RC per birre usando collaborative filtering Item-based

Andrea Iskander Belkhir | SII | 08/06/2023

Introduzione

Questa relazione è basata sul progetto per il corso Sistemi intelligenti per internet dove è stato creato un sistema di raccomandazioni usando il paradigma collaborative filtering applicato nella versione Item-based in un dominio di birre artigianali.

Dati

Per questo progetto è stato usato un dataset trovato online presso datatworld [1], questo dataset contiene 10 anni di reviews del forum beeradvocate e entry è definita da diversi attributi fra cui andremmo ad usare solo quelli per applicare un CF base ovvero user_name,beer_name e rate_overall a qui andremmo ad aggiungere degli ID

ANALISI DATI

La prima parte del progetto è stata uno studio di questi dati presenti nel data set.

Nel dataset sono presenti un totale di 33388 utenti unici e 56857 birre uniche.
sono anche stati effettuati degli istogrammi per vedere le relazioni fra numero birre e numero review e fra numero utenti e numero review

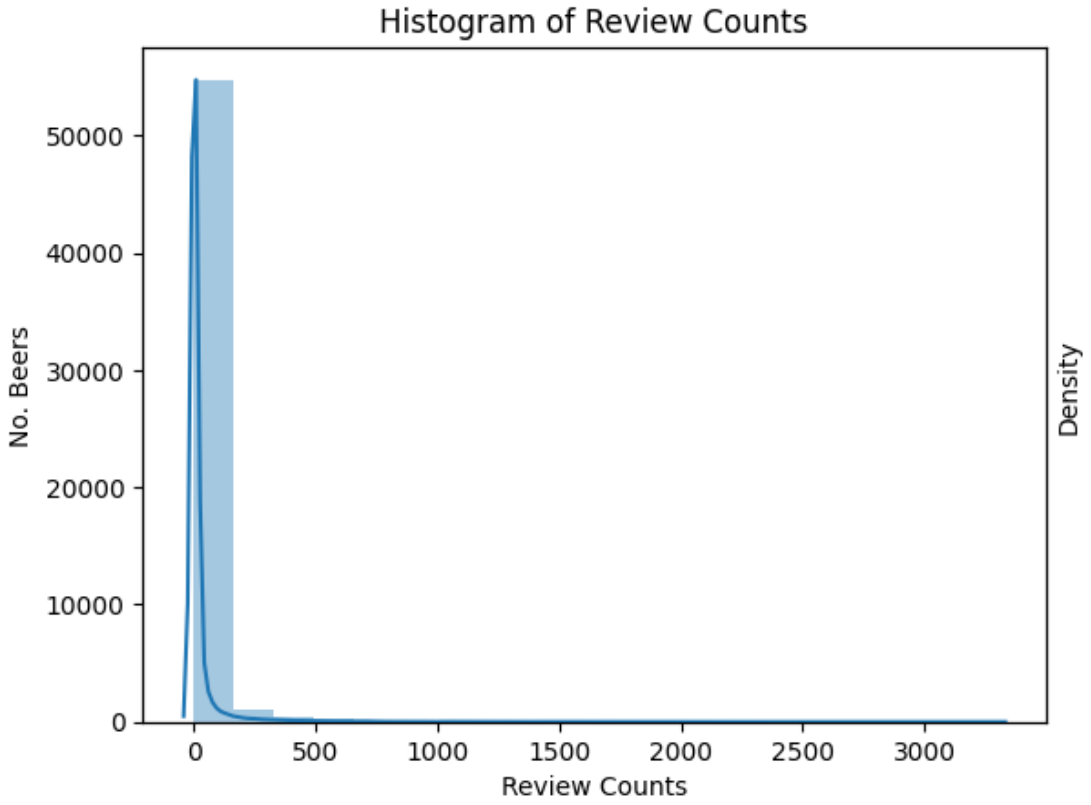

Figura 1 #beer & #reviews

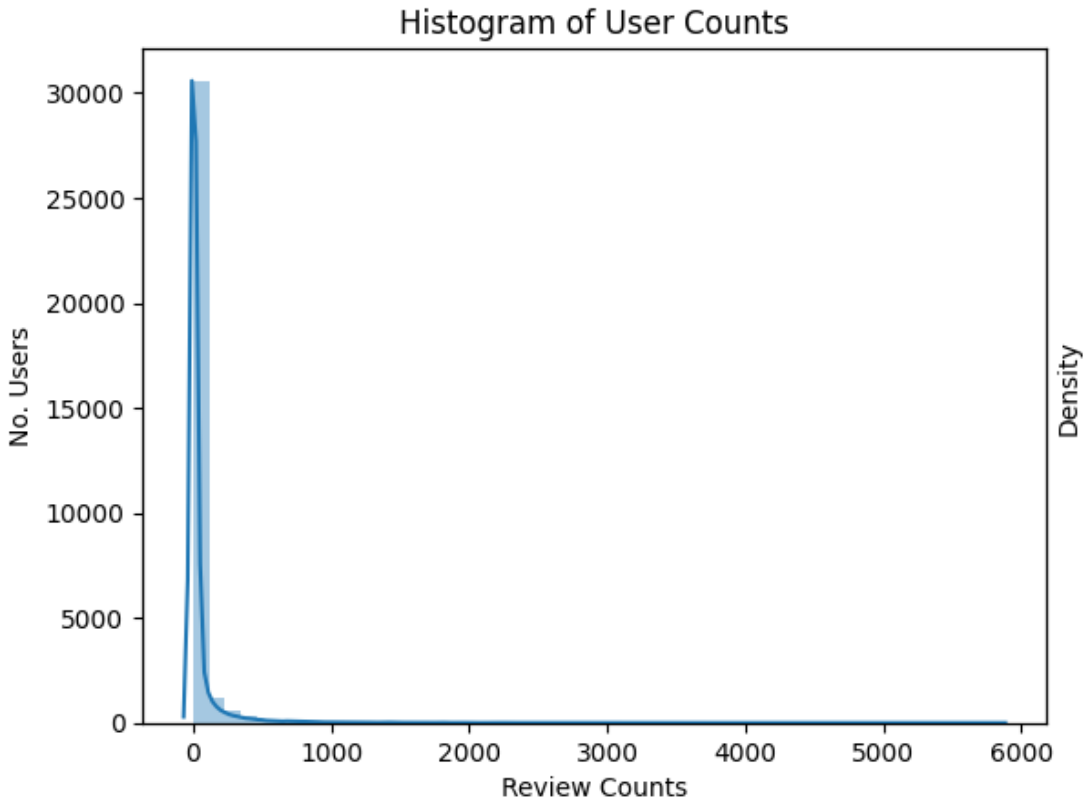

Figura 2 #user&#reviews

Fra le altre anali è stato visto piu precisamente (anche se non prendendo lo spettro completo) quante review hanno gli utenti:

- 10443 users rated 1 or less beers
- 14550 users rated 2 or less beers
- 16976 users rated 3 or less beers
- 18576 users rated 4 or less beers
- 19821 users rated 5 or less beers
- 23198 users rated 10 or less beers
- 24846 users rated 15 or less beers
- 25863 users rated 20 or less beers

Per poi andare a vedere attraverso l'uso dei quantili come si distribuiscono le reviews degli utenti e i rating delle birre andando a scoprire che circa 85% degli utenti ha almeno 20 review e il 15% delle birre ha almeno un voto di 3 su 5 mentre il 75% ha almeno 4.

Attraverso questo studio sui dati è stato effettuato un taglio del dataset andando quindi a creare un secondo avente le parti piu influenti del primo, questo è stato fatto sia per un motivo di rimozione rumore (ovvero andando a rimuovere birre con troppo poche review) ma soprattutto per problemi legati alla RAM durante il training, infatti il numero di birre uniche passa da piu di 56000 a 3640(ovviamente in un approccio reale questo non buono anche perche vengono usate solo birre con rating alte ignorando quelle basse).

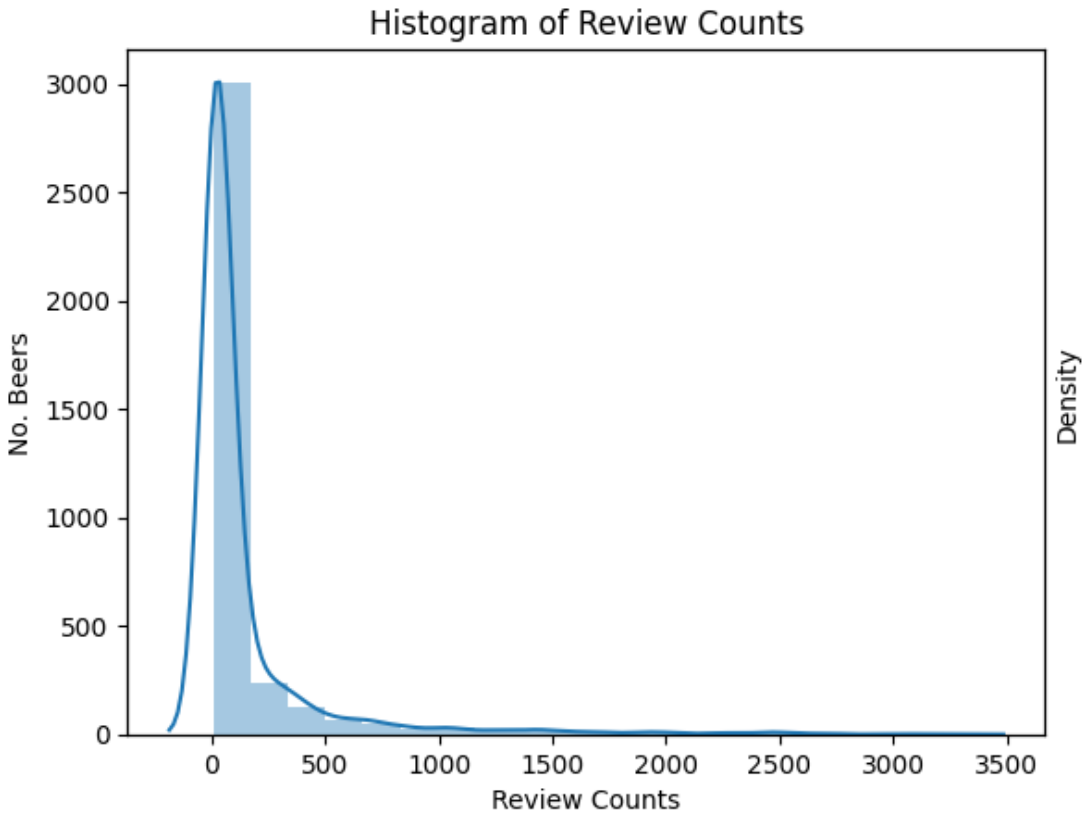

Figura 3 Istogramma nuovo per birre

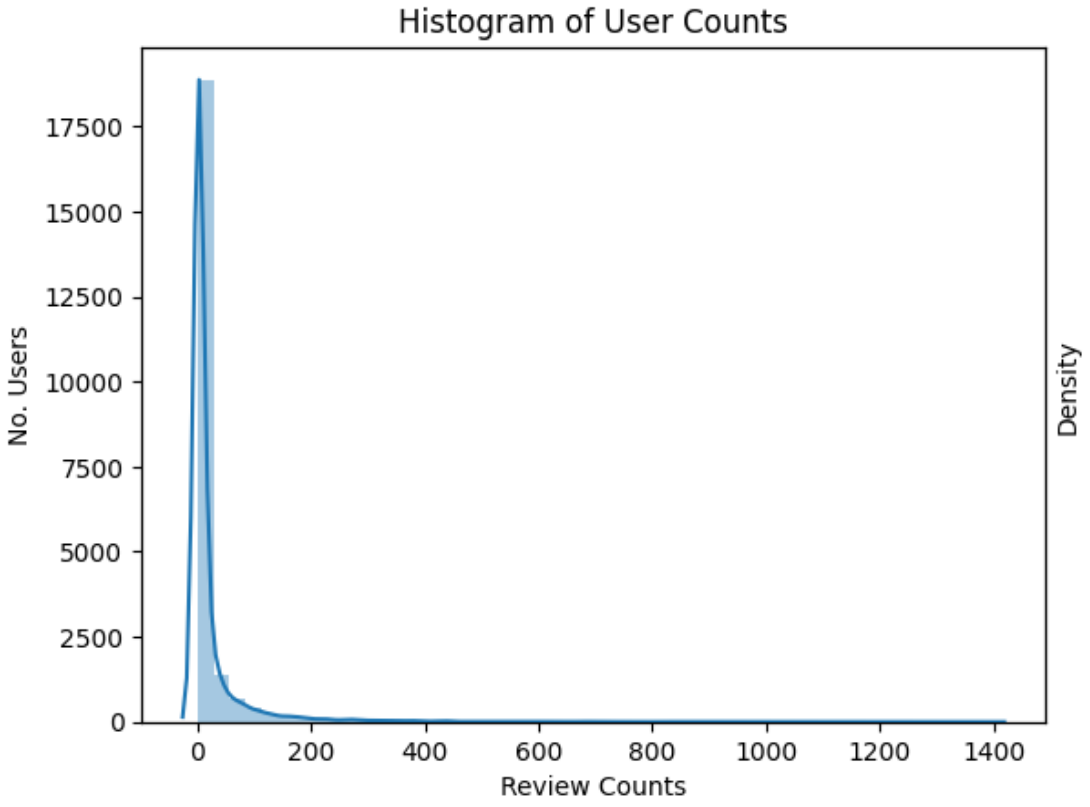

Figura 4 Istogramma nuovo per utenti

Training

Per questa parte è stato usato il framework Surprise [2] a cui sono state modificate delle funzioni per andare a funzionare nel nootebook ovvero come passare da ID a nome della birra e prendere K raccomandazioni per un item (questi cambiamenti sono giusta per interfacciarsi con le nostre variabili).

L'algoritmo è stato addestrato usando la similitudine di Pearsone usando la baseline al posto della media e un algoritmo KNNBaseline, mentre per fare una vadidation è stata usata la **root-mean-square error** e la **Mean Absolute Error.**

RACCOMANDAZIONI

Per vedere che l'addestramento non sia stato troppo bias durante le prove di raccomandazione ho preso le top 20 birre per rating e numero di reviews e ho incrociato questi set con le 20 raccomandazioni delle birre in cima a queste top 20 per vedere se le raccomandazioni sono bias ovvero raccomandavano semplicemente cio che è migliore o se raccomandavano qualcosa di simile(anche se non possiamo dimostrarlo) alla birra su cui abbiamo chiesto le raccomandazioni, queste intersezioni sono vuote in caso o con una sola birra nell'altro quindi possiamo dire che le raccomandazioni sono basate sul item argomento della raccomandazione e non su fattori esterni.

Riferimenti

[1] [Online]. Available: https://data.world/socialmediadata/beeradvocate.

[2] [Online]. Available: https://surpriselib.com/.