PROGETTO DI TEXT MINING: TEXT CLASSIFICATION, CLUSTERING & TOPIC MODELING

AMAZON FINE FOOD

Gabriele Carrara | Alberto Filosa | Simone Tufano

Gabriele Carrara
Matricola: 814720
e-mail:g.carrara12@campus.unimib.it

Alberto Filosa
Matricola: 815589
e-mail: a.filosa1@campus.unimib.it

Simone Tufano
Matricola: 816984
e-mail: s.tufano1@campus.unimib.it

In questo studio sono state analizzate le recensioni degli utenti Amazon attraverso tecniche di Text Mining. Il progetto si può idealmente dividere in tre parti: una prima in cui si è proceduto con la classificazione, per prevedere la valutazione di un prodotto analizzando il contenuto della recensione. Per questo scopo sono state utilizzate diverse rappresentazioni testuali per valutare eventuali differenze nelle performance. Una seconda parte affronta il problema in modo non supervisionato, attraverso diversi tipi di classificazione per osservare la presenza di legami tra la similarità delle recensioni e la valutazione finale. Infine si è proceduto con l'estrazione dei topic più rilevanti.

KEYWORDS:
Text Mining - Clustering - Classificazione - Topic Modeling - Amazon - Fine Food

1 | INTRODUZIONE

Il dataset iniziale è composto da 568'454 recensioni su 74'258 prodotti nella piattaforma di vendita Amazon nell'ambito del *fine food* nel periodo Ottobre 1999 - Ottobre 2012. Per *fine food* si intendono specialità gastronomiche di qualità, anche se da una analisi iniziale non si riesce a darne una definizione precisa, vista l'ampia gamma di prodotti presenti. Non sono state incluse le recensioni considerate duplicate secondo il seguente criterio: stesso id utente, stesso nome profilo, stesso periodo di tempo e testo. Questa operazione ha portato una riduzione del numero di osservazioni a 393'933.

Per l'analisi sono state considerate solamente le variabili *Rating* e *Text*. La variabile *Rating*, il target della classificazione, è una variabile categoriale che valuta il prodotto con un punteggio assegnato da un utente da 1 a 5. Di seguito si riporta la distribuzione della variabile in questione:

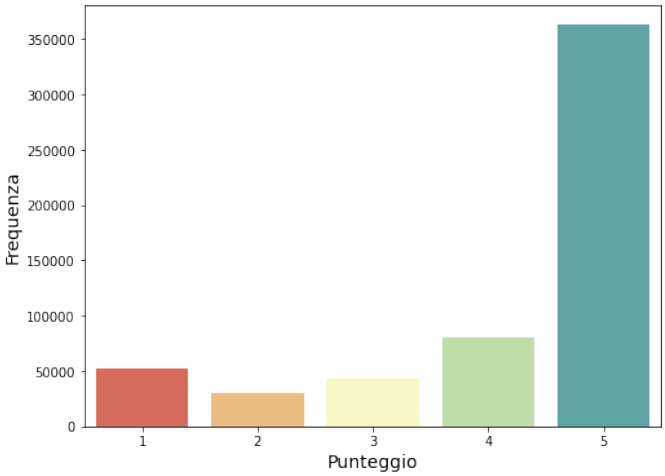

FIGURA 1 Distribuzione delle Valutazioni

La variabile *Text* è una variabile che contiene la recensione testuale del prodotto acquistato, utilizzata come variabile esplicativa principale per la classificazione. Di seguito si riporta la distribuzione del numero di caratteri presenti nelle recensioni:

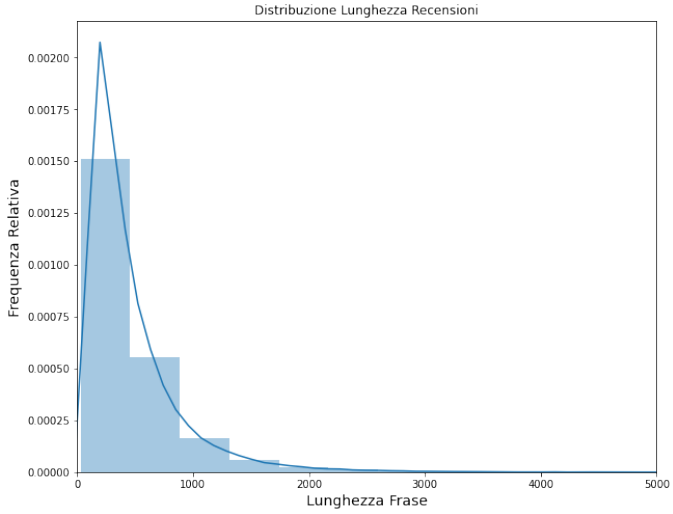

FIGURA 2 Distribuzione della Lunghezza delle Recensioni

Come emerge dalla Figura 2, il 90% della distribuzione della lunghezza delle recensioni ha meno di 1'000 caratteri.

Poiché il numero di osservazioni è particolarmente elevato ed il target è sbilanciato (circa il 60 % delle osservazioni ha una valutazione pari a 5), si è deciso di svolgere una strategia di *undersampling* selezionando 10'000 unità per ogni modalità della variabile risposta, in modo da ottenere un dataset definitivo con 50'000 osservazioni equi-distribuite.

1.1 | Domande di Ricerca

Sono state identificate le seguenti domande di ricerca:

1. *Data una recensione è possibile prevedere la valutazione data dall'utente?*
2. *Recensioni simili portano a valutazioni simili?*
3. *Quali sono gli argomenti più discussi nelle recensioni degli utenti?*

2 | PREPARAZIONE DEI DATI

2.1 | Pre Processing

Prima di procedere allo svolgimento dei task, sono state necessarie operazioni di *pre processing* al fine di uniformare il testo. Le tecniche utilizzate sono le seguenti:

- *Lower Case*, trasformazione di tutte i caratteri in minuscolo;
- Rimozione di spazi in eccesso, numeri e punteggiatura attraverso espressioni regolari;
- *Tokenization*, ciascuna frase è stata tokenizzata in unigrammi;
- Rimozione *Stopword* presenti nel dizionario del pacchetto `nltk` in lingua inglese;
- *Stemmatization* e *Lemmatization*, ovvero riduzione al lemma ed alla radice della parola in questione, sono stati svolti in parallelo. Per il resto dell'analisi, tutti i passaggi sono stati eseguiti su entrambi gli approcci.

2.2 | Text Representation

Le recensioni sono state rappresentate in forma strutturata secondo tre metodi:

- *Bag of Words*, ciascun documento è identificato secondo un vettore in cui è presente il numero di occorrenze di ciascuna parola;
- *Binary Representation*, a ciascun documento è associato un vettore binario in cui ogni elemento corrisponde alla presenza o meno della parola nel testo;
- *Tf-Idf*, il numero di occorrenze di ciascuna parola (*Term Frequency*) è pesato rispetto all'inverso della presenza della parola nel *corpus*(*Inverse Document Frequency*) in modo da ottenere una maggior caratterizzazione del documento.

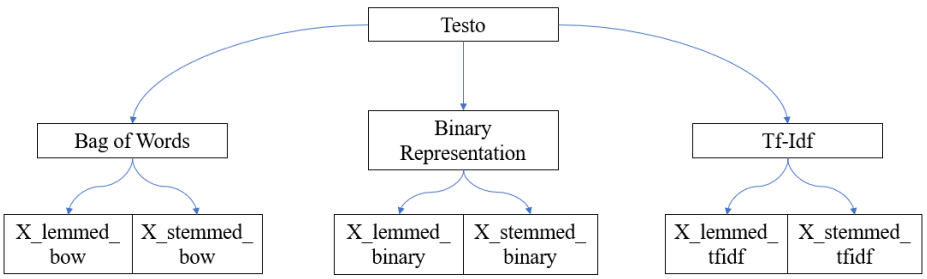

FIGURA 3 Divisione dei Dataset

2.3 | Dimensionality Reduction

Queste rappresentazioni hanno portato a 10000 variabili, scelte come parole più frequenti. Nonostante ciò, il peso computazionale è rimasto elevato, per cui si è optato per un'ulteriore riduzione della dimensionalità dei dataset tramite la tecnica della *Singular Value Decomposition* (**SVD**). Il numero di componenti mantenuto è di 300. Questo ha portato a dei nuovi dataset più densi la cui percentuale di varianza spiegata varia nel range 35% - 64%. Per quanto non sia un valore molto elevato e questa tecnica comporti una perdita di interpretabilità dei dati, la riduzione della dimensionalità porta ad un notevole vantaggio computazionale. Di seguito si riportano i valori delle varianze spiegate di ogni rappresentazione

Rappresentazione	Varianza Spiegata
X_stemmed_binary	0.538
X_stemmed_bow	0.641
X_stemmed_tfidf	0.384
X_lemmed_tfidf	0.356
X_lemmed_binary	0.505
X_lemmed_bow	0.613

TABELLA 1 Varianze Spiegate per ogni dataset

Si nota che la SVD fatica a spiegare la varianza laddove utilizzato il *Tf-Idf*, mentre rende molto meglio con le altre rappresentazioni.

3 | CLASSIFICATION

Le recensioni presentano una valutazione da 1 a 5 stelle, di conseguenza si tratta di una classificazione multinomiale. I modelli presi in considerazione sono:

- *Singular Vector Machine* (SVM);
- *k Nearest Neighbor* (kNN);
- *Random Forest* (RF).

Tutte le combinazioni dei dataset sono state sottoposte al processo di *Holdout*, ottenendo le partizioni di *Train* e *Test* (rispettivamente del 67% e 33% dei dati iniziali).

3.1 | Singular Vector Machine

Il primo modello utilizzato per risolvere il task è SVM, della libreria `sklearn`. Il kernel utilizzato è di tipo Radial Basis Function (`rbf`). Il tipo di rappresentazione dei dati che raggiunge le migliori performance, in termini di *Accuracy* sul test, è il *Tf-Idf* con le parole lemmatizzate:

	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

Accuracy
0.47

TABELLA 2 Statistiche Descrittive Modello SVM

3.2 | k Nearest Neighbor

Il secondo modello utilizzato è k-NN, presente nella libreria `sklearn`, con numero di vicini pari a 5 e metrica di distanza Euclidea. Il tipo di rappresentazione dei dati che raggiunge le migliori performance, in termini di *Accuracy* sul test, è la rappresentazione binaria con le parole lemmatizzate:

	Precision	Recall	F1-Score
1	0.31	0.56	0.40
2	0.25	0.23	0.24
3	0.24	0.17	0.20
4	0.30	0.21	0.24
5	0.38	0.31	0.35

Accuracy
0.30

TABELLA 3 Statistiche Descrittive Modello k-NN

3.3 | Random Forest

L'ultimo modello utilizzato è il Random Forest, presente nella libreria `sklearn`, con numero di alberi pari a 100 ed il criterio di misura di divisione *Gini*. Il tipo di rappresentazione dei dati che raggiunge le migliori performance, in termini di *Accuracy* sul test, è il *Tf-Idf* con le parole sia lemmatizzate che stemmatizzate:

	Precision	Recall	F1-Score
1	0.44	0.54	0.48
2	0.29	0.28	0.29
3	0.30	0.25	0.27
4	0.33	0.28	0.30
5	0.46	0.50	0.48

Accuracy
0.37

TABELLA 4 Statistiche Descrittive Modello Random Forest

3.4 | Conclusioni

Si nota che nessun modello riesce a prevedere particolarmente bene la variabile risposta. In generale, il modello più adeguato per risolvere il problema tra quelli proposti è la Singular Vector Machine, con una *Accuracy* pari al 47%. È ragionevole pensare che prevedere un punteggio multinomiale sia più complicato in quanto lo score viene assegnato sulla base di una scala soggettiva, per cui a due recensioni simili in termini di testo non è certo che corrisponda una pari valutazione in termini numerici. Nel caso in cui la variabile target fosse stata binomiale, la classificazione sarebbe probabilmente stata più precisa.

Infine, è possibile notare la modellazione del testo che mediamente ottiene il valore più elevato in termini di *Accuracy* è la rappresentazione *Tf-Idf* lemmatizzata, che pare quindi più adatta a questo tipo di problema.

4 | CLUSTERING

Le tecniche di clustering prese in considerazione sono:

- *k-Means*;
- *Hierarchical Clustering*.

In entrambe le partizioni è stato preso in considerazione un numero di gruppi pari a 5, in modo tale da verificare se l'appartenenza ad un cluster potesse corrispondere ad un punteggio attraverso la metrica della *Normalized Mutual Information*. Essa può essere interpretata come la correlazione tra l'appartenenza ad un gruppo e la variabile risposta.

4.1 | k-Means

Il primo approccio è stato di tipo *Prototype-Based* tramite l'utilizzo dell'algoritmo *k-means*. La rappresentazione dei testi che ha portato ad un valore maggiore di *Normalized Mutual Information* è stata quella lemmatizzata e pesata tramite *Tf-Idf*. Di seguito si riportano i risultati.

	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

Mutual Information Score
0.002013

TABELLA 5 Matrice di Confusione Cluster-Score

Si nota che non c'è alcuna correlazione tra appartenenza ad un gruppo e valutazione assegnata al prodotto.

4.1.1 | Caratterizzazione

Di seguito si riportano le *Wordcloud* dei 5 gruppi ottenuti per la rappresentazione *Tf-Idf* lemmatizzata. Poiché da una prima analisi è emerso che le parole più frequenti entro ciascun cluster fossero le stesse e di conseguenza poco significative, si è optato per escludere i primi 20 termini per frequenza dalla rappresentazione, mantenendo i 50 successivi per avere un maggior numero di parole che caratterizzano un cluster da un altro.

FIGURA 4 Wordcloud per ciascun Cluster

In generale, è difficile dare un'interpretazione alle Wordcloud.

4.2 | Hierarchical Clustering: Agglomerative Single linkage

L'ultima tecnica utilizza è il clustering gerarchico. Dapprima si è operato con un raggruppamento a legame singolo di tipo agglomerativo con metrica di distanza coseno. Di seguito si riportano i risultati.

	1	2	3	4	5
1	9999	0	0	0	1
2	10000	0	0	0	0
3	10000	0	0	0	0
4	9999	0	1	0	0
5	9998	1	0	1	0

Mutual Information Score
0.000160

TABELLA 6 Matrice di Confusione Gerarchico Single Linkage

Poichè questo approccio non ha portato a risultati utili, in quanto quasi tutte le unità appartenevano al medesimo cluster e i rimanenti gruppi risultavano popolati da sole una o due unità, si è proceduto con un nuovo tipo di clustering a legame medio con metrica di distanza euclidea.

4.3 | Hierarchical Clustering: Agglomerative Average linkage

Per problemi computazionali, è stato necessario dimezzare la dimensionalità, portando il dataset a 25'000 osservazioni. Di seguito si riportano i risultati.

	1	**2**	**3**	**4**	**5**
1	4940	0	0	1	1
2	5047	1	1	0	0
3	5067	0	0	0	0
4	4935	0	0	0	0
5	5007	0	0	0	0

Mutual Information Score
0.000320

TABELLA 7 Matrice di Confusione Gerarchico Average Linkage

Anche questo tipo di clustering non ha portato a risultati notevoli.

4.3.1 Caratterizzazione

Poiché in entrambe le analisi sono risultati dei gruppi popolati da una sola unità, si è cercato di comprenderne il motivo osservando le recensioni interessate. È emerso che a fronte di un Wordloud come in Figura 5, le frasi sono le seguenti.

Nome	Titolo	Recensione
D. Hodgson	A chocolate lover	I will never have a party without it, I made 16 chocolate Martini's and also served it hot with Bailey's, and mixed it with cream and made fondue for dipping fruit- directions are on the hang tag.My guests are still thanking me for the great time!
Shennah R. Teuscher	Jesus won	My daughter was in Peru on a mission trip this last summer and loved their Inca Kola. So for Christmas I decided to order her some. We all had some and liked it as well. It has a bubblegum/cream soda flavor. Different! But the price and shipping terribly high.
Curtis Gomez	Excellent taste	I really love this product. I use it in a Filipino style scramble with tofu, onion, tomato and garlic. Tastes exactly like the eggs my wife would make before we became vegan. I've used it in a vegan coconut custard cream (tofu) pie and stuffed red potatoes that were made like deviled eggs. I'm a newbie to this but I'm sure I'll think of some more things. You can make a large batch and keep it handy in the fridge. Make sure you follow the directions, there's no way you can whip this stuff by hand.
D. Figuero	Not With Corn Gluten	Crude Oil is natural, too, but I'm not going to feed it to my dog. Corn is bad for dogs and any vet will tell you that. At least whole corn is human-edible. Corn gluten meal's other use is as an herbicide (shudder). No thank you. Do the research, this also has 'animal fat', which has been listed as a very unhealthy ingredient as well. Purina can try dressing up this stuff as much as they like, it's still bad for dogs and it's filler.If you have enough money to buy yourself snack foods, if you're not eating ramen noodles for the last two weeks of every month and if you can drive through a fast-food joint, then feed your dog food that will extend his life.

TABELLA 8 Recensioni Isolate nei Cluster

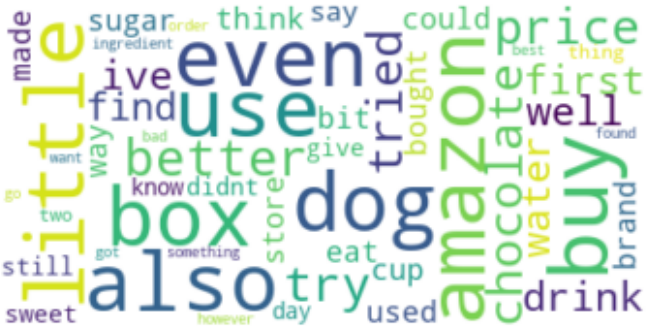

FIGURA 5 Wordcloud Gruppo 1 Gerarchico

Le frasi emerse non appaiono così distanti dal topic generale (*fine food*), però è possibile notare che gli alimenti in questione facciano riferimento a prodotti specifici non largamente utilizzati sul mercato (in particolare nei primi anni 2000, periodo di riferimento del dataset).

4.4 | Conclusioni

I metodi gerarchici non sembrano adatti ad affrontare questo tipo di problema, a differenza di quanto accade per il metodo delle *k-medie*, che identifica gruppi comunque poco significativi in un'ottica di interpretabilità. Inoltre la *Normalized Mutual Information* presenta valori particolarmente bassi, ciò implica che non vi sia alcun legame tra l'appartenenza ad un gruppo ed il punteggio ricevuto. A conferma di ciò, è possibile osservare che la divisione in cluster non ha portato alcuna separazione tra i commenti negativi e quelli positivi.

5 | TOPIC MODELING

L'ultimo task svolto è stato il Topic Modeling. In un primo momento tramite LSA (*Latent Semantic Analysis*) e successivamente tramite LDA (*Latent Dirichlet Allocation*). Queste tecniche sono state applicate esclusivamente al testo lemmatizzato in quanto un lemma è più facilmente interpretabile rispetto ad una radice di una parola.

5.1 | LSA

La LSA si basa sulla decomposizione della matrice iniziale tramite SVD per ottenere i topic. Si è deciso di troncare la matrice decomposta in modo da ottenere 10 tematiche. Di seguito si riportano i risultati.

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

TABELLA 9 Identificazione per Topic

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.

5.2 | LDA

Nel caso della LDA, il numero di topic non è stato fissato a priori ma valutato attraverso la metrica *Coherence* (in particolare attraverso il parametro *c_v* che porta ad un risultato compreso tra 0 e 1), 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); di seguito si riportano le composizioni dei topic.

A differenza di quanto accaduto in precedenza, i topic appaiono molto più interpretabili e coerenti. Ad esempio, il topic 2 mostra dei riferimenti a quelli che possono essere dei piatti principali, come il riso o la pasta. Allo stesso modo, i topic 5 e 7 sono chiaramente identificabili come riguardanti *specifiche alimentari tecniche/sportive* e *thè e tisane*. L'unico topic che sembra

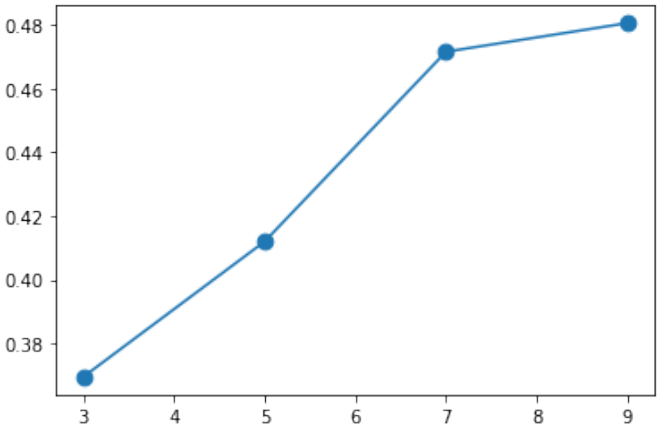

FIGURA 6 Variazione Coherence

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

TABELLA 10 Identificazione per Topic LDA

coincidere con i risultati della LSA, è il topic inerente ad *Amazon e spedizioni* (numero 4). Inoltre, è interessante notare come le parole specifiche dei topic della LDA tendano a ripetersi molto meno rispetto a quanto accade nella LSA.

5.3 | Conclusioni

Dal punto di vista dell'interpretabilità e della separazione dei topic, la tecnica della *Latent Dirichlet Allocation* è risultata migliore rispetto la *Latent Semantic Analysis*. A differenza di quanto emerso dalla Cluster Analysis, appare più chiaro il fatto che ci siano dei gruppi di recensioni simili.

6 | CONCLUSIONI FINALI

Recuperando le domande di ricerca, dopo un'analisi dei dati, è possibile affermare che:

1. *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.
2. *Recensioni simili portano a valutazioni simili?* Raggruppare recensioni simili in termini di contenuto non porta vantaggi dal punto di vista previsionale.
3. *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.

7 | RIFERIMENTI

1. Viviani Marco, *Text Mining and Search* [Lecture notes or PowerPoint slides], a.a. 2020-2021, [Approccio Metodologico];
2. Kaggle, *Amazon Fine Food Reviews*, 2021;
3. Yihui Xie, J. J. Allaire, Garrett Grolemund, R Markdown: The Definitive Guide, Chapman and Hall/CRC, 2018;
4. Kaggle, Notebook di *shashanksai*, 2021.
5. J. McAuley and J. Leskovec, *From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews*, WWW, 2013.