Sentiment Analysis su Amazon Reviews

Scopo del progetto

- Importare le recensioni e modellizzare le entità
- Analizzare e distinguere i topic che caratterizzano un utente
- Distinguere il sentiment dei diversi topic
- Creare una view che permette di visualizzare le analisi fatte

Modellizzazione Dati

Appoggiandosi ad un database **SQLite** e modellizzando i dati con **Django** è stato più semplice gestire le entità rilevate

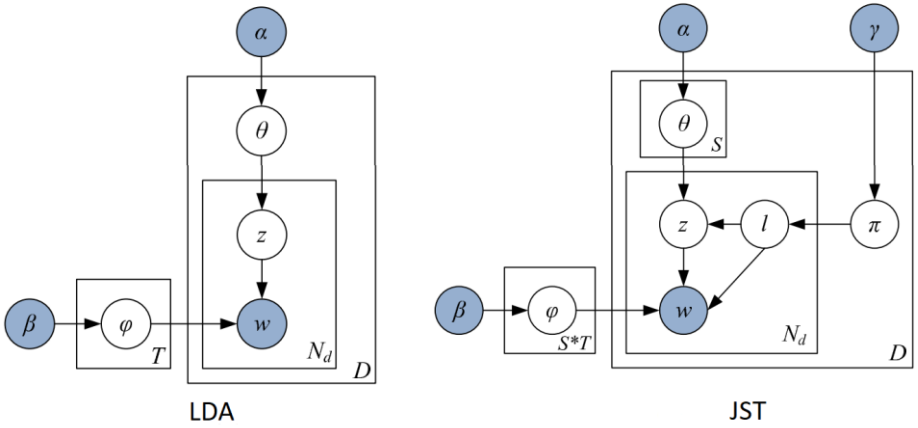
- **User**: (id, num_rating, average_score, variance, experience, pos_words, neg_words)
- **Product**: (id, num_rating, average_score, variance, experience, words)
- **Rating**: (id, *productid*, *userid*, score, text, timestamp)

Dataset alternativo

- Formato .tsv
- 85% voti positivi
- Pochissime recensioni per utente
- Assenza timestamp

- Formato .json
- Voti più distribuiti
- Utenti con parecchie recensioni
- Presenza timestamp

Join Sentiment/Topic Model



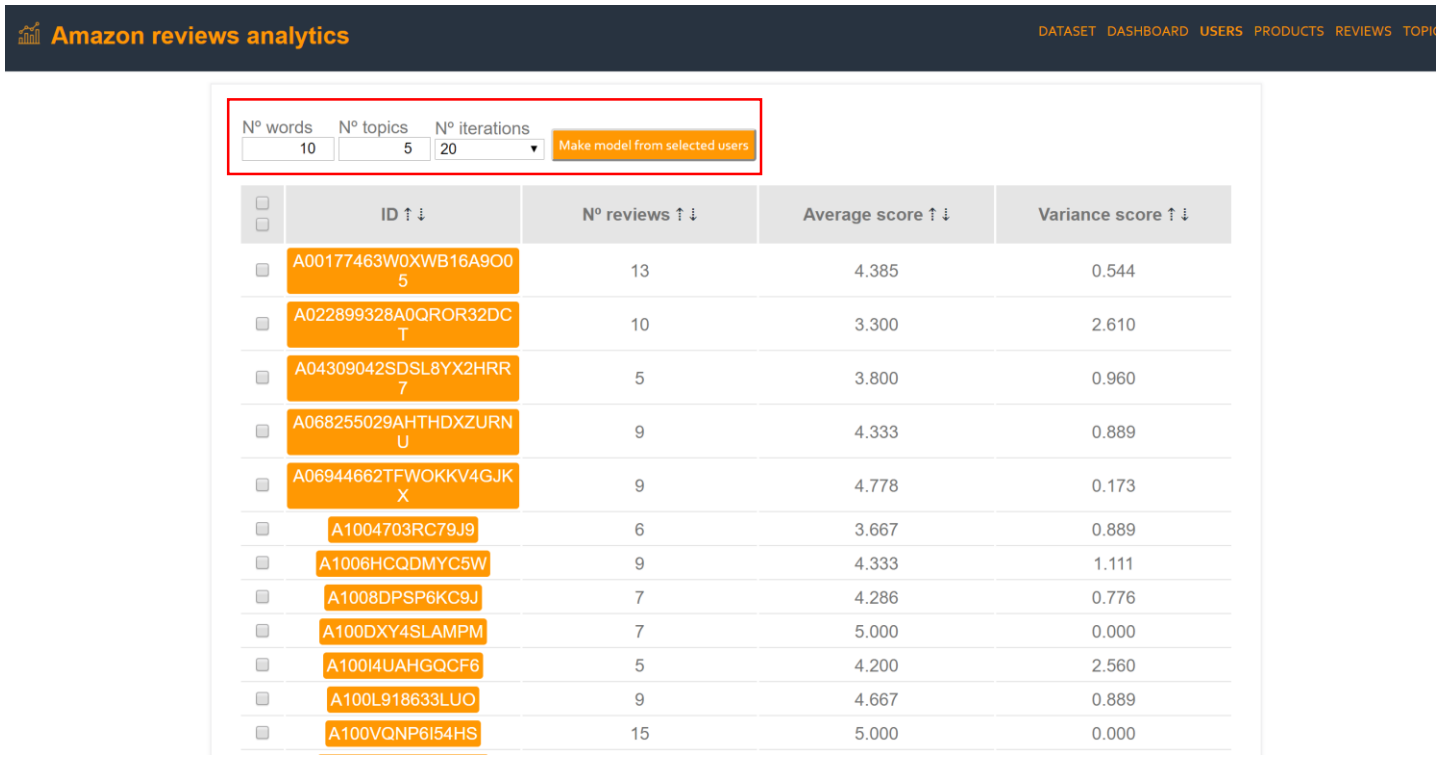
Modello basato su LDA.

Calcola sia topic che sentiment simultaneamente

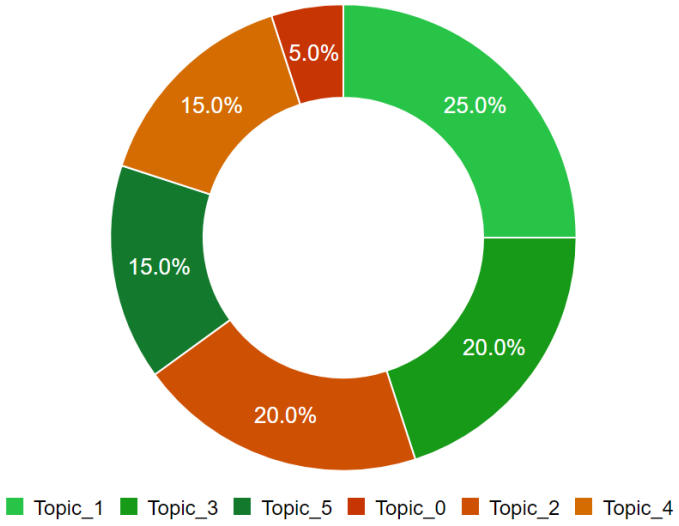
- Necessario fare diverse prove per gli iperparametri α, β e γ
- E’ importante sapere il numero di topic da trovare
- Selezionare un adeguato numero di parole per topic

Creazione modello

Possibilità di selezionare gli utenti da utilizzare per creare il modello



Prodotto



Distribuzione di topic per prodotto

- Parole che caratterizzano il prodotto calcolate con **LDA**
- Intersezione con i topic per capire la distribuzione

Utente

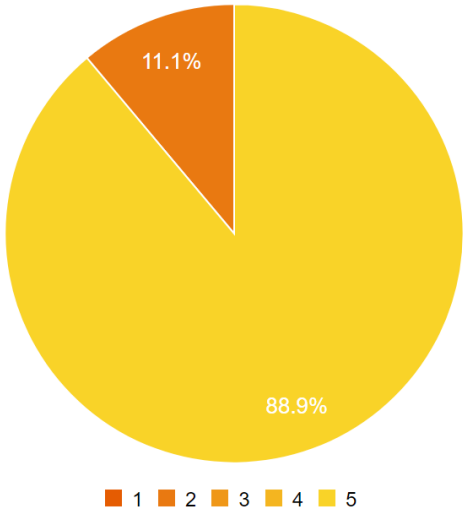
Parole positive (LDA)

flavor wrote differ realli celiac invest similiar textur buyer cream actual cracker ident everyon belowi better cooki

Parole negative (LDA)

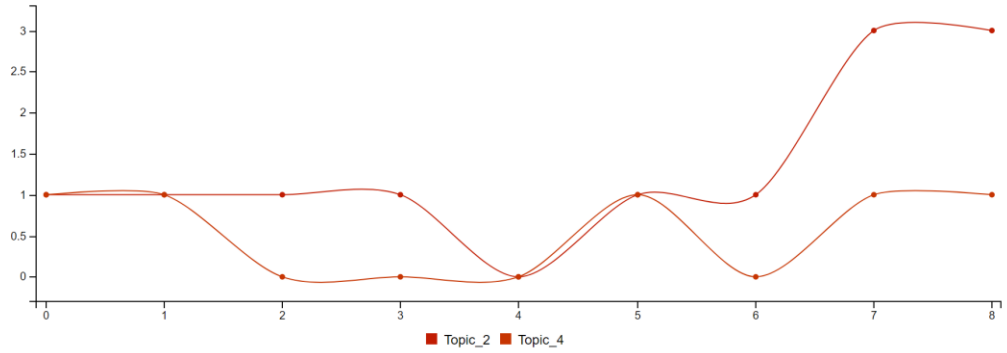
absolut anoth product small textur enfamil refus thing sever disappoint confus truli review posit

Distribuzione dei commenti per utente:



Utente (topic nel tempo)

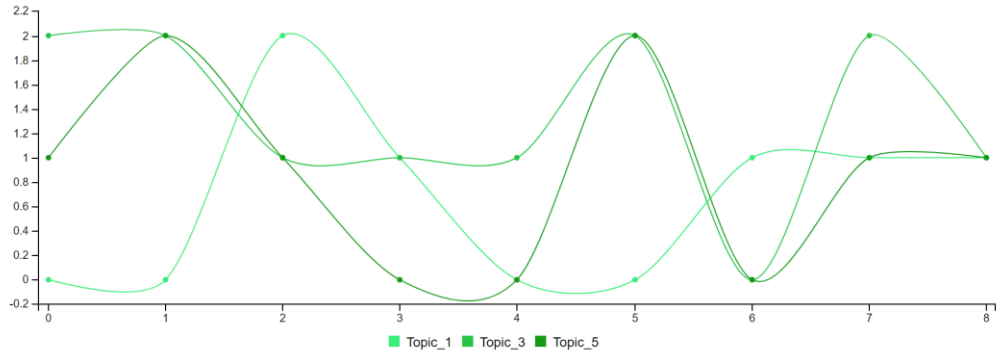
Negative topics



Topic_2	coffe tast cup bar chocol bad flavor calori product roast weak tea drink howev count dark far brew make squar
Topic_4	tast flavor buy bag use know review star size pea drink result price rice product seed time think order day

Topic_1	like good coffe tast cup sweet better love flavor tri enjoy brew roast bit water vue realli brand want brewer
Topic_3	tea use product great price green amazon sugar time food best know stuff love year like make tast drink tri
Topic_5	oil make fat use calori review product recip cook bag coconut piec contain eat box delici sauc meal read tasti

Positive topics



amazon

Conclusione

- Iperparametri fondamentali nella creazione del modello
- Importante selezionare con cura il numero dei topic
- Poche parole (ma anche troppe!) per topic portano il modello a non trovare la giusta distribuzione per utente
- Utilizzare un numero accettabile di iterazioni