MyTedx

HOMEWORK 2

DIEGO ROSSI MAT:1073945

GIORGIO CORNA MAT: 1074241

SERVIZIO AGGIUNTIVO WATCH NEXT

Dopo la visione del video giornaliero verrà condivisa con l'utente una playlist di video collegati al video appena visto.

La playlist sarà formata attraverso i collegamenti all'interno del file watch_next.csv

In questo modo l'utente dopo aver video il video giornaliero, avrà anche la possibilità di guardare altri video e di sceglierli direttamente dalla playlist fornita.

CRITICITA’

- Dato che il video giornaliero rispetta determinati vincoli impostati tramite i filtri, i video collegati nella playlist possono non rispettare questi vincoli.

- I video della playlist potrebbero non essere interessanti per l’utente.

JOB

Il dataset watch_next conteneva molti record uguali e per ogni talk era presente un record con un url non valido.

https://www.ted.com/session/new?context=ted.www%2Fwatch-later

Perciò prima di aggiungere il set al database, abbiamo fatto una pulizia del set rimuovendo tutti i record doppi e quelli non validi.

JOB

```
# READ WATCH_NEXT DATASET
watch_next_dataset_path = "s3://unibg-tedx-data-2023-homework/watch_next_dataset.csv"
watch_next_dataset_raw = spark.read.option("header","true").csv(watch_next_dataset_path)

# DATA CLEANING
watch_next_dataset = watch_next_dataset_raw.drop_duplicates() \
.where('url != "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"')

# ADD WATCH_NEXT TO AGGREGATE MODEL
watch_next_dataset_agg = watch_next_dataset.groupBy(col("idx").alias("idx_ref_watch_next")) \
.agg(collect_list("watch_next_idx").alias("watch_next"))
tedx_dataset_agg = tedx_dataset_agg.join(watch_next_dataset_agg,
tedx_dataset_agg._id == watch_next_dataset_agg.idx_ref_watch_next,
"left").drop("idx_ref_watch_next")
```

MONGO DB


```
  _id: "8d2005ec35280deb6a438dc87b225f89"
  main_speaker: "Alexandra Auer"
  title: "The intangible effects of walls"
  details: "More barriers exist now than at the end of World War II, says designer…"
  posted: "Posted Apr 2020"
  url: "https://www.ted.com/talks/alexandra_auer_the_intangible_effects_of_wal…"
▸ tags: Array
▾ watch_next: Array
    0: "5bd34fcc55d9e1267f605fa0c060d54e"
    1: "d9896b41b372ec60cdd3c662e57caad3"
    2: "078766d6cc461cf71d45dc268b66db95"
    3: "fe35edd737282ab3a325f2387cf1b50b"
```


PREVIOUS

1-20 of many results

NEXT

DURATION

Dopo aver collegato i video suggeriti all'interno di Mongo DB, abbiamo deciso di inserire anche la durata dei vari video, dato che è uno dei filtri per la ricerca del video giornaliero.

Per semplicità, la durata dei video l'abbiamo ricavato da un dataset trovato online.

Link : https://www.kaggle.com/datasets/jeniagerasimov/ted-talks-info-dataset

JOB DURATION

```
duration_dataset_path="s3://unibg-tedx-data-2023-homework/new/talks_info.csv"
duration_dataset_raw= spark.read.option("header","true").csv(duration_dataset_path)
duration_dataset_raw=duration_dataset_raw.drop('event','likes','published_date','recorded_date','related_videos','speakers','subtitle_languages','summary','title','topics','transcript','views','youtube_video_code')
duration_dataset_raw=duration_dataset_raw.drop("_id")
duration_dataset_raw=duration_dataset_raw.drop_duplicates()
tedx_dataset_agg=tedx_dataset_agg.join(duration_dataset_raw,
tedx_dataset_agg.url == duration_dataset_raw.page_url,"left").drop("page_url")
```

Mongo DB


```
_id: "8d2005ec35280deb6a438dc87b225f89"
main_speaker: "Alexandra Auer"
title: "The intangible effects of walls"
details: "More barriers exist now than at the end of World War II, says designer…"
posted: "Posted Apr 2020"
url: "https://www.ted.com/talks/alexandra_auer_the_intangible_effects_of_wal…"
▸ tags: Array
▸ watch_next: Array
duration: "698"
```

FINE