Amazon Fine Food Reviews:

Progetto di Text Mining

DI FRANCESCO ANELLO,
FRANCESCO FUSTINI
E LAURA RAPINO

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Gennaio 2022

Workflow e obiettivo

OVERVIEW

- Analisi esplorativa
- Pre-Processing: dati e testo
- Text Representation
- Riduzione della dimensionalità
- Text Classification
- Text Clustering

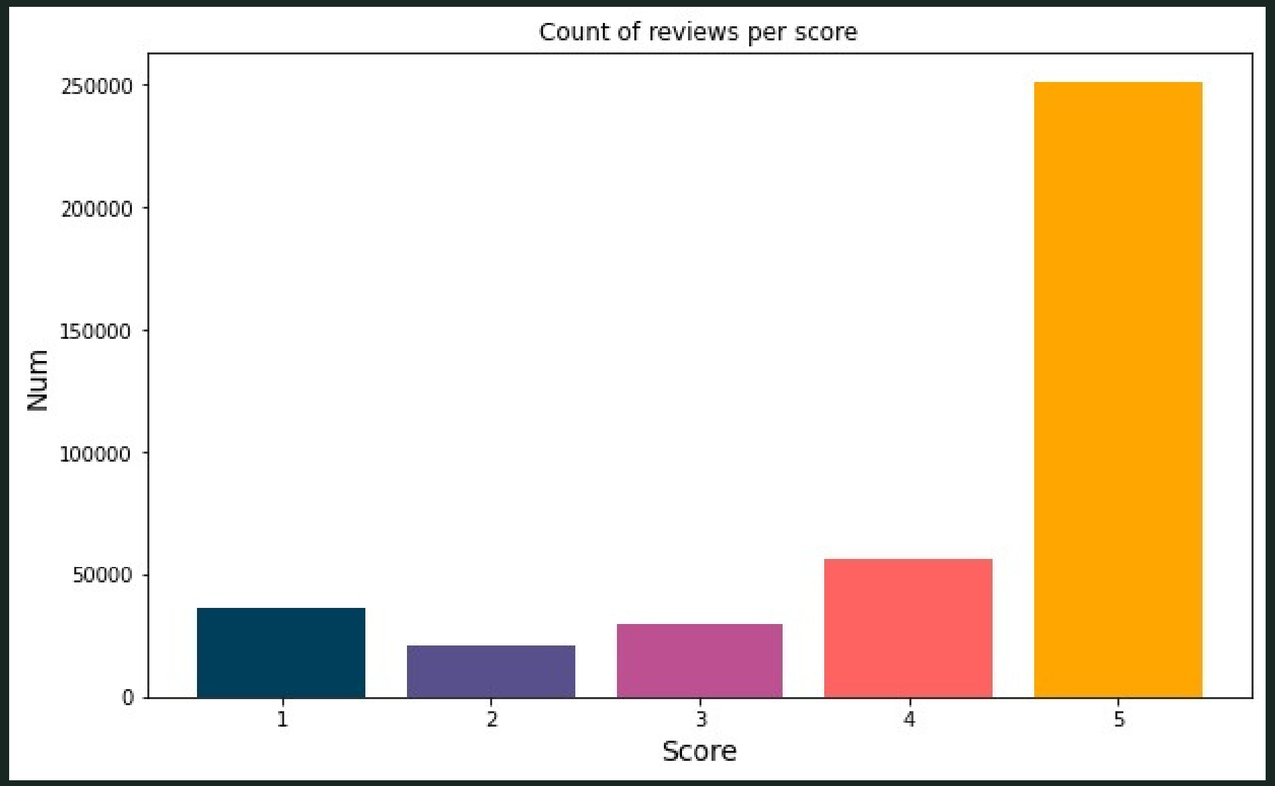
Analisi Esplorativa

VARIABILE SCORE

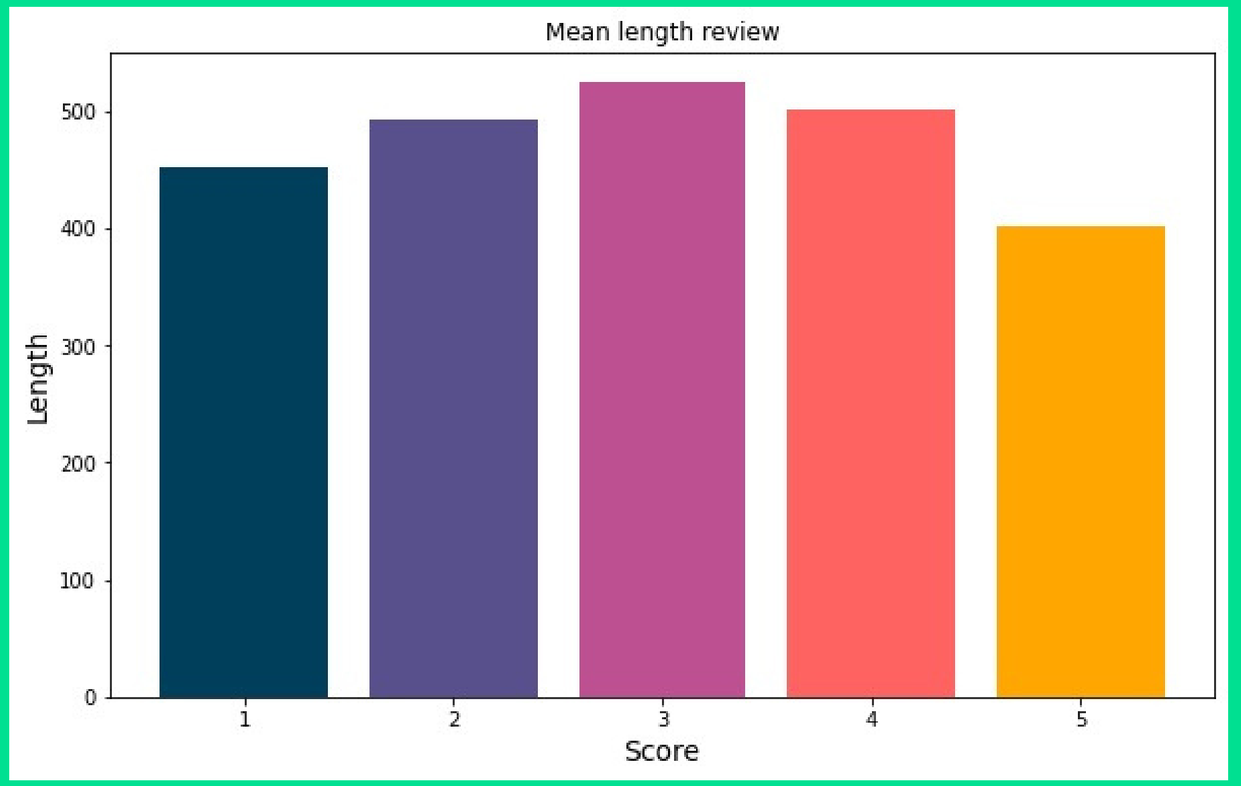
Distribuzione della variabile riguardante il punteggio assegnato alle recensioni studiate.

VARIABILE TEXT

Vizualizzazione della lunghezza media delle recensioni per punteggio di Score



Sbilanciamento distribuzione per Score = 5



Lunghezze maggiori per Score = 3

Pre-processing

DATI

- DATA BALANCING

TESTO

- PAROLE IN MINUSCOLO
- RIMOZIONE DI:
  - CONTRAZIONI
  - URLS
  - PUNTEGGIATURA
  - NUMERI
  - SPAZI
- TOKENIZATION
- RIMOZIONE DELLE STOPWORDS
- LEMMATIZATION & STEMMING

Text Representation & Riduzione dimensionalità

Strutturazione dati:

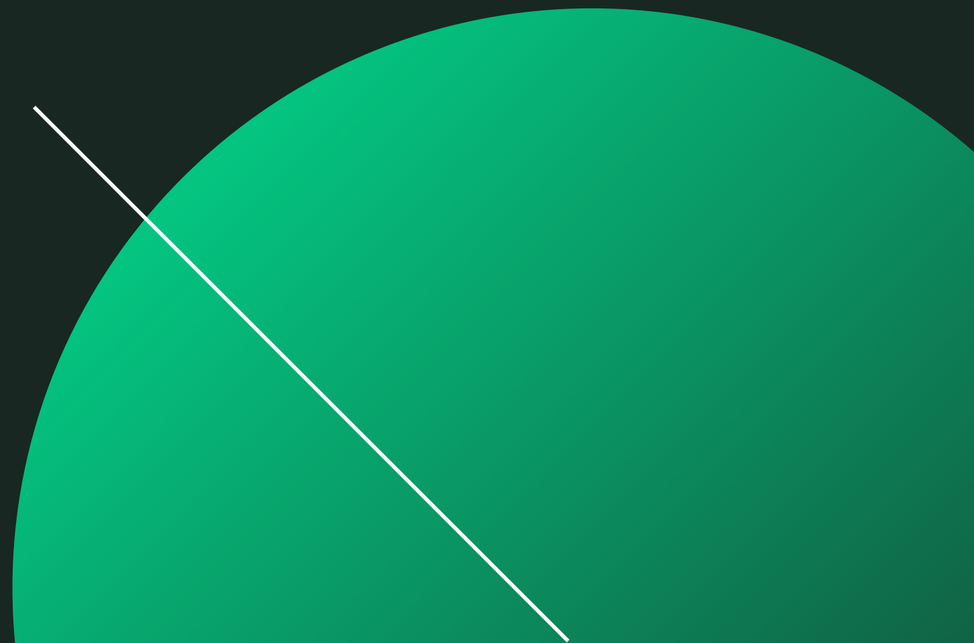
BAG OF WORDS

PESO TF-IFD

- uni e bi-grammi
- 30.000 più frequenti

Metodo usato:

SVD → 300 COMPONENTI



Text Classification & Text Clustering

I MODELLI UTILIZZATI

Text Classification

TRAINING E TEST SET

70% e 30% rispettivamente
con successiva standardizzazione

XGBOOST

SVM

K-NN

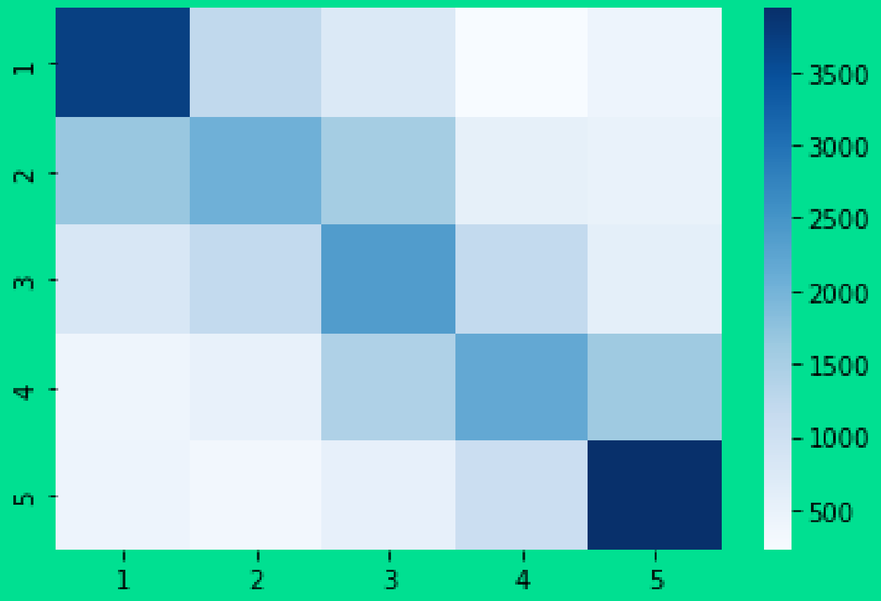
GRIDSEARCH CON CROSSVALIDATION

Tuning dei parametri sulla
rappresentazione BoW stemmata
considerando 3 fold

EVALUATION

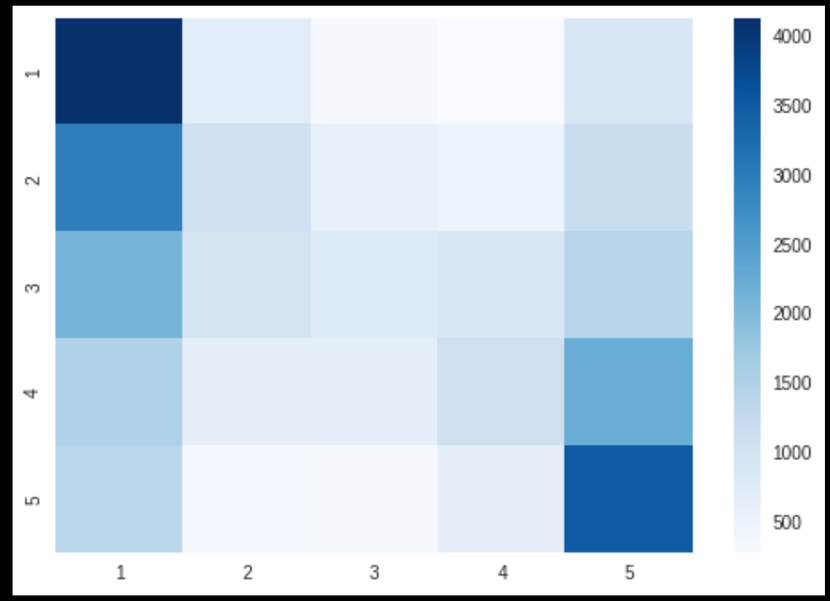
Precision, Recall, F-score, Accuracy

XGBOOST & SVM



CLASS	PRECISION	RECALL	F-SCORE
1	0.57	0.61	0.59
2	0.39	0.36	0.37
3	0.40	0.38	0.39
4	0.43	0.41	0.42
5	0.59	0.66	0.62

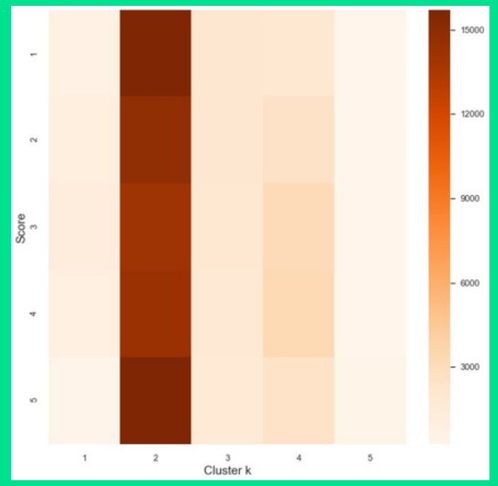
KNN



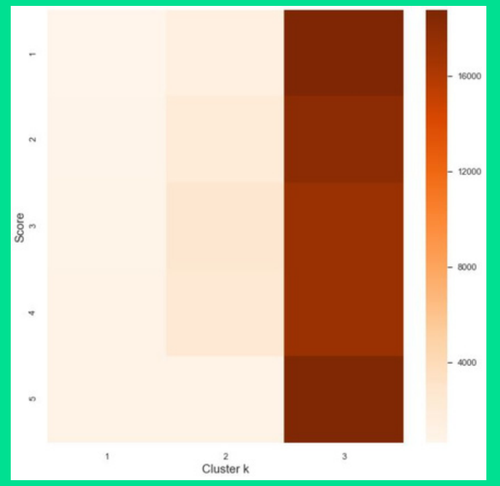
CLASS	PRECISION	RECALL	F-SCORE
1	0.34	0.65	0.45
2	0.28	0.17	0.21
3	0.29	0.13	0.18
4	0.32	0.18	0.23
5	0.38	0.56	0.45

Text Clustering

K-MEANS



k=5, Lemmed Bag of Words



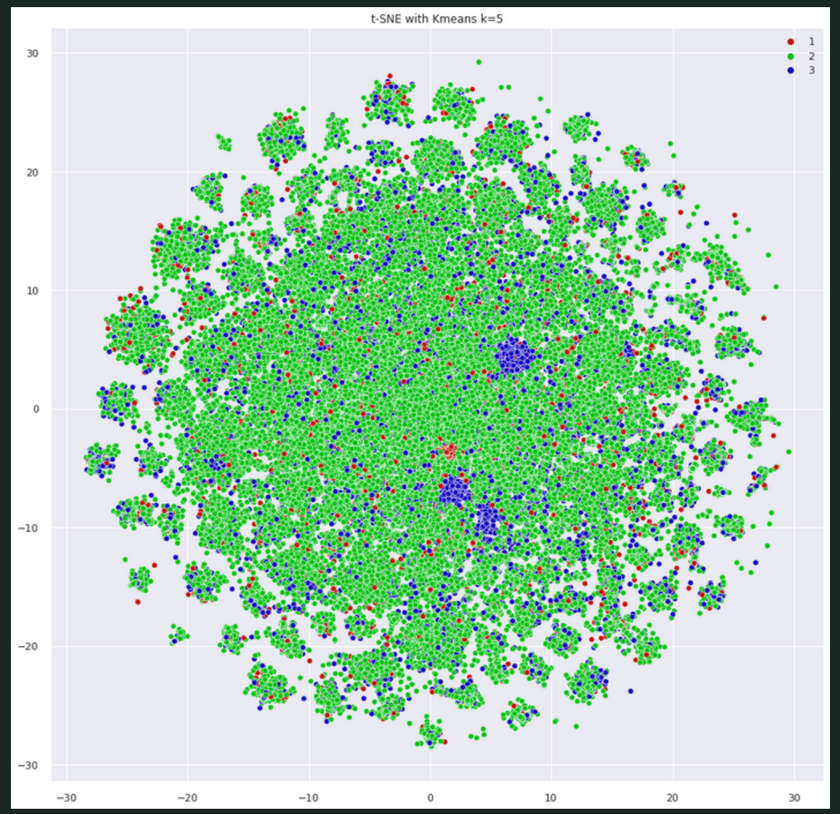
k=3, Lemmed Bag of Words

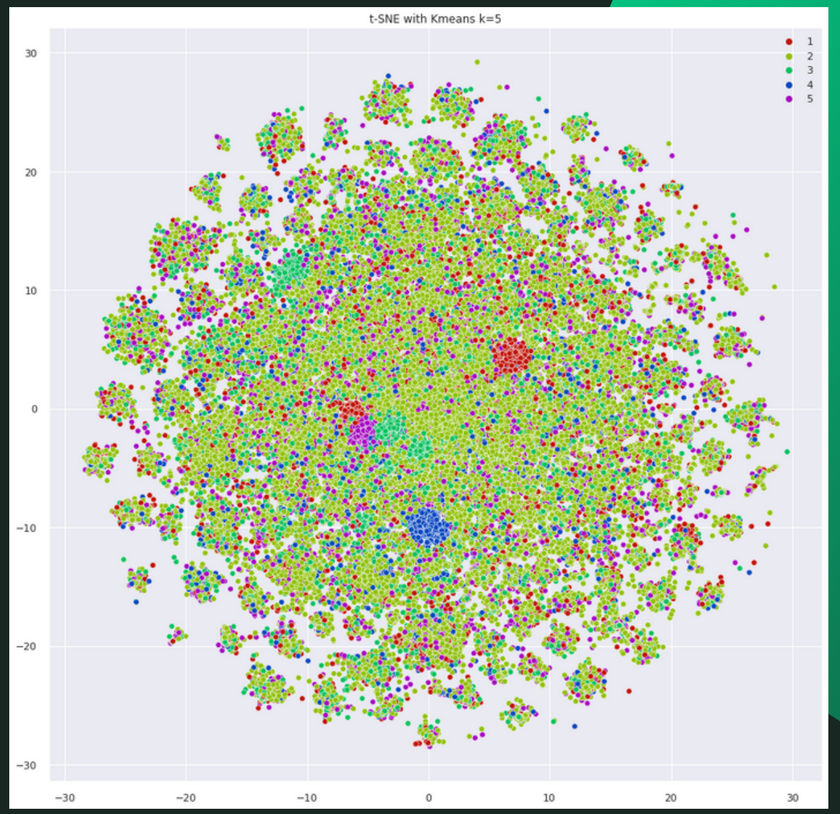
La tecnica di clustering k-Means non identifica gli stessi gruppi identificati dalla variabile *Score*. Infatti le osservazioni vengono principalmente associate ad un singolo gruppo sia nel caso k=3 che nel caso k=5.

CLUSTERING GERARCHICO

I cluster gerarchici agglomerativi testati con legame semplice e completo, metrica euclidea e di Manhattan non danno risultati migliori, infatti accentuano maggiormente il fenomeno dell'associazione delle osservazione ad un gruppo principale osservato nel k-Means.

t-SNE → 2 componenti → grafico di dispersione





Conclusioni

TEXT CLASSIFICATION:

- XGBoost-SVM vs k-NN per performance

- XGboost vs SVM per quantità di risorse necessarie

- Possibilità di ridurre la classificazione a 2-3 classi

TEXT CLUSTERING:

- k-means e gerarchico agglomerativo inefficaci