UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

MyTEDx

Filippo Barbieri - 1054060
Francesco Satriani - 1051361
Andrea Uberti - 1052992

Scopo

Il nostro scopo è stato quello di ottenere una nuova collezione di dati (tedz_wn) contenente un ulteriore array nel quale viene specificata la lista dei talks correlati al talk appena visto.

Procedimento

Per prima cosa abbiamo caricato sul bucket di S3 il file *watch_next_dataset.csv* contenente i talks correlati per ogni video sulla base dell'id.

Successivamente abbiamo creato un nuovo job che impostasse la nuova collezione di dati. Nello specifico vogliamo avere per ogni talk le seguenti informazioni :

- Id
- Presentatore
- Titolo
- Dettagli
- Data di pubblicazione
- Url
- Tags
- Url dei video correlati
- Numero di visualizzazioni

Job PySpark

I dati che abbiamo utilizzato sono: la lista dei talks, i tag e la lista dei watch next. Tramite una query ben strutturata siamo riusciti a creare una collezione di dati contenente i video descritti dai tag alla quale appartengono e i relativi video correlati (i watch next).

```
# Carico il file csv che contiene i watch next
wn_dataset_path = "s3://mytedx-data/watch_next_dataset.csv"
# Leggo il file
wn_dataset = spark.read.option("header","true").csv(wn_dataset_path)
# Rimozione dei duplicati
wn_dataset = wn_dataset.dropDuplicates()

# Seleziono dall'intero dataset solo la colonna dell'id del video attuale e l'URL del watch next
wn_dataset_agg = wn_dataset.select(col("idx"),col("url")).groupBy(col("idx").alias("idx_wn")).agg(collect_list("url").alias("watch_next"))

wn_dataset_agg.printSchema()

# Il dataset risultato sarà la join tra le tabelle dei tag, del dataset di tutti i video e dei watch next
tedx_dataset_agg = tedx_dataset.join(tags_dataset_agg, tedx_dataset.idx == tags_dataset_agg.idx_ref, "left") \
    .join(wn_dataset_agg, tedx_dataset.idx == wn_dataset_agg.idx_wn, "left") \
    .drop("idx_ref") \
    .drop("idx_wn") \
    .select(col("idx").alias("_id"), col("*")) \
    .drop("idx")
```

Risultato del Job


```
_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: "https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimag..."
   1: "https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimag..."
   2: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
   3: "https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_th..."
   4: "https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_th..."
   5: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
   6: "https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_..."
   7: "https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_..."
   8: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
   9: "https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_b..."
   10: "https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_b..."
   11: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
   12: "https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vert..."
   13: "https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vert..."
   14: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
   15: "https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_borde..."
   16: "https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_borde..."
   17: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
```


Presenza di duplicati

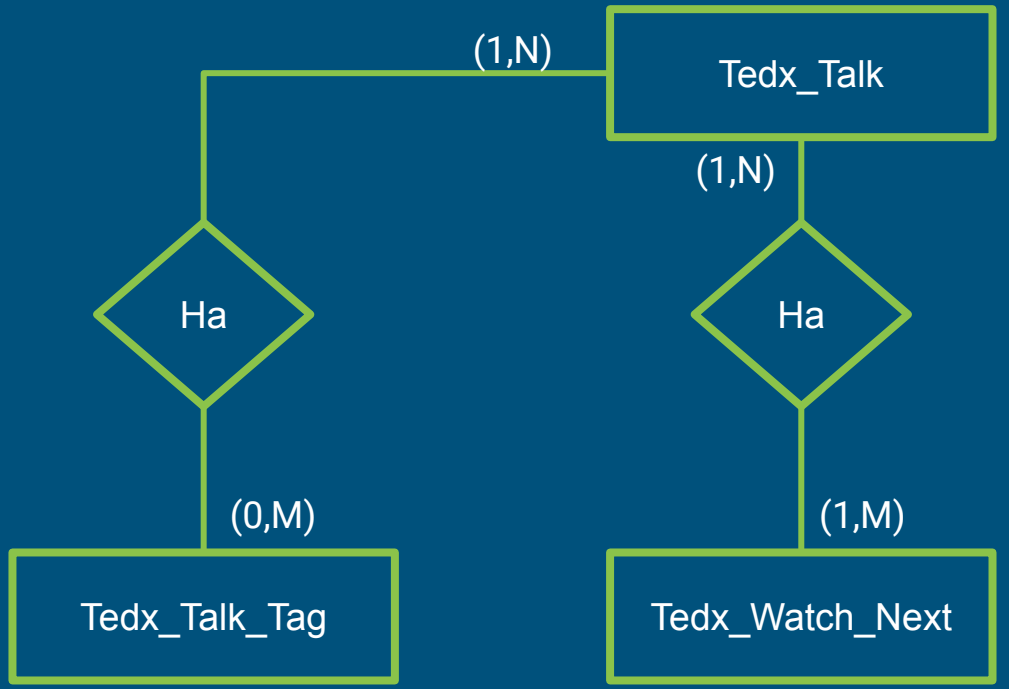
wn_dataset.dropDuplicates()


```
_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: "https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_th..."
   1: "https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vert..."
   2: "https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_b..."
   3: "https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimag..."
   4: "https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_..."
   5: "https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_borde..."
   6: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
```


Assenza di duplicati

Modello dei dati

- Un talk può avere 1 o N tag e un tag può appartenere a nessun talk o a M talks.
- Un talk può avere 1 o N watch next e un talk Watch Next può riferirsi ad uno o M talks



Incongruenze


```
_id: "Elisabeth est zythologue"
main_speaker: " une des quelques femmes experte en bière. Son travail et ses recherch..."
title: " chercheurs et cuisiniers ainsi que ses conférences et ses programmes ..."
details: "Posted Mar 2020"
posted: "https://www.ted.com/talks/elisabeth_pierre_l_histoire_inedite_des_femm..."
url: "0"
```


A sinistra si nota come sia tutto spostato verso l'alto, ovvero l'URL del video è nel campo 'posted' e la data è nel campo 'details' e così via.


```
_id: "4adc9fee977fa04c357ed4c9b52aa3cc"
main_speaker: "Butterscotch"
title: ""Accept Who I Am""
details: "Firing off her formidable beatboxing skills, musician Butterscotch ser..."
posted: "Posted Apr 2020"
url: "https://www.ted.com/talks/butterscotch_accept_who_i_am"
num_views: "0"
› tags: Array
⌄ watch_next: Array
    0: "https://www.ted.com/talks/tom_thum_and_matthew_broadhurst_what_happens..."
    1: "https://www.ted.com/talks/tom_thum_and_matthew_broadhurst_what_happens..."
    2: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
    3: "https://www.ted.com/talks/madame_gandhi_and_amber_galloway_gallego_mus..."
    4: "https://www.ted.com/talks/madame_gandhi_and_amber_galloway_gallego_mus..."
    5: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
    6: "https://www.ted.com/talks/a_tribe_called_red_we_are_the_halluci_nation"
    7: "https://www.ted.com/talks/a_tribe_called_red_we_are_the_halluci_nation"
    8: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
```


- Presenza di attributi scorretti nei records del db → possibile dataset contenente errori;
- Il view non è specificato per ogni video, quindi se non compare non viene inserito neanche il campo;
- Presenza di duplicati nell'array watch_next. Soluzione : wn_dataset = wn_dataset.dropDuplicates()

Criticità tecniche

Qualora verranno caricati nuovi video nel database si deve gestire l'aggiornamento automatico relativo alla lista di video correlati per ogni talk.
Il job implementato richiede diversi minuti prima di mostrare i risultati e aggiornare il db, questo potrebbe portare a problemi di tempistiche.

Possibili evoluzioni

Nelle prossime implementazioni abbiamo programmato di gestire l’aggiornamento automatico dei video
In aggiunta gestiremo anche la cronologia dei video visualizzati.