Salvetti Davide – 1057596
Verzeroli Matteo – 1057926

TikTEDx

Link a GitHub:
https://github.com/davidesalvetti-unibg/unibg_cloud_e_mobile_2020/tree/main/Homework_2

TikTEDx

Watch Next

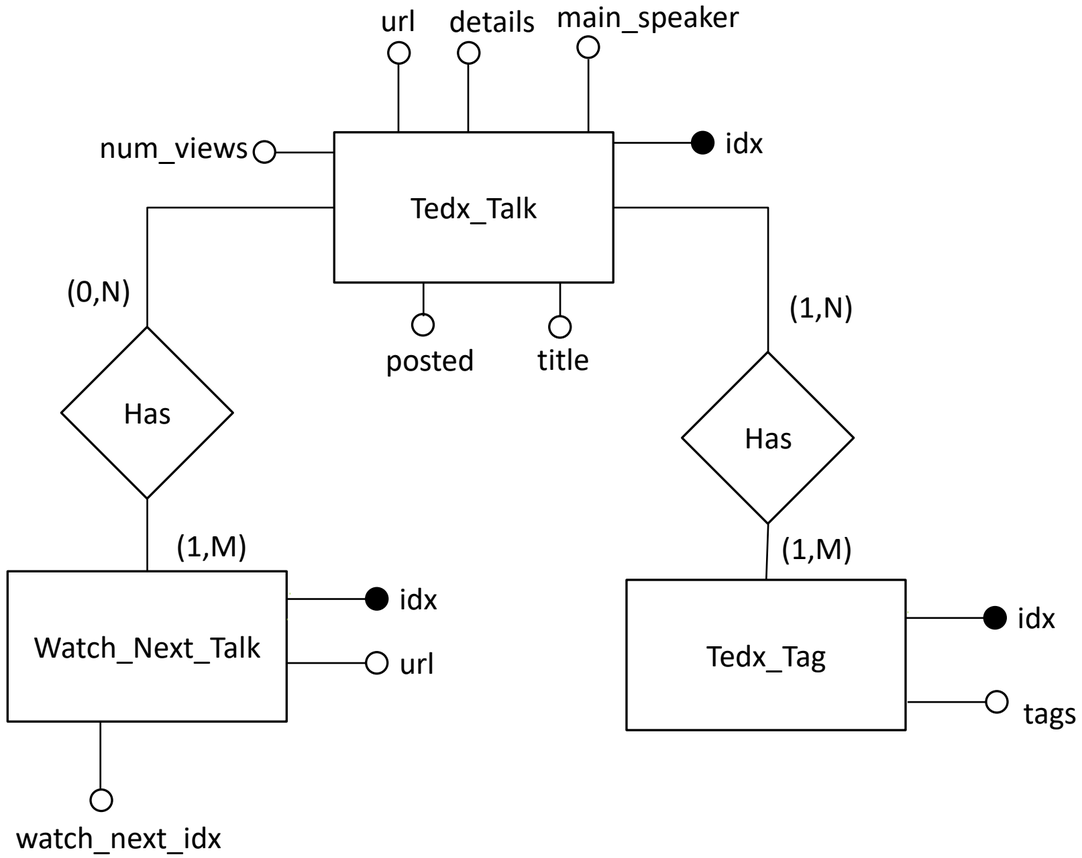
Abbiamo aggiunto il codice in grado di estrarre i watch_next dal dataset "watch_next_dataset.csv", raggrupparli in base alla colonna “idx” e fare il join con i dati estratti da "tedx_dataset.csv".
Abbiamo deciso di inserire anche tutte le informazioni dei ted talks disponibili nel dataset "tedx_dataset.csv" ("main_speaker", "title", "details") nell’oggetto watch_next e non solo gli id dei talk ‘watch next’. In questo modo abbiamo a disposizione direttamente tutte le informazioni dei ted talks successivi, i quali verranno proposti all’utente subito dopo aver terminato la visualizzazione di un video.

```
# READ WATCH NEXT DATASET
watch_next_path  = "s3://unibg-data-2021-davide/watch_next_dataset.csv"
watch_next_dataset = spark.read.option("header","true").csv(watch_next_path)

watch_next_dataset.printSchema()

# FILTER OPERATIONS
print(f"Before deleting duplicates the rows are: {watch_next_dataset.count()}")
watch_next_dataset = watch_next_dataset.dropDuplicates()
print(f"After deleting duplicates the rows are: {watch_next_dataset.count()}")
watch_next_dataset = watch_next_dataset.filter('url LIKE "https://www.ted.com/talks/%"')
print(f"After deleting urls the rows are: {watch_next_dataset.count()}")

# ADD WATCH NEXT TO TEDX_DATASET
watch_next_dataset = watch_next_dataset.select(col("idx").alias("id_ref"), col("watch_next_idx"), col("url").alias("url_wn"))

watch_next_dataset = tedx_dataset.join(watch_next_dataset, tedx_dataset.idx == watch_next_dataset.watch_next_idx, "right") \
    .drop("idx") \
    .drop("url") \
    .drop("posted") \
    .drop("num_views") \

watch_next_dataset = watch_next_dataset.groupBy(col("id_ref")).agg(collect_list(struct("watch_next_idx", "url_wn", "main_speaker", "title", "details")).alias("watch_next_obj"))

ted_tags_wn = ted_tags.join(watch_next_dataset, ted_tags.idx == watch_next_dataset.id_ref, "left") \
    .drop("id_ref") \

ted_tags_wn.printSchema()
```

TikTEDx

Schema

Schema ER dei dati forniti nei file tags_dataset.csv, tedx_dataset.csv, watch_next_dataset.csv




```
_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_obj: Array
  ⌄ 0: Object
      watch_next_idx: "d9896b41b372ec60cdd3c662e57caad3"
      url_wn: "https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_..."
      main_speaker: "Julia Dhar"
      title: "How to disagree productively and find common ground"
      details: "Some days, it feels like the only thing we can agree on is that we can..."
  ⌄ 1: Object
      watch_next_idx: "8576654442b6633b1dc0eb48a989172a"
      url_wn: "https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vert..."
      main_speaker: "Alex Honnold"
      title: "How I climbed a 3,000-foot vertical cliff – without ropes"
      details: "Imagine being by yourself in the dead center of a 3,000-foot vertical ..."
  ⌄ 2: Object
      watch_next_idx: "5bd34fcc55d9e1267f605fa0c060d54e"
      url_wn: "https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimag..."
      main_speaker: "Ronald Rael"
      title: "An architect's subversive reimagining of the US-Mexico border wall"
      details: "What is a border? It's a line on a map, a place where cultures mix and..."
  ⌄ 3: Object
      watch_next_idx: "5134ae81a27c94354173f38e84289ad5"
      url_wn: "https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_b..."
      main_speaker: "Anna Heringer"
      title: "The warmth and wisdom of mud buildings"
      details: ""There are a lot of resources given by nature for free -- all we need ..."
  ⌄ 4: Object
      watch_next_idx: "078766d6cc461cf71d45dc268b66db95"
      url_wn: "https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_borde..."
      main_speaker: "Will Hurd"
      title: "A wall won't solve America's border problems"
      details: ""Building a 30-foot-high concrete structure from sea to shining sea is..."
  ⌄ 5: Object
      watch_next_idx: "fe35edd737282ab3a325f2387cf1b50b"
      url_wn: "https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_th..."
      main_speaker: "Megan Campisi and Pen-Pen Chen"
      title: "What makes the Great Wall of China so extraordinary"
      details: "The Great Wall of China is a 13,000-mile dragon of earth and stone tha..."
```

Question Dataset

Dato che TikTEDx ti offre la possiblità di rispondere a delle domande al termine della visione di un talk, abbiamo deciso di aggiungere il dataset “question_ted_dataset.csv” su S3.
Il dataset contiene i seguenti campi:

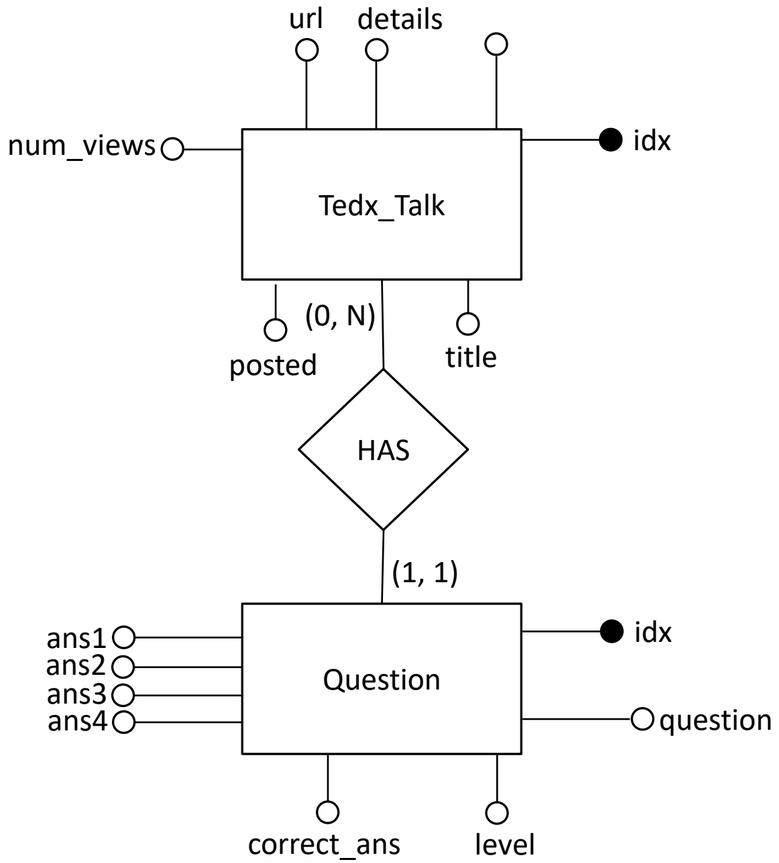
- idx: id del talk a cui fa riferimento la domanda
- question: la domanda vera e propria
- ans1, ans2, ans3, ans4: le possibili risposte che vengono mostrate agli utenti
- correct_ans: la risposta corretta
- level: il livello di difficoltà della domanda

```
# READ QUESTION DATASET
question_path  = "s3://unibg-data-2021-davide/question_ted_dataset.csv"
question_dataset = spark.read.option("header","true").csv(question_path)

question_dataset.printSchema()

# ADD QUESTION TO TED_X DATASET
question_dataset = question_dataset.groupBy(col("idx").alias("id_ref")).agg(collect_list(struct("question", "ans1","ans2","ans3","ans4","correct_ans","level")).alias("questions_obj"))

question_dataset.printSchema()

ted_tags_wn_question = ted_tags_wn.join(question_dataset, ted_tags_wn.idx == question_dataset.id_ref, "left") \
    .drop("id_ref") \
    .select(col("idx").alias("_id"), col("*")) \
    .drop("idx") \
```

TikTEDx

Schema

Schema ER dei dati forniti nel file question_ted_dataset.csv




```
_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_obj: Array
v questions_obj: Array
  v 0: Object
      question: "Whats's the speaker name?"
      ans1: "Alexandra Auer"
      ans2: "Sonia Shah"
      ans3: "Alex Gendler"
      ans4: "Noeline Kirabo"
      correct_ans: "Alexandra Auer"
      level: "3"
  v 1: Object
      question: "Whats's the topic?"
      ans1: "Windows"
      ans2: "Walls"
      ans3: "Mac"
      ans4: "Travels"
      correct_ans: "Walls"
      level: "2"
  v 2: Object
      question: "More barriers exist now than at the end of"
      ans1: "Worls War II"
      ans2: "World War I"
      ans3: "Cold War"
      ans4: "Other"
      correct_ans: "Worls War II"
      level: "4"
```

Criticità

1

Analizzando i dataset ci siamo accorti che i dati presentavano alcuni errori. La presenza di errori nei dataset comporta inconsistenza nei dati ed è da limitare. Abbiamo così introdotto alcune operazioni di filtraggio:

- Per il dataset "tedx_dataset.csv" abbiamo inserito il filtro `tedx_dataset.filter("idx NOT LIKE '%[ ]%'")`. In questo modo vengono rimosse tutte le righe che contengono degli spazi nel campo id. Infatti il campo idx corretto è solamente composto da una stringa alfanumerica.
- Per il dataset "tedx_dataset.csv" abbiamo inserito il filtro `tedx_dataset.filter("idx is not null")`. In questo modo vengono rimosse tutte le righe che contengono il valore null nel campo idx.
- Per il dataset "watch_next_dataset.csv" abbiamo inserito il filtro `watch_next_dataset.dropDuplicates()` per eliminare i duplicati.
- Per il dataset "watch_next_dataset.csv" abbiamo inserito il filtro `watch_next_dataset.filter('url LIKE "https://www.ted.com/talks/%"')` per eliminare eventuali righe che dei watch next che contenevano dei link invalidi.

2

Abbiamo inserito al caricamento del dataset "tedx_dataset.csv" l'opzione `option("multiline", "true").` Infatti abbiamo notato che all'interno della colonna details nel dataset "tedx_dataset.csv" era presente un record multilinea, che veniva erroneamente interpretato al momento del caricamento. L'opzione introdotta evita che eventuali record multiline vengano considerati come record differenti.

3

Velocità di sviluppo limitata. Il job impiega molto tempo per essere eseguito ed il tempo di attesa per verificare la correttezza dello script risulta essere troppo ampio. In fase di sviluppo sarebbe meglio utilizzare altre piattaforme che supportano PySpark.

Implementazioni future

1. Modifica del dataset delle domande per supportare la generazione automatica delle domande e la possibilità di inserire un numero variabile di risposte ad una domanda.

2. Possibilità di inserire domande di vario genere (vero o falso, risposta multipla, completamento della frase, domande aperte).

3. Automatizzare la personalizzazione dei watch next in base ai contenuti visti dall'utente e l'aggiornamento periodico dei watch next recuperati dal sito di Tedx.

4. Aggiungere ulteriori filtri per verificare la correttezza dei dataset, come ad esempio verificare l'effettivo funzionamento dei URL proposti