

Guida operativa

Federico Luzzi, Marco Peracchi, Christian Uccheddu, Gabriele Centemeri (TTC)

Di seguito la guida operativa per eseguire il codice in modo da replicare i risultati ottenuti:

Presa dati Dicembre

Eseguire il file:

`scraper/scraper_csv.py`

Questo script serve ad effettuare una rilevazione dati ogni 6h. I dati così estratti vengono trasformati in csv. Trasformare i dati da csv in json usando:

`csv_to_json/main.py`

Per eseguire lo script è necessario fornire il seguente parametro:

- `-d "directory_dei_dati"`

Caricare quindi i json su mongo usando:

`json_to_mongo/json_to_mongo.py`

A questo script vanno forniti i seguenti parametri:

- `-d "directory dei dati"`
- `-u "utente mongo"`
- `-p "password utente"`
- `-port "porta in cui è attivo l'utente"`
- `-db "Nome del database in output"`
- `-c "collection in cui vengono inseriti i dati"`

Presa dati periodo Covid

Aprire il servizio mongo da terminale.

Lanciare in due terminali contemporaneamente:

- `scraper/scraper_consumer.py`
- `scraper/scraper_producer.py`

Questo effettua una rilevazione dati ogni 6h attraverso il servizio kafka. L'utilizzo di kafka non è indispensabile, inizialmente però avevamo deciso di prendere i dati sia dei canali che dei video in live stream quindi kafka aveva senso in quanto venivano usati due topic diversi e fatte le prime operazioni preliminari. Abbiamo deciso di tenerlo per non stravolgere la pipeline di esecuzione.

Presa dati Covid

Scaricare i dati in formato csv da questo sito

Eseguire il codice:

`covid/cleaner.py`

Questo script permette di eseguire una pulizia dei dati in modo da renderli integrabili con i json raccolti in precedenza.

Integrazione dei dati

Per eseguire l'integrazione tra i dati covid e i dati di youtube bisogna eseguire il seguente script:

clean_store_data/merge_to_mongo.py

A questo script vanno forniti i seguenti parametri:

- -d "directory dei dati"
- -u "utente mongo"
- -p "password utente"
- -port "porta in cui è attivo l'utente"
- -db "Nome del database in output"
- -c "collection in cui vengono inseriti i dati"

Questo script integra i due dataset e carica tutto su mongo.

Query mongo

Per le visualizzazioni che intendiamo fare abbiamo bisogno di poter distinguere quando un video contenga nel titolo o nei tag una delle parole che si rifanno al coronavirus. Per controllare questo è stata costruita la seguente espressione regolare:

```
/(corona|covid|virus|pandemi[aec]|epidemi[aec]|tampon[ei]*/sierologico/mascherin[ae]/  
코로나 바이러스/fase\s*(2|due)|iorestoacasa|stayathome|lockdown|[qc]uar[ae]nt[äae]i*n[ea]/  
कोरोनाविरस/वेरेनहाइम/massisolation/distanziamento\s*sociale|social\s*distancing/감염병  
세계적 유행/パンデミック/コロナウイルス/सर्वव्यापी महामारी/मरघटिआपी मरभारी/пандемия/  
коронавирус/social\s*distancing|distanciamento\s*social/코로나/코비드/वेहड/vaccin[oe]*/  
isolamento|intensiv[ao]|assemblament[io]|quant[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 사태|비밀|범 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|フ  
クチン|隔離|集会|集中治療|緊急|बूँदें|फाहे|मास्क|लॉकडाउन|सोशल डिस्टन्सिंग|टीका|गहन देखभाल|  
समारोहों|आपातकालीन|gotas|cotonetes|m[äa]scaras|ficioemcasa|vac[iu]na|reuni[öo]n*es/emerg  
[ëe]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|檢疫|संगरोध|[kK]арантин)/i
```

Figura 1: Regular expression usata

Creare due nuovi campi chiamati **covid_title** e **covid_tags** settati entrambi a **False**

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

Eseguire le seguenti due query che controllano se l'espressione regolare è presente nel campo title o in uno dei tag per ogni video.

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

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