TEDx CS

Homework 2: link a GitHub

Allievi Giulia – 1058231

Fanton Martina – 1059640

TEDx dataset

- Prima di integrare il Watch Next dataset, si è aggiunta l'istruzione *.option("multiline", "true")* in modo da consentire a Spark di riconoscere dati su più righe come un unico dato, visto che alcune righe vengono erroneamente separate.

```
#### READ INPUT FILES TO CREATE AN INPUT DATASET
tedx_dataset = spark.read \
    .option("header","true") \
    .option("quote", "\"") \
    .option("escape", "\"") \
    .option("multiline", "true") \
    .csv(tedx_dataset_path)

tedx_dataset.printSchema()
```

Watch Next dataset

- Per l'integrazione in S3 dei dati relativi ai talk Watch Next con il dataset di partenza, è risultato necessario compiere delle operazioni preliminari di filtraggio: prima sono stati eliminati i dati duplicati con il comando *dropDuplicates()* e poi sono stati rimossi i dati caratterizzati da un URL non appartenente alla sezione dei talk di TEDx con il comando *filter('url LIKE "https://www.ted.com/talks/%"')*.
- Successivamente i tre campi contenuti nel Watch Next dataset (ID del talk, URL del talk e ID del talk Watch Next) sono stati aggregati in base al primo campo.
- Infine è stato eseguito un join del dataset *tedx_dataset_agg* con il dataset dei Watch Next appena creato, denominato *watchnext_dataset_agg*.

```python
## READ WATCHNEXT DATASET
watchnext_dataset_path = "s3://tcm-tedx-mf99-dati/watch_next_dataset.csv"
watchnext_dataset = spark.read.option("header","true").csv(watchnext_dataset_path)

## FILTER WATCHNEXT DATASET
watchnext_dataset = watchnext_dataset.dropDuplicates()
watchnext_dataset = watchnext_dataset.filter('url LIKE "https://www.ted.com/talks/%"')

## CREATE THE AGGREGATE MODEL, ADD WATCHNEXT DATASET
watchnext_dataset_agg = watchnext_dataset.groupBy(col("idx").alias("idx_ref_watchnext")) \
    .agg(collect_list("url").alias("url_watchnext"), collect_list("watch_next_idx") \
    .alias("watch_next_idx_watchnext"))
watchnext_dataset_agg.printSchema()
tedx_dataset_agg = tedx_dataset_agg.join(watchnext_dataset_agg, tedx_dataset_agg._id == \
    watchnext_dataset_agg.idx_ref_watchnext, "left") \
    .drop("idx_ref_watchnext")

tedx_dataset_agg.printSchema()
```

Codice aggiunto al job.


```
_id: "8c1fad5ce0dab8908dee527f88697ce2"
main_speaker: "Todd Dufresne"
title: "History vs. Sigmund Freud"
details: "Working in Vienna at the turn of the 20th century, he began his career…"
posted: "Posted Mar 2020"
url: "https://www.ted.com/talks/todd_dufresne_history_vs_sigmund_freud"
▾ tags: Array
    0: "TED"
    1: "talks"
    2: "education"
    3: "animation"
    4: "TED-Ed"
    5: "psychology"
    6: "history"
    7: "brain"
    8: "science"
▾ url_watchnext: Array
    0: "https://www.ted.com/talks/janja_lalich_why_do_people_join_cults"
    1: "https://www.ted.com/talks/alex_gendler_history_vs_christopher_columbus"
    2: "https://www.ted.com/talks/julia_galef_why_you_think_you_re_right_even_…"
    3: "https://www.ted.com/talks/mark_robinson_and_alex_gendler_history_vs_he…"
    4: "https://www.ted.com/talks/rick_doblin_the_future_of_psychedelic_assist…"
    5: "https://www.ted.com/talks/julia_shaw_a_memory_scientist_s_advice_on_re…"
▾ watch_next_idx_watchnext: Array
    0: "44996a194bdb3498a61f3cae3a5a1389"
    1: "43014f52663d1cf317a606b9c4dfe2fd"
    2: "e13408af724b61e69205333f09e16c6d"
    3: "9d7e0fbafa4023ef33c5b2e34984b823"
    4: "ba6ec23bd21321897864007e384b96f8"
    5: "bbc025842ca7c9dc5a5a74ab1dd19dd2"
```


Esempio di un oggetto della collezione.

TEDxCS dataset

- Visto che la nostra applicazione prevede di calcolare una percentuale che indica il grado di apprezzamento di un talk, abbiamo deciso di caricare in S3 un nuovo dataset, *tedx_cs_dataset.csv*.
- Questo dataset contiene due campi: il primo corrisponde all'ID del talk, mentre il secondo è relativo alla percentuale di gradimento iniziale del talk corrispondente.
- Per aggiungere la nuova informazione ai dati già presenti nella collezione, è sufficiente fare un join fra il dataset del modello aggregato, *tedx_dataset_agg*, e il dataset appena caricato, *cs_dataset*, sull'ID del talk.

```
## ADD PERCENTAGE

## READ CS DATASET
cs_dataset_path = "s3://tcm-tedx-mf99-dati/tedx_cs_dataset.csv"
cs_dataset = spark.read.option("header","true").csv(cs_dataset_path)

## CREATE THE AGGREGATE MODEL, ADD CS DATASET
tedx_dataset_agg = tedx_dataset_agg.join(cs_dataset, \
    tedx_dataset_agg._id == cs_dataset.idx, "left") \
    .drop("idx")

tedx_dataset_agg.printSchema()
```

Codice aggiunto al job.


```
_id: "8c1fad5ce0dab8908dee527f88697ce2"
main_speaker: "Todd Dufresne"
title: "History vs. Sigmund Freud"
details: "Working in Vienna at the turn of the 20th century, he began his career…"
posted: "Posted Mar 2020"
url: "https://www.ted.com/talks/todd_dufresne_history_vs_sigmund_freud"
▾ tags: Array
    0: "TED"
    1: "talks"
    2: "education"
    3: "animation"
    4: "TED-Ed"
    5: "psychology"
    6: "history"
    7: "brain"
    8: "science"
▾ url_watchnext: Array
    0: "https://www.ted.com/talks/mark_robinson_and_alex_gendler_history_vs_he…"
    1: "https://www.ted.com/talks/rick_doblin_the_future_of_psychedelic_assist…"
    2: "https://www.ted.com/talks/janja_lalich_why_do_people_join_cults"
    3: "https://www.ted.com/talks/julia_galef_why_you_think_you_re_right_even_…"
    4: "https://www.ted.com/talks/julia_shaw_a_memory_scientist_s_advice_on_re…"
    5: "https://www.ted.com/talks/alex_gendler_history_vs_christopher_columbus"
▾ watch_next_idx_watchnext: Array
    0: "9d7e0fbafa4023ef33c5b2e34984b823"
    1: "ba6ec23bd21321897864007e384b96f8"
    2: "44996a194bdb3498a61f3cae3a5a1389"
    3: "e13408af724b61e69205333f09e16c6d"
    4: "bbc025842ca7c9dc5a5a74ab1dd19dd2"
    5: "43014f52663d1cf317a606b9c4dfe2fd"
gradimento: "NaN"
```


Esempio di un oggetto della collezione.

Criticità

- Analizzando i dati sono stati notati alcuni aspetti critici (che sono stati risolti come spiegato nelle precedenti slide):
  1. La disposizione di alcuni dati su più righe;
  2. La presenza di duplicati nei dati relativi agli Watch Next;
  3. La presenza di alcuni URL degli Watch Next non conformi a quelli dei talk di TEDx.
- Nel dataset non è presente il numero di visualizzazioni dei talk. Questo dato lo si voleva utilizzare per definire la percentuale di gradimento iniziale. Di conseguenza si è deciso di inizializzare questa percentuale a NaN per tutti i talk.
- In AWS Glue l'esecuzione del job richiede del tempo e di conseguenza si ha uno sviluppo del codice piuttosto lento.

Sviluppi futuri

- Per inizializzare la percentuale di gradimento di un talk, si potrebbe sfruttare il dato relativo al numero di like, visto che il dato relativo al numero di visualizzazioni non è disponibile. Il numero di like si ottiene facendo scraping.
- Si potrebbe aggiungere, oltre che all'ID e all'URL dei Watch Next talk, anche il campo relativo alla percentuale di gradimento, così che nelle fasi successive di sviluppo risulti più facile ordinare i Watch Next talk sulla base della percentuale, in quanto questo dato risulta già disponibile.