Master di I livello in Analisi Dati per la Business Intelligence e Data Science

A.A. 2017/2018

Titolo della tesi: “Libri e social network: criteri per la costruzione di una top ten degli autori”

Autore: Fabio Bruno

Abstract

Scopo del progetto di tirocinio è stato quello di realizzare uno studio di fattibilità che ha come contesto di applicazione il mondo dell’editoria, nello specifico un grande distributore di prodotti editoriali.

Il presente lavoro ha come obiettivi quelli di arricchire le informazioni a disposizione sui prodotti editoriali e di pervenire ad uno *score* per ogni autore sulla base dei risultati ottenuti da diverse fonti web, quali Google Trends, Facebook, Twitter: ciò permetterà di ottenere una classificazione con gli autori del momento.

Il progetto può essere diviso in 3 fasi:

- Fase 1: arricchimento delle informazioni sui libri

  Questa fase ha come scopo quella di arricchire le informazioni a disposizione sui libri. Il dataset di partenza è stato costruito estraendo in modo casuale alcuni titoli dal database del cliente e arricchendolo con i libri nelle prime posizioni delle classifiche di vendita attuali. Per ogni libro si ha a disposizione il codice ISBN, il titolo del libro e l’autore: il codice ISBN è indispensabile per ottenere, richiamando le API, il codice univoco che Google assegna ai libri presenti all’interno del database di Google Libri. Nell’ordine di arricchire il contenuto informativo della scheda del libro vengono richiamate nuovamente le API di Google Libri, inserendo come input di ricerca il codice univoco ottenuto precedentemente. L’output è una tabella che contiene, oltre alle informazioni bibliografiche, anche descrizione e categorie.

  A partire dalla descrizione vengono estratte delle parole chiave che possano identificare i contenuti e le tematiche del libro. A tale scopo sono stati utilizzati il servizio Text Analytics messo a disposizione da Azure, la piattaforma cloud pubblica di Microsoft, e il servizio di analisi semantica del testo Dandelion API.

- Fase 2: calcolo del social engagement
  La seconda fase pone al centro del progetto gli autori. In particolare l’attenzione si sposta, in primo luogo, sull’interesse di ricerca che è in grado di generare su Google calcolato da Google Trends: per ottenere i dati da Google Trends è stato utilizzato il pacchetto R “gtrendsR”.

In secondo luogo viene calcolato l'engagement e il sentiment generati dagli autori su Facebook e Twitter. Per quanto riguarda l'ottenimento dei dati utili a calcolare l'engagement vengono richiamate le API delle piattaforme attraverso Power Query e RStudio. Attraverso lo stesso procedimento sono stati ricavati i commenti, i quali rappresentano l'input da inviare al servizio Text Analytics di Azure per il calcolo del sentiment.

A conclusione di questa seconda fase si è cercato di riassumere i risultati ottenuti in uno score, pesato sulle varie fonti.

- Fase 3: costruzione di una *dashboard* con *PowerBI*

  La terza e ultima fase ha lo scopo di dare valore ai risultati ottenuti nelle fasi precedenti attraverso la costruzione di dashboard che rendano i dati fruibili ed esplorabili.

Gli strumenti utilizzati sono stati Excel, R, Power Query e Power BI.