News ed engagement - Articoli Virali

SEZIONI

Introduzione

Questa relazione ha come obbiettivo l'analisi di un dataset arbitrario da scegliere tra quelli disponibili su Kaggle rispettando alcuni vincoli tra cui:

- Utilizzo di libreriaìe pandas, numpy
- Utilizzo d Boolean/Fancy indexing
- Presenza di grafici
- Lettura e scrittura di file

Individuato il dataset Internet news data with user engagement ci si è posta la seguente domanda:

> Gli articoli "virali" sono isolati oppure hanno hanno degli effetti duraturi dull'engagement? Ci sono delle avvisaglie o si presentano senza preannuncio?

Per l'analisi si è tratta forte ispirazione dal notebook Melboune Real Estate Market comprehensive Analysis

Preparazione dei dati

In questa sezione si fa riferimento al notebook: Preparazione Dataset

1. Scrematura colonne

Già da un primo squardo al dataset si nota che molte delle colonne non saranno necessarie alla nostra analisi, essendo noi interessati alla relazione tra il titolo e le interazione con l'articolo, quindi vengono eliminate.
Le colonne *source_id* e *source_name* contengono di fatto la stessa informazione; viene quindi creta una tabella di dettaglio indicizzata su *source_id* per eliminare dal dataset *source_name*. Cosi facendo si nota un refuso nei dati, che viene eliminato.
Si aggiunge una colonna contenente il numero di articoli disponibili per ogni testata alla tabella di dettaglio, per verificare il numero di articoli per testata. Gli 82 articoli di ESPN potrebbero essere considerati non sufficienti per proseguire con l'analisi, ma si decide di tenerli almeno fino ad una prima valutazione delle interazioni.

Vengono poi aggiustati i nomi delle colonne ed i relativi dtype per comodità di utilizzo.

2. Gestione dei null

Per comodità di visualizzazione si disegna una heatmap per evidenziare la distribuzione dei valori nulli, in questo modo è facile vedere che appartengono tutti agli stessi articoli. Essendo così pochi e concentrati, è facile pensare a qualche errore avvenuto in fase di raccolta dei dati, quindi vengono scartati.