YouTube al tempo del Covid-19
Un'analisi dei video in tendenza

Gabriele Celeri, Federico Luzzi, Marco Peracchi, Christian Uccheddu

Introduzione e obiettivi

Negli ultimi mesi il Covid-19 ha avuto un impatto notevole sulle vite di tutti noi. L'obiettivo di questo lavoro è capire se la piattaforma YouTube sia stata influenzata dalla presenza di questo virus che ha costretto a rimanere nelle proprie case gran parte della popolazione mondiale. Per valutare l'impatto della pandemia e della quarantena obbligatoria, abbiamo confrontato i video in tendenza di Youtube nel periodo Dicembre-Gennaio rispetto Marzo-Maggio, con l'obiettivo di verificare se i contenuti presenti sulla piattaforma e la loro tipologia siano cambiati. Inoltre, ci siamo preposti, come secondo obiettivo, di verificare se l'andamento dei contagi, e delle notizie in merito, specialmente se negative, influenzassero in alcun modo la fruizione di video riguardanti il Covid-19.

Keyword: Covid-19, YouTube

Scelta degli strumenti

Per rispondere alle nostre domande di ricerca dobbiamo comprendere su quali delle tre "V" della Data Management avremmo dovuto concentrare i nostri sforzi. La conclusione a cui siamo giunti si focalizza sulla *Volume* e *Variety*. Il trattamento di grandi quantità di dati è risultato fondamentale, in quanto tramite le API di Youtube è stato possibile ricavare una discreta quantità di informazioni riguardanti i video in tendenza ($\sim 3Gb$). I dati riguardanti la pandemia sono stati ricavati in formato *csv*, di conseguenza l'integrazione con i dati di Youtube, in formato *json*, è stata fondamentale.

Per la gestione dei dati da Youtube ci siamo affidati al software MongoDB, sul quale sono stati caricati i dati mediante script python. Le API di Youtube non permettevano di scegliere il periodo, ma fornivano i dati in tempo reale, per questo abbiamo utilizzato Apache Kafka per avere i dati salvati per poi poterli caricare su MongoDB.

Raccolta dati

La raccolta dati è basata su due fonti principali: YouTube e OurWorldInData della Oxford University.

Youtube API

I video in tendenza su Youtube variano ogni 15 minuti circa (fonte: `https://support.google.com/youtube/answer/7239739?hl=it`). Tuttavia questo non significa che cambino effettivamente i contenuti presenti, infatti solitamente si assiste solo a qualche video che scompare dalle tendenze o viceversa nuovi video che entrano. Va specificato inoltre che il numero assoluto di video in tendenza è di circa 200, con qualche calo durante la notte più rilevante.

Le API di Youtube ci hanno permesso di raccogliere i dati necessari riguardanti i video in tendenza, con delle limitazioni sul numero di richieste gratuite che si potessero fare mediante il Google Developer.
Sono state affrontate due sessioni di *scraping* con metodi leggermente differenti:

1. dal 23 dicembre 2019 al 5 gennaio 2020 (richieste ogni 30 minuti)
2. dal 18 marzo 2020 al 6 maggio 2020 (richieste ogni 6 ore)

Abbiamo deciso di raccogliere dati delle tendenze dei seguenti paesi:

Italia, USA, Regno Unito, India, Germania, Canada, Francia, Corea del sud, Russia, Giappone, Brasile, Messico

Prima sessione

L'idea che ha guidato la fase iniziale del progetto era quella di comprendere come l'algoritmo delle tendenze di Youtube sceglie i video, utilizzando caratteristiche come visualizzazioni, likes, dislikes e commenti. La raccolta dati ha seguito il seguente algoritmo, messo in pratica dallo script *scraper_csv.py*:

Algorithm 1: Scraper csv

Data: country - insieme dei paesi prescelti; videos - insieme di video scaricati da un determinato paese

```
for every 30 minutes do
    foreach country do
        videos = APIrequest(max(50 video), country)
        while tendency videos are finished do
            videos += APIrequest(max(50 video), country)
        videos_fix = arrangeData(videos)
        saveToCsv(videos_fix)
        videos = ∅
```

Nota: per ogni richiesta API è possibile scaricare un massimo di 50 video, inoltre ogni paese ha un numero di video differente (solitamente 150-200).

I dati vengono salvati in formato *csv* secondo il seguente schema:

Attributo	**Descrizione**
timestamp	data, ora e minuto della nostra rilevazione
video_id	identificativo unico del video
title	nome del video per esteso
publishedAt	data di pubblicazione
channelId	identificativo unico del canale che ha pubblicato il video
channelTitle	nome del canale per esteso
categoryId	identificativo unico della categoria
trending_date	data in cui il video è in tendenza
tags	stringa contenente i tag usati, separati dal carattere "|"
view_count	numero di visualizzazioni
likes	numero di like (mi piace)
dislikes	numero di dislike (non mi piace)
comment_count	numero di commenti sotto il video
thumbnail_link	url all'immagine di copertina del video
comments_disabled	booleano che dichiara se i commenti sono disabilitati
ratings_disabled	booleano che dichiara se i like/dislike sono disabilitati
description	descrizione del video

Tabella 1: schema degli attributi dei dati csv

I dati così raccolti sono salvati in una cartella che può essere definita al momento dell'avvio dello script mediante l'argomento *-o*.

Seconda sessione

Lo scoppio della pandemia da Covid-19 ci ha permesso di cambiare approccio e domande di ricerca, cercando di valutare come la pandemia abbia influenzato Youtube. Con l'esperienza della presa dati precedente abbiamo cambiato metodo di raccolta, preferendo immagazzinare i dati direttamente, attraverso una pipeline, in un database MongoDB.

Inoltre abbiamo effettuato cambiamenti all'algoritmo di scraping, preferendo effettuare le richieste API ogni sei ore, per un totale di quattro rilevazioni al giorno, invece che ogni 30 minuti come prevedeva precedentemente lo schema.

A scopo didattico abbiamo deciso di implementare una piccola pipeline Kafka in cui dividiamo la fase di raccolta dati (*scraper_producer.py*) e la fase di immagazzinamento (*scraper_consumer.py*).

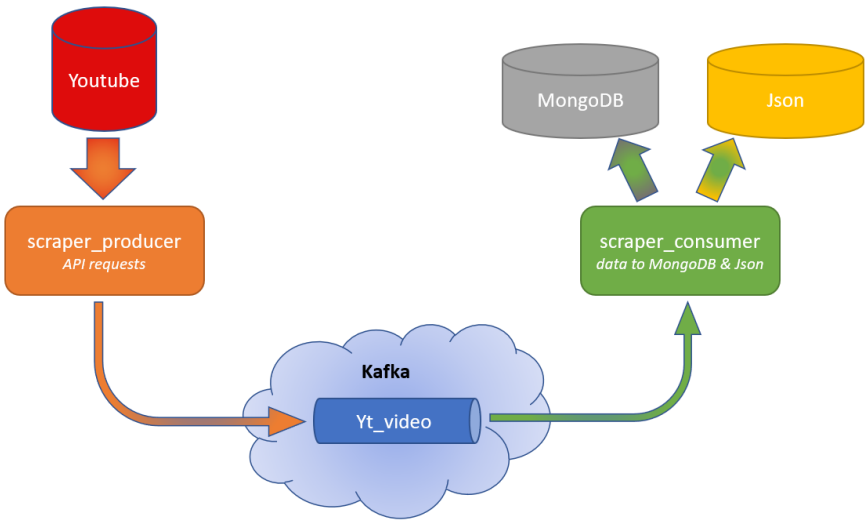

Figura 1: pipeline di raccolta dati - seconda sessione

Di seguito si può vedere nel dettaglio il procedimento dei due algoritmi:

Algorithm 2: scraper producer

Data: country - insieme dei paesi prescelti; videos - insieme di video scaricati da un determinato paese; KafkaProducer(data, channel) - sends data to channel kafka

```
1 for every 6 hours do
2     foreach country do
3         videos = APIrequest(max(50 video), country)
4         KafkaProducer(videos, yt_video)
5         while tendency videos are finished do
6             videos = APIrequest(max(50 video), country)
7             KafkaProducer(videos, yt_video)
```

Algorithm 3: scraper consumer

Data: KafkaConsumer(channel) - receives data from channel kafka

```
1 while loop do
2     if KafkaConsumer(yt_video) ≠ ∅ then
3         videos = KafkaConsumer(yt_video)
4         videos_fix = arrangeData(videos)
5         saveToJson(videos_fix)
6         saveToMongoDB(videos_fix)
```

I dati così raccolti sono stati salvati in formato json, è possibile definire la cartella di output in cui vengono salvati i file mediante il comando *-o*.

Nota: I dati sono stati salvati in json per avere una maggior facilità nel caricare i dati su MongoDB. Il caricamento è stato eseguito mediante lo script *json_to_mongo.py*

Lo schema logico dei documenti json è sostanzialmente invariato rispetto a quello della prima sessione, eccetto per due variazioni:

- **tags** immagazzinati come array di stringhe.

  Ad esempio:

  tags: ["covid-19", "quarantena"]

- informazioni relative ai likes, dislikes, view_count e comment_count come documenti innestati nella chiave **statistics**.

  Ad esempio:

  statistics: {view_count: 14235, likes: 513, dislikes: 34, comment_count: 254 }

Dati Covid-19

I dati relativi alla pandemia, come il numero di contagi totali o giornalieri, sono stati ottenuti dal sito OurWorldInData dell'università di Oxford. Successivamente abbiamo eseguito una prima manipolazione per ottenere i dati che interessavano i paesi e le date che abbiamo considerato. I risultati vengono visualizzati in un file *csv*:

Attributo	**Descrizione**
iso_code	codice unico riferito al paese
location	nome esteso del paese
date	data della rilevazione
total_cases	numero totale dei casi
new_cases	nuovi casi registrati in quella giornata.

Tabella 2: Schema degli attributi dei dati sul Covid-19.

Qualità dati

La grande disponibilità di dati ha rappresentato il problema più rilevante della verifica di qualità. Fortunatamente i dati non presentavano problemi di *missing values* che avrebbero rappresentato difficoltà non trascurabili. Le principali problematiche emerse sono due:

- Ridondanza: I dati di Dicembre-Gennaio presentano richieste effettuate ai server di Youtube ogni mezz'ora, a differenza del periodo successivo. Di conseguenza sono stati rilevati molti dati simili tra loro, senza alcuna sostanziale variazione nelle varie fasce delle giornata. Per risolvere il problema abbiamo optato per scegliere quattro rilevazioni distaccate di sei ore ciascuna, in maniera da uniformare i dati con quelli di marzo-maggio. Pe ulteriori dettagli è possibile visualizzare il notebook jupyter con cui è stato affrontato il problema.

- Saturazione richieste: Google non permette di effettuare troppe richieste nella stessa giornata. Per arginare il problema abbiamo utilizzato tre chiavi differenti da alternare durante la presa dati. Nonostante questo espediente ci sono stati alcuni momenti dove non è stato possibile effettuare le richieste ai server. Come risoluzione di questo problema abbiamo duplicato i dati della richiesta precedente, poiché i dati così ravvicinati non presentavano effettive variazioni.

Integrazione dati

Per poter rispondere alle nostre domande di ricerca abbiamo dovuto effettuare un'integrazione tra i dati dei video di Youtube e i dati relativi alla pandemia da Covid-19. L'integrazione è avvenuta prima del caricamento dei dati su MongoDB, lo script che mostra l'operazione è *merge_to_mongo.py*.

Per ricavare le informazioni relative all'andamento della pandemia nella giornata considerata abbiamo effettuato un'**integrazione temporale**, dove le chiavi considerate sono state il **timestamp** e il **country_name**, cioè la data e il paese. Il procedimento è il seguente: viene ricercata la data del video considerato e il paese di appartenenza, e successivamente viene creato un dizionario con tutte le informazioni della pandemia nella data e paese appena cercato. Infine viene creata una nuova chiave *covid* contenente un nuovo documento innestato, che è il dizionario creato precedentemente. Solo a questo punto i documenti vengono caricati su MongoDB.

Alcune date presentano un fuso orario che non ha alcun riscontro nel dataset covid, perché alcuni paesi presentano diverse fusi orari. Come risoluzione sono stati spostati tutti gli orari in base al fuso orario della capitale del paese di appartenenza del video. Questa correzione è stata effettuata all'interno dello script *scraper_consumer* direttamente. I fusi orari utilizzati sono presenti all'interno del file *country_name.json*.

L'operazione di integrazione complessivamente richiede:

- senza sharding: 1419 s (23,6 min circa)

- con sharding: 1431 s (24 min circa)

Espressione regolare

Per poter distinguere quali video possano essere considerati legati al Covid-19 o meno, abbiamo definito la seguente espressione regolare:

```
/(corona|covid|virus|pandemi[aec]|epidemi[aec]|tampon[ei]*|sierologico|mascherin[ae]|코로나 바
이러스|fase\s*(2|due)|iorestoacasa|stayathome|lockdown|[qc]uar[ae]nt[äae]i*n
[ea]|कोरोनावाइरस|वैश्विक महामारी|सर्वव्यापी महामारी|महामारी|пандемия|коронавирус|social\s*distancing|감염병 세계적 유행
|パンデミック|コロナウイルス|सर्वव्यापी महामारी|सर्वव्यापी महामारी|пандемия|коронавирус|social\s*distancing|
distanciamiento\s*social|코로나|코비드|바이러스|vaccin[oe]*|isolamento|intensiv[ao]|assembrament[io]|
guant[oi]|dpi|disinfettante|swabs|emergenza|emergency|droplets*|aerosol|isolation|intensive
\s*care|crowd|gloves*|disinfectant|감염병 유행|완충기|마스크|나는 집에있어|폐쇄|사회적 거리두기|백신|모임|비상 사
태|비말|봄 혈증|écouvillon|masques*|restealamaison|confin[ae]mento*|distanciation\s*sociale|
soins\s*intensifs|rassemblements|désinfectant|urgence|gouttelettes|飛沫|タンポン|マスケリン|封鎖|
人混みを避ける|ワクチン|隔離|集会|集中治療|緊急|बूंदें|फाहे|मास्क|लॉकडाउन|सोशल डिस्टन्सिंग|टीका|गहन देखभाल|समारोहों|आपातकालीन|
gotas|cotonetes|m[áa]scaras|ficoemcasa|vac[iu]na|reuni[õo]n*es|emerg[êe]ncia|капли|тампоны|
маски|карантин|социальное\s*дистанцирование|вакцина|интенсивная\s*терапия|сходы|
чрезвычайное\s*происшествие|hisopos|mequedoencasa|cierre|Tröpfchen|Tupfer|Masken|
bleibezuHause|Ausgangssperre|soziale\s*Distanzierung|Impfstoff|Intensivstation|
Versammlungen|Notfall|건강\s*격리|検疫|संगरोध|кК]арантин|hand|man[io]s*|소유|手|[Pp]уки|Hände|
mãos)/i
```

Come è possibile vedere, si è cercato di includere tutte le parole relative alla pandemia, e la loro traduzione nelle lingue di tutti i paesi che abbiamo considerato.

L'espressione regolare è stata applicata sia ai titoli dei video, sia ai tags scelti per descrivere il video. Di seguito le query applicate:

1. Vengono create due nuove variabili che identificano se nel video sono presenti riferimenti al Covid-19 o meno, una per il titolo e una per i tags. Questa variabile viene inizializzata come *false*:

   db.video_merge.update({},{$set : {covid_tags : false, covid_title : false}},{multi : true})

2. I video vengono analizzati singolarmente, e se l'espressione regolare (REGEX) restituisce un match positivo la variabile viene modificata in *true*, prima per i tags:

   db.video_merge.update({tags : {$in : [REGEX]}}, {$set : {covid_tags: true}}, {multi : true})

3. Successivamente per il titolo:

   db.video_merge.update({title : {$in : [REGEX]}}, {$set : {covid_title: true}}, {multi : true})

Scalabilità dell'algoritmo

Una delle V su cui è stata posta la nostra attenzione è la Volume, ovvero come trattare e gestire grandi quantità di dati. Avendo a disposizione il gestore MongoDB abbiamo deciso di implementare il metodo dello **Sharding**. Sono stati costruiti tre shard, tutti in modalità replica set per garantire ridondanza in caso di guasti e frammentazione per rendere le query più efficienti nel momento in cui si andava a interrogare il *router mongos*. I *config server* sono stati anch'essi configurati come replica set. Come chiave di sharding è stata scelta il campo **country_name**, perché le query per rispondere alle nostre domande vengono fatte sul singolo paese. Questo approccio è stato pensato anche nell'ottica di dividere la grande mole di dati in server disposti in ciascun paese.

Lo schema logico applicato è il seguente:

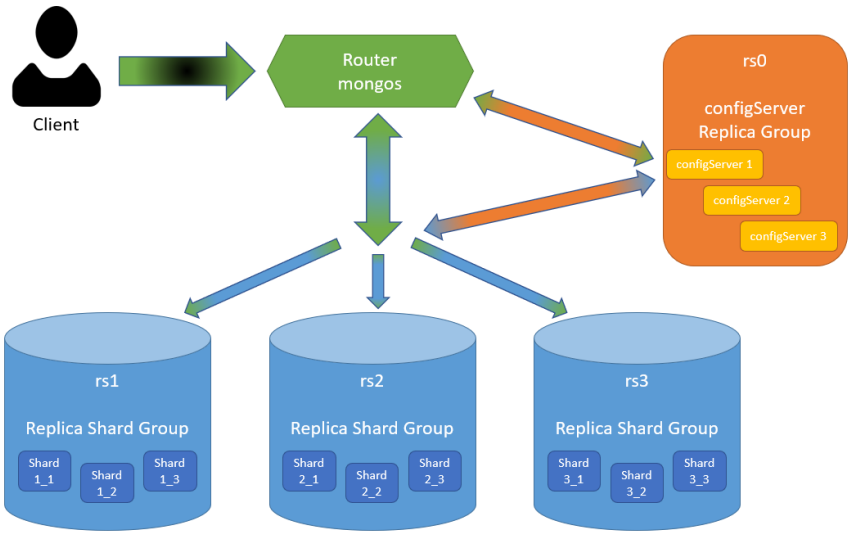

Figura 2: Pipeline sharding

Query

Abbiamo eseguito alcune query per testare il funzionamento del database con e senza sharding. In particolare nella cartella *query* si nota come la query 4, 5, 6 presentino dei risultati interessanti.

In particolare:

- *query 4*: elenco video a tema covid della categoria intrattenimento d'Italia
- *query 5*: elenco video a tema covid della categoria intrattenimento e musica in Russia
- *query 6*: elenco video a tema covid della categoria news & politics in Italia e Germania

Per quanto riguarda tempo di esecuzione in millisecondi(*executionTimeMillis*):

Query	**No Sharding**	**Sharding**
Query 4	635	126
Query 5	802	199
Query 6	1119	538

Tabella 3: differenze tempi in millisecondi

Riguardo al numero di documenti esplorati per ottenere la risposta:

Query	**No Sharding**	**Sharding**
Query 4	444207	36950
Query 5	444207	28715
Query 6	444207	444207

Tabella 4: differenze numero di documenti esplorati

Si può notare come per la query 4 e 5 vengono esplorati un numero di video molto inferiori nel database con sharding rispetto a quello senza, questo si traduce in un miglioramento dei tempi di esecuzione. La query 6, poiché necessità di informazioni riguardanti tutti i paesi, esplora tutti i dati a prescindere degli shard, quindi non ha alcun miglioramento prestazionale.

Per il dettaglio vedere i file di risposta nella cartella *query*.

Visualizzazione

La scelta delle visualizzazioni è stata guidata dalle domande di ricerca che ci siamo posti:

- Come variano le tipologie dei video in tendenza dal periodo precedente al coronavirus alla quarantena?

- È vero che la fruizione di video su Youtube riguardanti il Covid-19 segue l'andamento dei dati sull'epidemia?

Scelta features

Per rispondere in modo coerente alle nostre domande di ricerca abbiamo deciso di concentrarci sulle seguenti features del nostro dataset.

- View Count
- Covid Title
- Covid Tags
- Trending Date
- Title
- Cases New

Scelta della visualizzazione

Abbiamo utilizzato due infografiche diverse per le domande di ricerca, poiché ci è sembrata incompatibile un'unica visualizzazione per entrambe.

Prima infografica La prima infografica consiste nella combinazione di due diverse visualizzazioni:

- Un lollipop chart temporale che rappresenta il numero video entrato in tendenza ogni giorno. Sottolineamo che l'asse orizzontale rappresente le date di dicembre-gennaio e di marzo-maggio, in accordo con la nostra raccolta dati.
- Un bubble chart, dove ogni bolla è un video, la sua grandezza rappresenta il numero di visualizzazioni e il colore la categoria di appartenenza.

La combinazione di queste due visualizzazioni permette di capire se il contenuto dei video entrati in tendenza nel periodo precedente al Covid-19 e durante la quarantena differiscano significativamente. Per farlo è sufficiente selezionare due giorni contemporaneamente e vedere il cambiamento nelle bolle. L'esplorazione di questa infografica avviene per passi guidati attraverso una storia, in modo da introdurre all'utente tutte le informazioni che questa visualizzazione può offrire. Di seguito una visione sommaria dell'infografica comprensiva dei contesti:

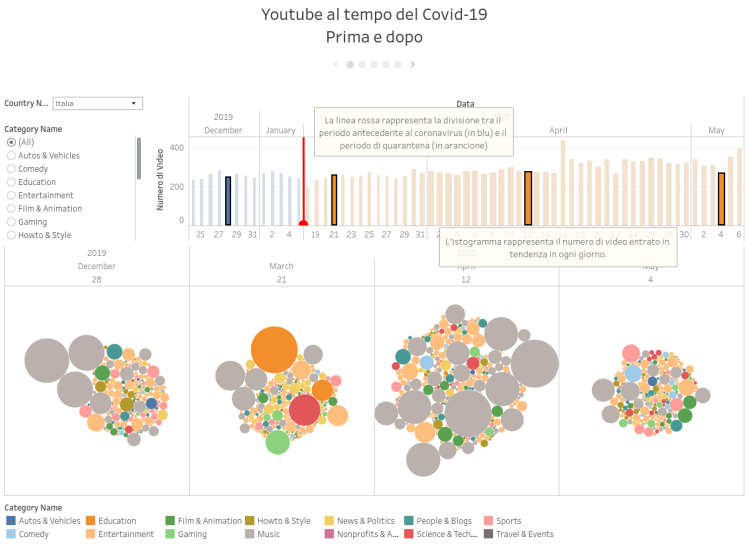

Figura 3: Prima infografica.

Seconda infografica Per quanto riguarda la risposta alla seconda domanda di ricerca abbiamo deciso di utilizzare una infografica composta da due visualizzazioni come precedentemente. In particolare:

- Un misto tra un bar chart e un line chart temporale. In questa visualizzazione abbiamo fatto risaltare la differenza percentuale tra una rilevazione e quella del giorno precedente. Il line chart riguarda l'aumento percentuale del numero di video in tendenza relativi al Covid-19 mentre il bar chart riguarda l'aumento percentuale dei nuovi casi di Covid-19 rispetto al giorno precedente. Abbiamo utilizzato questo tipo di grafico in modo da poter vedere se le notizie negative o positive dei dati riguardanti il Covid-19 abbia influito sulla fruizione online dei video concernenti lo stesso argomento. L'utilizzo di un bar chart e un line chart è dovuto al fatto che dai questionari è risultata più apprezzata questa scelta per riconoscere le due variabili.
- Uno stacked bar chart che rappresenta il numero di video per categoria che riguardano il Covid-19. La visualizzazione può essere filtrata per giorno semplicemente interagendo con la prima visualizzazione.

La combinazione di queste due visualizzazioni consente di rispondere alla seconda domanda di ricerca che ci siamo posti precedentemente. Proponiamo di seguito una visione sommaria dell'infografica comprensiva dei contesti.

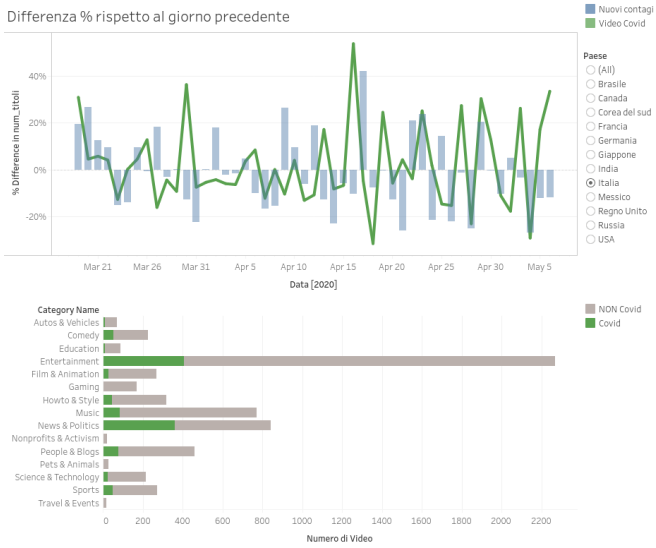

Figura 4: Seconda infografica.

Valutazione della qualità

La valutazione della qualità si è articolata in tre passaggi:

User Test Durante questa fase ci siamo occupati di sottoporre la nostra infografica a sei persone lasciando completa libertà di esplorazione. Le varie interazioni sono state registrate in modo da far sì che potessero emergere le diverse problematiche di cui non ci siamo accorti in fase di realizzazione delle infografiche. Esponiamo le problematiche emerse durante questa fase di valutazione e le correzioni applicate:

- *Problema 1:* Il fatto di avere due variabili sotto forma di linea nella prima infografica rende difficoltoso distinguerle nonostante il colore.
  Soluzione 1: Abbiamo deciso di assegnare ad una variabile la forma "linea" e all'altra variabile la forma "barra".
- *Problema 2: Come comprendere che cliccare sul bianco significa non avere nessun giorno selezionato, e quindi una visione complessiva.*
  Soluzione 2: Introdurre l'infografica mediante storie che possano guidare l'utente nell'esplorazione.
- *Problema 3: Come comprendere quali video sono nella categoria Covid e quali no.*
  Soluzione 3: Una legenda semplificata e l'utilizzo di bar chart sovrapposti di colori diversi.

Risultati dei task Durante questa fase ci siamo occupati di sottoporre tre diverse richieste per ogni infografica a 24 utenti; questi task devono essere soddisfatti esplorando in maniera interattiva. I task da risolvere sono stati i seguenti:

1. Per la *prima infografica:*
    - Nella giornata del 20 marzo quale video ha avuto più visualizzazioni e a quale categoria appartiene? *Science & technology, IPad Pro*
    - Della categoria Entertainment quali sono i canali che hanno fatto più visualizzazioni a Natale e a Pasqua? *The Late Show with Corbin, Mr Beast*
    - Nella categoria education confronta il 29 dicembre e 21 marzo in Germania. Qual è il titolo dei video più visti per ciascun giorno? *How to move the Sun, Recognizing Ignaz*
2. Per la *seconda infografica:*
    - Trova la categoria che ha avuto più video Covid-19 il giorno 27 Marzo in Italia? *News and Politics*
    - Quanti video Covid-19 della categoria "People and Blogs" ci sono stati negli USA? *12*
    - Quanti nuovi contagi ha avuto la Corea del Sud il 12 Aprile? *62*

Sono stati registrati i tempi in cui gli utenti riuscivano a completare questi obiettivi e sono stati visualizzati i risultati nei seguenti box plot: Questa visualizzazione è utile per capire se le nostre infografiche sono troppo dispersive oppure se riescono a essere facili e intuitive.

Questionari Per quanto riguarda l'ultima fase è stato somministrato un questionario di valutazione della qualità a 24 persone articolato nella seguente maniera:

- Come valuti la chiarezza dell' infografica?
- Come valuti l'utilità dell'infografica?
- Quanto valuti la bellezza dell'infografica?
- Come valuti l'intuitività dell'infografica?
- Quanto è stata informativa l'infografica?
- Come valuti complessivamente l'infografica?

Le risposte sono state registrate grazie al tool "Questionari di Google". Una volta registrati i risultati è stata controllata che la valutazione dell'infografica fosse coerente con una ricostruzione complessiva della regressione con i coefficienti di Cabitza-Locoro. L'utilizzo di un box plot è risultato molto comodo per la registrazione delle risposte poiché la media è un indicatore di tendenza centrale e non fornisce alcuna informazione sulla distribuzione di questi dati. Per quanto riguarda invece la coerenza della valutazione rispetto alla ricostruzione data dai coefficienti abbiamo avuto la seguente distribuzione: