Analisi delle correlazioni tra le tendenze di youtube e quelle di instagram

Gabriele **inserire cognome**, Federico Luzzi, Marco Peracchi , Christian Uccheddu

---

Introduzione e obiettivi

Questo lavoro si occupa di analizzare le correlazioni presenti tra le tendenze di youtube e quelle di instagram nel tempo. In particolare attraverso l'uso di indicatori creati ad hoc ci si è occupati di cercare il prototipo di personaggio con video ideale per un dato periodo dell'anno, in questo caso le festività natalizie.

Raccolta dati

In questa sezione del report verrà esposta la procedura con cui sono stati acquisiti i dati e come sono state affrontate le problematiche riscontrate.

Scelta strumenti

Per prima cosa ci siamo dovuti occupare della scelta degli strumenti più adatti per effettuare la nostra presa dati. Per farlo abbiamo dovuto per prima cosa capire su quali V dei Big Data il nostro progetto sarebbe andato a focalizzarsi. Nel nostro caso è risultato subito evidente come fossero la V di Velocity e quella di Volume. In questo caso abbiamo deciso di utilizzare un database NoSQL quale MongoDB per lo storage dei dati. Il fatto che tramite le API di Youtube riuscissimo a scaricare i dati in formato JSON ci ha indotti ad utilizzare MongoDB che è pensato apposta per i database document based. Per rendere ottimizzato questo procedimento abbiamo usato due script python separati: *scraper_ consumer.py* e *scraper_ producer.py.* In particolzare abbiamo così effettuato una presa dati che può risultare efficace anche con una raccolta dati di volume maggiore. Il nostro oggetto producer si occupa infatti di scaricare i dati da Youtbue attraverso le API

Qualità dati

Mostrare le analisi di qualità effettuate sul dataset Esempio: c'erano missing values? se sì come abbiamo deciso di replicarli?

Scalabilità dell'algoritmo

Visto che una delle V che abbiamo deciso di usare concernenti i Big Data è stata quella relativa al volume dobbiamo occuparci di vedere come si comporta la nostra elaborazione dati con volumi di dati sempre crescenti. Per farlo ci occupiamo di effettuare la nostra analisi su partizionamenti sempre maggiori del nostro dataset e registriamo il tempo di esecuzione del nostro programma, effettuiamo poi una semplice regressione per vedere di che tipo di crescita stiamo parlando, ovviamente più la crescita è minore più il nostro algoritmo scala bene per volumi di dati maggiori

Visualizzazione

Inserire qua le domande di ricerca e come abbiamo pensato di rispondere a tali domande.

Scelta features

Inserire qua le scelte relative alle caratteristiche del nostro dataset da usare e i motivi che ci hanno spinto a usare proprio quelle .

Scelta della visualizzazione

Inserire qua le scelte relative alla visualizzazione proposta e perché abbiamo deciso di usare proprio quelle.

Valutazione della qualità

inserire le problematiche emerse durante la trasposizione delle infografiche e i metodi usati per risolverle.
Inserire i diagrammi emersi dalla valutazione della qualità con il questionario
Sottoporre il questionario ad una ventina di persone e valutare l'infografica in base alla scala Cabitza Locoro. Ricordiamo che è molto importante la dispersione dei dati, ricordo quindi di visualizzare i risultati in un box plot.