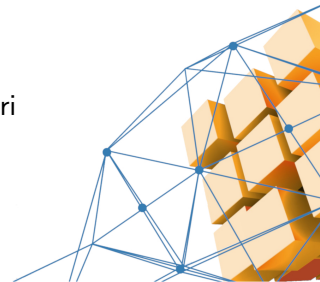
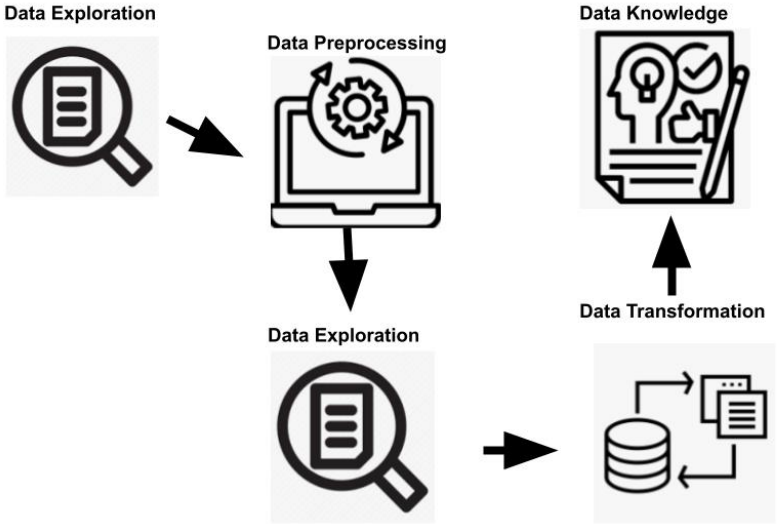
Politecnico di Torino
1859
Dipartimento di Scienze Matematiche "G. L. Lagrange"
DISMA
ECCELLENZA 2018•2022

Progetto di Analisi di Dati: Amazon rewiews

Gruppo7: Giorgia delle Grazie (300879), Elisa Salvadori (302630)



Pipeline dell'analisi



Descrizione Dataset

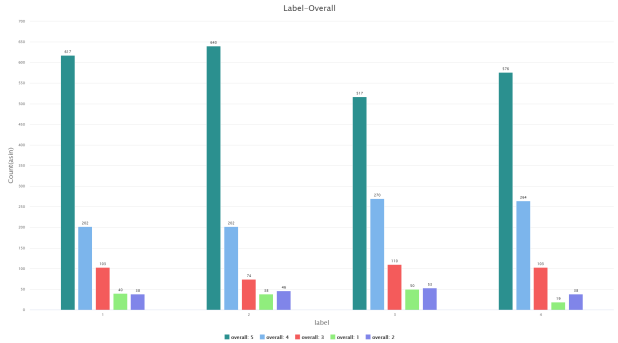
Il dataset è composto da 4000 recensioni Amazon, ogni riga è costituita dai seguenti attributi:

- **reviewerID**: codice identificativo dell'utente;
- **asin**: codice identificativo del prodotto recensito;
- **reviewerName**: nome dell'utente;
- **helpful**: valutazione sull'utilità della recensione (formato: [a, b]):
  - b rappresenta il numero di volte totale in cui la recensione è stata valutata;
  - a rappresenta il numero di volte totale in cui la recensione è stata valutata utile;
- **reviewText**: testo della recensione;
- **overall**: valutazione sul prodotto;
- **summary**: riassunto della recensione;
- **unixReviewTime, reviewTime**: data della recensione;
- **label**: categoria a cui il prodotto appartiene.

Data exploration

overall

Nel dataset sono presenti 4 categorie: Cd e Vinili (1), Casa e Cucina (2), Video Games (3), Kindle Store (4).



Overall = 5 è maggiormente presente .

Preprocessing

L'attributo **summary** e l'attributo **reviewerName** presentavano dei *missing value* perciò sono stati rimossi in quanto non fornivano ulteriori informazioni utili all'analisi prevista.

Inoltre è stato rimosso l'attributo **unixReviewTime**.

L'attributo reviewTime è stato scomposto in **day**, **mese**, **anno**.

È stato generato un nuovo attributo **season** diminuendo la granularità dell'attributo **mese**.

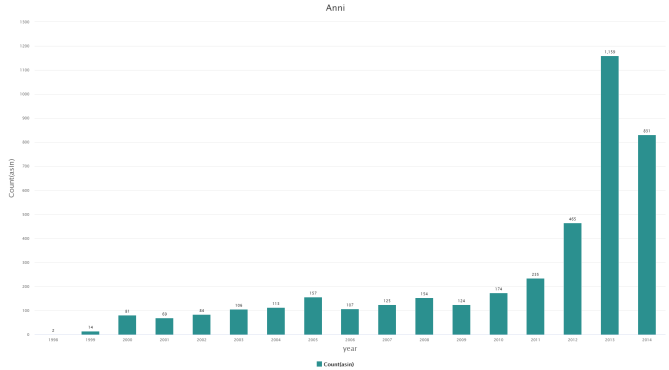
Preprocessing

reviewText

- Tokenize;
- Transform cases: lower case;
- Filter Stopwords;
- Stem Snowball;
- Filter tokens (by length): parole con lunghezza minima 4 e massima 10;
- Filter tokens (by content): sono tenute le parole che non contengono la parola Amazon.

Distribuzione temporale Dataset

Il dataset in esame è costituito da recensioni comprese tra il 1998 e il 2014.



Le recensioni prima del 2005 non sono significative per la nostra analisi, per questo si è deciso di lavorare con recensioni dal 2005 in poi (3531 record).

Processing

utility

Dato l'attributo **helpful** è stato eseguito il rapporto a/b per determinare l'utilità della recensione, in particolare è stato creato l'attributo **utility** con valore pari a **1** se $a/b \geq 0.6$ e pari a **0** altrimenti.

accuracy: 64.25% +/- 3.29% (micro average: 64.25%)

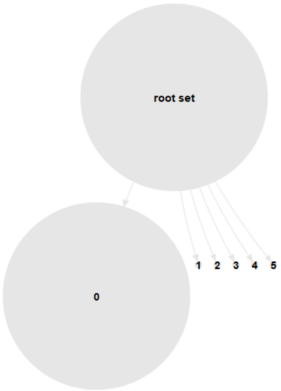
	true 0	true 1	class precision
pred. 0	248	346	41.75%
pred. 1	309	929	75.04%
class recall	44.52%	72.86%	

Con un Decision Tree (maximal depth = 30, minimal gain = 0.01) è stato predetto l'attributo **utility** quando b = 0.
Per l'analisi sono stati usati solo i record con **utility** = 1 (2487 record).

Cluster Analysis

Agglomerative Clustering-Flatten Clustering

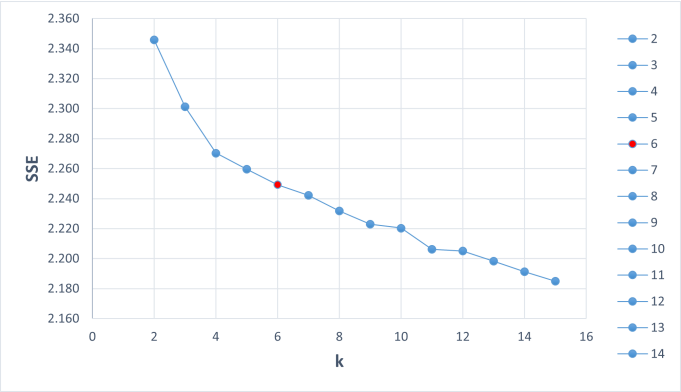
Measure types: Numerical measures (Cosine Similarity), numero di cluster pari a 6.
Si ottengono 6 cluster disomogenei e poco significativi.



Cluster Analysis

K-means

Si decide di utilizzare il k-means con k = 6, selezionato dopo un'analisi dell'SSE e dell'omogeneità dei cluster (Numerical measures-Cosine Similarity, max runs: 50, max optimization steps: 100).



Cluster-Kmeans

Number of Clusters: 6

Cluster 0 418

multiplay is on average **494.98%** larger, **gameplay** is on average **490.23%** larger, **weapon** is on average **482.04%** larger

Cluster 1 421

novel is on average **400.67%** larger, **heroin** is on average **381.38%** larger, **stori** is on average **345.05%** larger

Cluster 2 358

book is on average **420.79%** larger, **seri** is on average **377.18%** larger, **read** is on average **236.44%** larger

Cluster 3 420

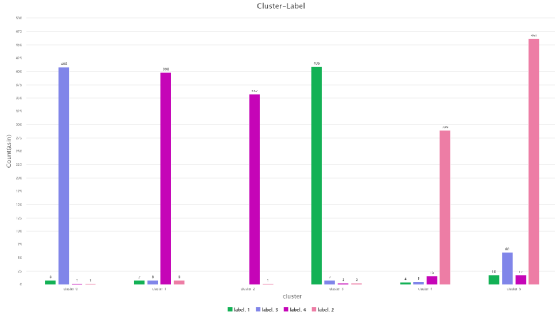
album is on average **490.34%** larger, **instrument** is on average **490.19%** larger, **vocal** is on average **481.69%** larger

Cluster 4 314

coffe is on average **657.39%** larger, **clean** is on average **624.91%** larger, **water** is on average **527.91%** larger

Cluster 5 556

vacuum is on average **333.12%** larger, **sleep** is on average **263.15%** larger, **kitchen** is on average **246.26%** larger



Cluster0

Video games

Size	Support	Item 1
1	0.952	game
1	0.670	play
1	0.388	time
1	0.364	good
1	0.352	great
1	0.344	make
1	0.301	graphic
1	0.273	look
1	0.242	thing
1	0.227	love
1	0.220	want
1	0.215	find
1	0.213	control
1	0.213	take
1	0.203	charact
1	0.201	enjoy
1	0.194	come
1	0.189	gameplay

Disomogeneo, formato da molte parole, ma in gran parte poco frequenti.

Cluster1-Cluster2

Kindle Store

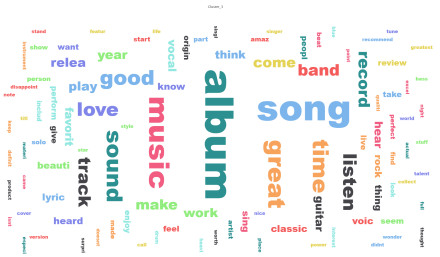
Size	Support	Item 1
1	0.651	stori
1	0.525	read
1	0.489	book
1	0.463	love
1	0.418	charact
1	0.264	good
1	0.245	make
1	0.235	author
1	0.226	enjoy
1	0.226	time
1	0.214	great
1	0.197	want
1	0.173	look
1	0.171	come
1	0.171	life
1	0.166	feel
1	0.164	find
1	0.162	know

Size	Support	Item 1
1	0.944	book
1	0.578	read
1	0.349	love
1	0.335	seri
1	0.223	enjoy
1	0.223	great
1	0.212	good
1	0.196	stori
1	0.193	charact
1	0.190	recommend
1	0.173	author
1	0.170	wait
1	0.168	want
1	0.165	start
1	0.154	know
1	0.154	make
1	0.137	look
1	0.128	time

Entrambi i cluster risultano essere abbastanza omogenei.

Cluster3

Cd e Vinili



Size	Support	Item 1
1	0.588	album
1	0.545	song
1	0.438	music
1	0.367	great
1	0.338	time
1	0.331	sound
1	0.317	good
1	0.298	listen
1	0.286	love
1	0.271	track
1	0.245	band
1	0.236	record
1	0.221	make
1	0.219	come
1	0.205	work
1	0.202	play
1	0.202	releas
1	0.200	year

Molto numeroso e abbastanza omogeneo.

Cluster4-Cluster5

Casa e Cucina

Size	Support	Item 1
1	0.395	easi
1	0.379	clean
1	0.306	make
1	0.252	great
1	0.236	time
1	0.220	work
1	0.197	love
1	0.182	water
1	0.172	good
1	0.169	keep
1	0.166	thing
1	0.159	coffe
1	0.156	want
1	0.143	look
1	0.140	cook
1	0.140	take
1	0.137	nice
1	0.127	recommend

Size	Support	Item 1
1	0.272	great
1	0.250	work
1	0.225	good
1	0.191	make
1	0.189	product
1	0.187	love
1	0.185	time
1	0.182	nice
1	0.176	look
1	0.147	perfect
1	0.144	price
1	0.140	purchas
1	0.122	bought
1	0.119	recommend
1	0.115	qualiti
1	0.115	want
1	0.113	size
1	0.106	thing

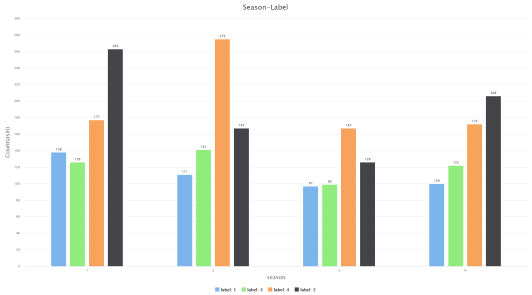
I cluster risultano abbastanza omogenei.

Applicazione

Azienda

Come notato in precendenza l'overall più presente è 5 e le parole più frequenti all'interno dei cluster sono tutte positive, quindi i prodotti sono stati valutati in maniera positiva.
Si noti che nella stagione 1 (inverno) e nella 4 (autunno) i prodotti più recensiti appartengono alla categoria Casa e Cucina, mentre nella stagione 2 (primavera) e nella 3 (estate) i prodotti più recensiti appartengono alla categoria Kindle Store.

Sfruttando quindi l'analisi svolta un'azienda potrebbe prendere in considerazione tali dati per incrementare la produzione di una categoria di prodotto in una specifica stagione.



Applicazione

Amazon

Inoltre si noti che le persone che hanno recensito più di un prodotto, tendono ad effettuare recensioni e quindi acquisti relativi alla stessa categoria di prodotto.

Quindi Amazon, al momento della pubblicazione di una recensione per un prodotto, potrebbe suggerire agli acquirenti i 5 prodotti più recensiti della categoria a cui appartiene il prodotto.