News ed engagement - Articoli Virali

SEZIONI

1. Introduzione
2. Preparazione dataset
    1. Scrematura
    2. Gestione dei null
3. Analisi dei dati
    1. Distribuzione interazioni
    2. Interazioni per fonte
    3. Normalizzazione interazioni
    4. Grafici
4. Conclusioni
5. Possibili Sviluppi

Introduzione

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

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

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

> Qual'è, se esiste, l'orario migliore per pubblicare un articolo di giornale?

Per rispondervi si è considerato il rapporto tra il numero di interazioni dei lettori e l'orario di pubblicazione degli articoli.

Preparazione dataset

notebook: Preparazione Dataset

Si inspeziona il dataset con lo scopo di alleggerirlo il più possibile. Vengono inoltre gestiti i pochi casi di dati mancanti.

1. Scrematura

Già da un primo squardo al datasetsi nota che molte delle colonne non ci saranno d'aiuto. Di tutte le informazioni relative al contenuto degli articoli viene mantenuto solo il titolo, necessario per identificare gli articoli e mantenere un contesto.
Le colonne *source_id* e *source_name* sono ridondanti. Viene quindi creata una tabella di dettaglio che mappa l'idenrificativo della fonte al nome completo, questa tabella conterrà tutte le informazioni delle testate giornalistiche.\ Esaminandola si identifica una *source* fittizia, che viene eliminata.

Alla tabella di dettaglio viene aggiunto il numero di articoli presenti per ogni *source*. In questo modo è possibile valutare quanto la singola testata giornalistica sia rappresentata nel dataset.
ESPN, con 82 articoli, risulta essere è significativamente meno presente rispetto alle altre e potrebbe essere eliminata. Si sceglie però di tenerla per osservarne l'engagement.

2. Gestione dei null

Si traccia una heatmap che mette in evidenza la distribuzione dei valori nulli nel dataset. Sono pochi e fortunatamente appartengono tutti agli stessi tre articoli, che vengono perciò scartati.

Analisi dei dati

notebook: Analisi Dati 1

Distribuzione interazioni

La distribuzione delle interazioni su tutti gli articoli, non divisi per testata, risulta pesantemente spostata verso lo zero, con valori massimi piottosto elevati. Ciò significa che un numero ridotto di articoli conta la vasta maggioranza delle interazioni. Potremmo definire questi articoli "virali".

Risulta inoltre evidente che la colonna *comment_plugin* presenta un numero di interazioni molto ridotto rispetto alle altre. Per approfondire questa anomalia si procede al conteggio degli articoli con engagement non nullo, divisi per tipo di interazione.
Con solo 50 aricoli che presentano un'interazione ed un massimo di 15 interazioni per articolo si rafforza *comment_plugin* non sia un tipo di interazione che non rappresenta fedelmente l'engagement dei lettori, e che sia perciò da scartare.

Studiando la distribuzione delle interazioni divise per fonte, l'ipotesi di cui sopra ha la sua conferma, in quanto la maggior parte delle testate non registrano interazioni per *comment_plugin*, che viene quindi scartata.
In questa fase si nota anche che la testata ESPN, oltre ad essere sottorappresentata per numero di articoli, ha engagement pari a zero. Si decide di scartarla.

Interazioni per fonte

Si realizza quindi una serie di istogrammi per rappresentare la distribuzione delle interazioni divise per fonte e per tipo, mettendo in evidenza l'articolo con engagement maggiore.
Guardando questi grafici si ipotizza che la grande differenza di interazioni tra l'articolo migliore ed il resto degli articoli sia legata ad una notizia o avvenimento di particolare interesse.
Se così fosse introdurremmo un bias nella nostra indagine analizzando questi articoli insieme agli altri in quanto è ragionevole assumere che questi siano pubblicati "il prima possibile" rispetto all'avvenimento in questione, il che influenzerebbe la relazione orario-engagement.

Viene quindi stilata una lista dei titoli ed orari di pubblicazione degli articoli migliori divisi per ogni fonte e tipo di interazione.
Da questa lista si scopre che gli articoli migliori coprono notizie differenti gli uni dagli altri. Non solo, per una stessa testata l'articolo con più interazioni spesso cambia a seconda del tipo di interazione. Si abbandona quindi l'ipotesi di un bias introdotto dagli articoli virali.

notebook: Analisi Dati 2

Normalizzazione interazioni

Vista la grande sproporzione del numero di interazioni per i tre tipi considerati, *reactions*, *comments* e *shares*, si rende necessario stabilire un criterio comune di valutazione dell'engagement.

Una possibilità sarebbe considerare ogni tipo di interazione separatamente. In questo caso si otterrebbe una rappresentazione più accurata del coinvolgimento dei lettori nelle diverse fasce orarie in quanto sarebbe possibile distinguere le categorie di interazione.
Tuttavia per rispondere alla domanda che ci siamo posti inizialmente questa distinzione non ci interessa, anzi comporterebbe un appesantimento nell'interpretazione dei risultati. Si sceglie quindi di combinare le interazioni delle diverse categorie normalizzandole a 1, in modo da bilanciare la sproporzione di numero tra le categorie.

Grafici

Tracciati i grafici ora-interazione si nota immediatamente l'assenza totale di interazioni tra le 19 e mezzanotte. Questa è quasi certamente una mancanza del dataset selezionato e non si può fare altro che prenderne nota.

Da un primo sguardo ai due grafici a dispersione delle interazioni di tutte le sorgenti si nota un aumento considerevole del'engagement nella fascia oraria compresa tra le 9:00 e le 18:00, come era raginevole aspettarsi.
Osservando con attenzione il grafico "Scatter Interazioni", in cui le *source* sono rappresentate separatamente, si nota che i minimi di engagement mostrano due massimi di entità confrontabile localizzati alle 11:00 ed alle 15:00.

"Scatter Interazioni Totali" raffigura la media tra le interazioni delle fonti per ogni ora, di conseguenza i punti di questo grafico sentiranno maggiormente il peso degli articoli virali.
Tuttavia, anche in questo caso si ripresentano i due massimi evidenziati in precedenza, ma quello delle 15:00 risulta molto più marcato arrivando a contare oltre il 10% delle interazioni. L'unico altro punto a superare il 10% è alle 14:00.
Si nota la comparsa di un terzo massimo alle 5:00, tuttavia è di intensità inferiore rispetto agli altri. Dato il marcato isolamento del punto di massimo si suppone che sia quasi interamente dovuto ad un articolo virale.

Rivolgendo l'attenzione alla provenienza delle testate giornalistiche ci si accorge che due, **The Irish Times** e **Al Jazeera English** non sono americane. Purtroppo il dataset non fornisce informazioni sul fuso orario con cui sono stati raccolti i dati, quindi non è possibile sapere se tutti gli orari si riferiscono allo stesso fuso ne tentare aggiustamenti.
Si possono però escludere le due testate sospette e tracciare dei grafici locali per osservare eventuali cambiamenti.

Dal nuovo "Scatter Interazioni Totali" si nota un abbassamento del numero di interazioni per quasi tutti i punti precedenti alle 13:00 comprese, con conseguente aumento per quelli successivi.
Il punto delle 5:00, che prima era un massimo ed un punto isolato, si è abbassato notevolmente spostando il picco verso le 4:30.

Conclusioni

Esaminando i grafici ottenuti si giunge alla conclusione che il momento milgiore per pubblicare un articolo in termini di user engagement sia il primo pomeriggio, tra le 14:00 e le 15:00.

La mattina riscontra in generale meno interazioni, che si massimizzano attorno alle 11:00.

Possibili Sviluppi

Articoli virali

I grafici presentati presentano picchi spigosi in concomitanza degli articoli virali. Si potrebbe formulare un'analisi simile a questa eliminando però gli articoli virali dal dataset. Così facendo l'andamento temporale dell'engagement rifletterebbe meglio qeullo de "l'articolo medio".
Sarebbe poi interessante affiancare una rappresentazione della frequenza degli articoli virali per ogni ora.

Top article

In questo studio non si è tenuto conto del fatto che la testata giornalistica avesse pubblicato gli articoli come *top_article* o meno. Si potrebbero riscontrare differenze rilevanti nell'andamento dell'engagement trattando i due tipi di articoli separatamente