Laurea Magistrale in
DATASCIENCE

AMAZON FINE FOOD

A TEXT MINING AND SEARCH PROJECT

Text Mining and Search

A.A. 2020 - 2021

Carrara Gabriele - 814720

Filosa Alberto - 815589

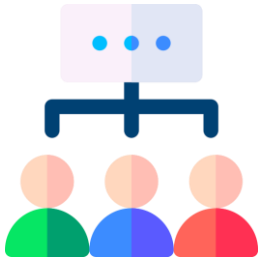
Tufano Simone - 816984

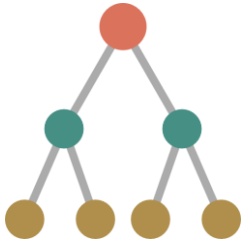
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

PIPELINE

Acquisizione Dati

Classification





Clustering

Analisi Esplorativa

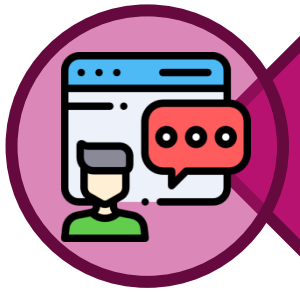
Preparazione Dati

Topic Modeling

DOMANDE DI RICERCA

Data una recensione è possibile prevedere la valutazione data dall'utente?

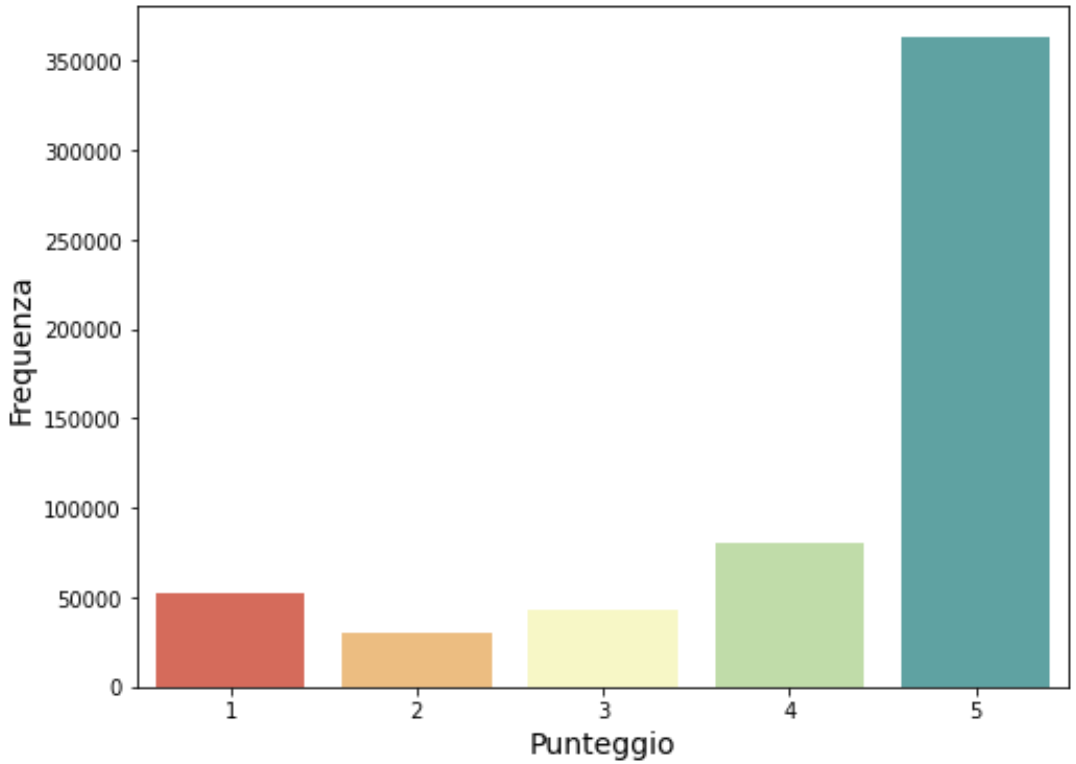
Recensioni simili portano a valutazioni simili?

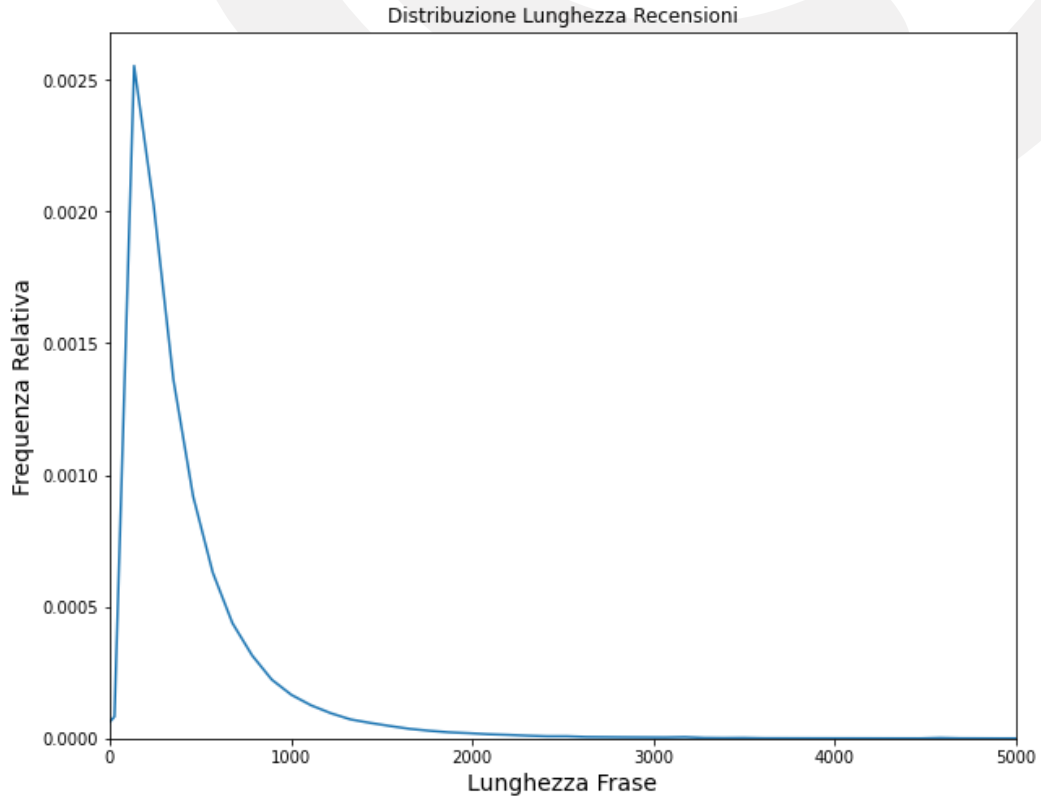


Quali sono gli argomenti più discussi nelle recensioni degli utenti?

ANALISI ESPLORATIVA

Si osserva uno sbilanciamento del punteggio delle recensioni. Le recensioni con la frequenza maggiore sono relative al punteggio massimo. È stata osservata la distribuzione delle lunghezze delle frasi: il 90% della distribuzione della lunghezza delle recensioni ha meno di 1'000 caratteri.





PREPARAZIONE DEI DATI

PRE - PROCESSING

- Bilanciamento dei Dati
- *Lower Case*
- Rimozione di spazi in eccesso
- *Tokenization*
- Rimozione *Stopword*
- *Stemmatization*
- *Lemmatization*

TEXT REPRESENTATION

- *Bag of Words*
- *Binary Representation*
- *Tf-Idf*

DIMENSIONALITY REDUCTION (SVD)

Rappresentazione	Varianza Spiegata
Stemmed Bynary	0.538
Stemmed BoW	0.641
Stemmed Tf-Idf	0.384
Lemmed Tf-Idf	0.356
Lemmed Binary	0.505
Lemmed BoW	0.613

CLASSIFICATION

Il Dataset delle recensioni è stato diviso in Train (67 %) e Test (33 %). Sono stati costruiti i seguenti modelli di classificazione multinomiale:

- Singular Vector Machine;
- k Nearest Neighbor (*k = 5*);
- Random Forest.

I modelli hanno prodotto le seguenti performance:

Modello	Accuracy
SVM	0.47
K-NN	0.30
RF	0.37

Accuracy dei modelli

	Precision	Recall	F1-Score
1	0.55	0.60	0.58
2	0.38	0.35	0.36
3	0.37	0.36	0.36
4	0.43	0.40	0.41
5	0.58	0.64	0.61

Altre Metriche di Classificazione (SVM Lemmed Tf-Idf)

CLUSTERING

Le tecniche di clustering prese in considerazione sono:

- *k*-Means;
- Hierarchical Clustering: *Agglomerative Single linkage*;
- Hierarchical Clustering: *Agglomerative Average linkage*.

In entrambe le partizioni è stato preso in considerazione un numero di gruppi pari a 5, così da intendere il clustering come classificazione non supervisionata, valutata tramite la metrica *Normalized Mutual Information*.

	1	2	3	4	5
1	7670	415	451	749	715
2	7218	583	663	632	904
3	7139	580	664	702	915
4	7071	618	650	677	984
5	7182	718	529	790	781

Matrice di Confusione (k-Means Lemmed Tf-Idf)

NMIS
0.002013

Il metodo delle *k*-medie non è un buon classificatore, in quanto l'appartenenza in un gruppo non implica un punteggio di recensione simile.

Inoltre, i metodi di tipo gerarchico (sia *Single* che *Average*) non riescono a dividere le unità in gruppi, in quanto generavano 4 gruppi composti da una sola unità ed il restante comprendente tutte le altre.

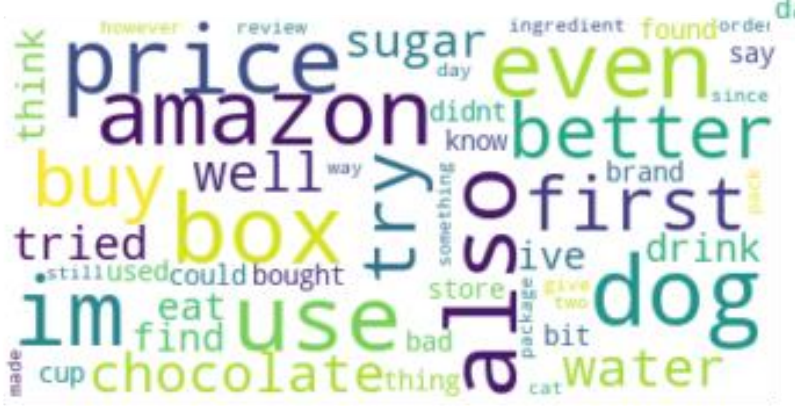
CARATTERIZZAZIONE GRUPPI

A conferma di quanto detto in precedenza, le Wordcloud dei gruppi appaiono molto confuse.

Wordcloud Cluster 1

Wordcloud Cluster 3

Wordcloud Cluster 2



Wordcloud Cluster 4

Wordcloud Cluster 5

TOPIC MODELING - LSA

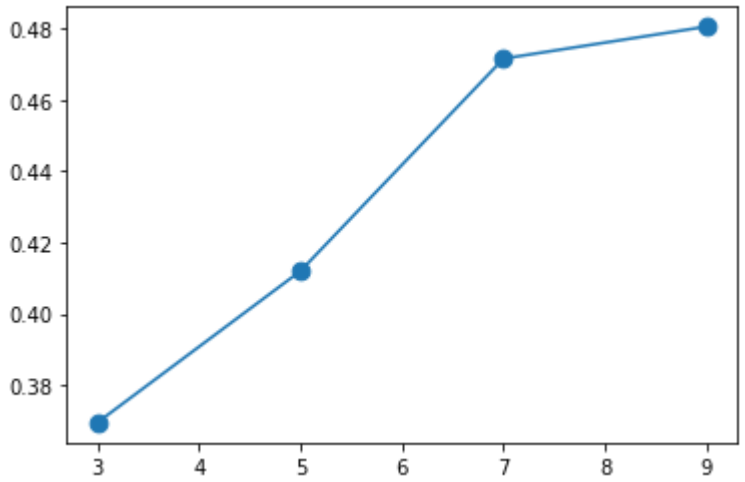
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Like	Tea	Tea	Dog	Product	Water	Chocolate	Cat	Product	Great
Taste	Coffe	Green	Food	Amazon	Cat	Bar	Food	Chocolate	Love
Coffee	Cup	Dog	Coffee	Prize	Food	Box	Eat	Taste	Flavor
Product	Flavor	Bag	Treat	Box	Drink	Cooky	Bar	Bar	Food
Tea	Green	Food	Tea	Order	Ingredient	Bag	Chocolate	Food	Price
Flavor	Drink	Leaf	Cat	Shipping	Use	Dark	Box	Cat	Good
Good	Strong	Box	Bag	Item	Sauce	Treat	Brand	Tea	Chocolate
One	Taste	Cat	Love	Ordered	Bottle	Candy	Dry	Coffee	Bar

Identificazione per Topic (LSA)

La procedura ha identificato 10 topic, i quali non sono tutti interpretabili. Ad esempio, nei topic 5 e 7 la lista di parole appare chiaramente riconducibile ad argomenti comuni (*Amazon, Order, Shipping e Chocolate, Candy, Cookie*), mentre in altri, come il topic 8 la tematica non è ben definita.

TOPIC MODELING - LDA

Nel caso della LDA, il numero di topic non è stato fissato a priori ma valutato attraverso la metrica Coherence, con numero di topic pari a 3, 5, 7 e 9.
Il valore maggiore di *Coherence* si è ottenuto per un numero di topic pari a 9 (0.48).



Variazione Coherence

A differenza di quanto accaduto in precedenza, i topic appaiono molto più interpretabili e coerenti.

Topic	Punteggio
Topic 1	$0.034\times$ *taste* $+ 0.030 \times$ *like* $+ 0.023 \times$ *flavor* $+ 0.015 \times$ *drink* $+ 0.014 \times$ *water* $+ 0.012 \times$ *good* $+ 0.011 \times$ *br* $+ 0.010 \times$ *really* $+ 0.009 \times$ *dont* $+ 0.009 \times$ *sugar*
Topic 2	$0.025 \times$ *oil* $+ 0.025 \times$ *sauce* $+ 0.020 \times$ *salt* $+ 0.019 \times$ *rice* $+ 0.018 \times$ *use* $+ 0.014 \times$ *make* $+ 0.014 \times$ *add* $+ 0.013 \times$ *br* $+ 0.012 \times$ *pasta* $+ 0.012 \times$ *recipe*
Topic 3	$0.096 \times$ *coffee* $+ 0.026 \times$ *cup* $+ 0.021 \times$ *flavor* $+ 0.018 \times$ *like* $+ 0.016 \times$ *taste* $+ 0.013 \times$ *bean* $+ 0.012 \times$ *good* $+ 0.012 \times$ *vanilla* $+ 0.009 \times$ *roast* $+ 0.009 \times$ *strong*
Topic 4	$0.022 \times$ *product* $+ 0.016 \times$ *price* $+ 0.016 \times$ *amazon* $+ 0.015 \times$ *great* $+ 0.013 \times$ *store* $+ 0.012 \times$ *one* $+ 0.011 \times$ *good* $+ 0.011 \times$ *buy* $+ 0.011 \times$ *box* $+ 0.010 \times$ *find*
Topic 5	$0.066 \times$ *br* $+ 0.019 \times$ *sugar* $+ 0.019 \times$ *product* $+ 0.016 \times$ *ingredient* $+ 0.013 \times$ *fat* $+ 0.013 \times$ *protein* $+ 0.012 \times$ *organic* $+ 0.012 \times$ *calorie* $+ 0.010 \times$ *high* $+ 0.009 \times$ *low*
Topic 6	$0.022 \times$ *like* $+ 0.019 \times$ *taste* $+ 0.018 \times$ *chocolate* $+ 0.017 \times$ *good* $+ 0.012 \times$ *flavor* $+ 0.012 \times$ *love* $+ 0.011 \times$ *great* $+ 0.011 \times$ *bar* $+ 0.011 \times$ *eat* $+ 0.011 \times$ *snack*
Topic 7	$0.106 \times$ *tea* $+ 0.021\times$ *green* $+ 0.018 \times$ *flavor* $+ 0.016 \times$ *hot* $+ 0.013 \times$ *br* $+ 0.012 \times$ *taste* $+ 0.011 \times$ *like* $+ 0.010 \times$ *ginger* $+ 0.010 \times$ *good* $+ 0.008 \times$ *cup*
Topic 8	$0.023 \times$ *br* $+ 0.014 \times$ *use* $+ 0.014 \times$ *work* $+ 0.011 \times$ *time* $+ 0.010 \times$ *day* $+ 0.009 \times$ *get* $+ 0.009 \times$ *one* $+ 0.008 \times$ *make* $+ 0.008 \times$ *well* $+ 0.007 \times$ *keep*
Topic 9	$0.037 \times$ *food* $+ 0.037 \times$ *dog* $+ 0.020 \times$ *treat* $+ 0.018 \times$ *cat* $+ 0.015 \times$ *love* $+ 0.012 \times$ *like* $+ 0.010 \times$ *one* $+ 0.009 \times$ *eat* $+ 0.007 \times$ *baby* $+ 0.007 \times$ *hair*

Identificazione per Topic (LDA)

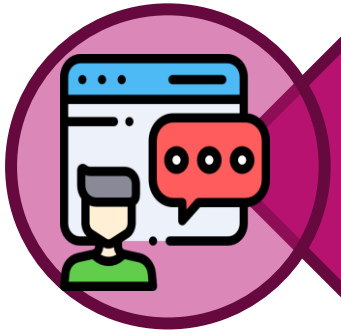
RISPOSTA ALLE DOMANDE DI RICERCA

Data una recensione è possibile prevedere la valutazione data dall'utente?

Esiste sicuramente una relazione tra il testo della recensione e la valutazione in termini numerici, però non è di tipo deterministico ma soggettivo. Quindi un modello di Machine Learning può essere utile solo in parte per questo genere di previsione.

Recensioni simili portano a valutazioni simili?

Raggruppare recensioni simili in termini di contenuto non porta vantaggi dal punto di vista previsionale.



Quali sono gli argomenti più discussi nelle recensioni degli utenti?

I topic sono direttamente legati alle categorie di prodotti venduti, ad eccezione di una tematica legata alle spedizioni effettuate da Amazon.