientamento definito.

4.1 Analisi sulla rete: videogiochi

??? manca Figura????? Il nostro approccio alla rete ha riguardato il *dataset* prodotti; in particolare abbiamo scelto di rappresentare un grafo per la categoria "videogiochi", in cui i nodi fossero rappresentati dagli id dei prodotti e gli archi dall'attributo `also viewed`. A questo punto abbiamo proceduto con la rilevazione delle comunità tramite il metodo **Louvian**, che sarà approfondito nella sezione seguente.

4.1.1 Metodo Louvain

Il metodo *Louvain* per il rilevamento di comunità è un metodo per estrarre comunità da grandi reti. L'ispirazione per questo metodo di rilevamento della comunità è l'ottimizzazione della modularità man mano che l'algoritmo progredisce. La modularità è un valore di scala compreso tra -0,5 (clustering non modulare) e 1 (clustering completamente modulare) che misura la densità relativa dei bordi all'interno delle comunità rispetto ai bordi esterni alle comunità. L'ottimizzazione di questo valore si traduce teoricamente nel

migliore raggruppamento possibile dei nodi di una determinata rete, tuttavia passare attraverso tutte le possibili iterazioni dei nodi in gruppi non è pratico, quindi vengono utilizzati algoritmi euristici. Nel metodo *Louvain* di rilevamento della comunità, le prime piccole comunità vengono trovate ottimizzando localmente la modularità su tutti i nodi, quindi ogni piccola comunità viene raggruppata in un nodo e il primo passaggio viene ripetuto. Il metodo è simile al metodo precedente di Clauset, Newman e Moore [3] che collega le comunità la cui fusione produce il maggiore aumento della modularità.

4.2 Identificazione comunità

Lo scenario che si presentava era composto da quasi 100 comunità; numero ancora troppo elevato per le nostre analisi. La nostra scelta è stata quindi quella di raggruppare nuovamente le comunità per il campo console. L'operazione ha portato all'identificazione delle seguenti *community*

- Altro
- Nintento Switch
- PS4
- Xbox one
- Nintento 3DS
- PS3
- PSVita
- Nintento Classic Mini
- Nintento Wii
- Xbox 360

Capitolo 5

Conclusioni

??? Riassunto delle risposte ottenute per le domande ?????

Capitolo 6

TODO

- info sul dataset -> introduzione
- aggiunta tabelle valori
- analisi su modelli: confronto con modelli + ultima
- info rete
- conclusioni

Bibliografia