
TEDxSCHOOL 2
AWS GLUE
Rota Leonardo 1086029
Bonomelli Pietro 1087035
1

Obiettivi

Analizzare i dati

Analisi e Coerenza del dataset per la sua preparazione e gestione.

Aggiungere funzionalità *Watch Next*

Implementare un sistema di raccomandazione per suggerire video correlati (basato su *related_videos.csv*).

Sistemi di filtro per materia

Implementare funzionalità dedicate ai Professori per la navigazione del catalogo tramite materie curricolari.

Pulizia dei Talk

Problematica

Caratteri di interruzione di riga (\n) in description (details.csv) causavano errori di parsing.

Collegamenti Inaffidabili:
L'ID suggerito in related_videos.csv spesso non corrisponde a un ID valido nella lista principale.

Correzione

Lettura con opzione multiLine: true e normalizzazione con regexp_replace.

Risoluzione tramite lookup (slug-to-ID) e scarto dei suggerimenti non validi.

```
# --- LOGGING PRE-PULIZIA ---
details_with_newline = details_dataset.filter(col("description").contains("\n"))
count_newline = details_with_newline.count()

print(f"WARNING: Trovate {count_newline} righe in details.csv che richiedono
pulizia (contengono '\\n').")
```

```
# OUTPUT
WARNING: Trovate 2,341 righe in details.csv che richiedono pulizia (contengono
'\n').
```

Esempio di Prima e Dopo

Prima

```
_id: "567935"
slug: "how_to_be_a_better_human_how_to_set_boundaries_and_find_peace_
speakers: "How to Be a Better Human"
title: "How to set boundaries and find peace (w/ Nedra Glover Tawwab
url: "https://www.ted.com/talks/how_to_be_a_better_human_how_ .. "
description: "Telling other people what you want — or need — can be a
duration: null
publishedAt: null
tags: Array (4)
    0: "relationships"
    1: "mental health"
    2: "boundaries"
    3: "therapy"
```

Dopo

```
_id: "567935"
slug: "how_to_be_a_better_human_how_to_set_boundaries_and_find_peace_w_nedra_g
speakers: "How to Be a Better Human"
title: "How to set boundaries and find peace (w/ Nedra Glover Tawwab)"
url: "https://www.ted.com/talks/how_to_be_a_better_human_how_ .. "
description: "Telling other people what you want — or need — can be a really d
duration: "2027"
publishedAt: "2025-03-03T17:51:067"
tags: Array (4)
    0: "relationships"
    1: "mental health"
    2: "boundaries"
    3: "therapy"
subjects: Array (1)
    0: "medicina"
related_videos_data: Array (4)
    0: Ojbect
        id: "542391"
        title: "How to develop the habits you want — and get rid of the ones y
        slug: "how_to_be_a_better_human_how_to_develop_the_habits_you_want_and
        url: "https://www.ted.com/talks/how_to_be_a_better_human_how_to_develo
        speaker: "How to Be a Better Human"
        image_url: "https://talkstar-assets.s3.amazonaws.com/production/ .."
    1: Object
        id: "565154"
        title: ".."
        ...
```

- **Abbiamo ripulito le descrizioni dai caratteri di controllo (\n, \r, \t) e unito i file CSV separati per recuperare metadati essenziali come durata e data di pubblicazione, inizialmente mancanti o mal formattati.**
- **Abbiamo tradotto i tag tecnici di TEDx in 14 materie curricolari. Questo permette all'app di offrire filtri immediati per i professori (es. Scienze, Storia, Geografia) eliminando la confusione dei tag originali.**
- **Per rendere l'app più veloce, abbiamo inserito titoli, URL e immagini dei video correlati direttamente dentro il documento principale. Questo evita all'app di dover fare decine di query separate, mostrando tutto il contenuto con un unico caricamento.**

Risoluzione dati e funzionalità “Watch Next”

- **Mappa di Lookup:** Abbiamo creato una tabella temporanea da *final_list.csv* che associa ogni slug al suo ID corretto.

- **Risoluzione:**
  - è stato eseguito un INNER JOIN tra *related_videos.csv* e la mappa di lookup, utilizzando lo slug come chiave di riferimento.
  - L'INNER JOIN scarta automaticamente tutti i suggerimenti video per i quali non è stato trovato uno slug corrispondente in *final_list.csv*, garantendo così l'integrità dei dati (il link suggerito esisterà sempre).
  - È stato sostituito l’ID originale (quello errato in *related_videos*) con l'ID corretto risolto dalla mappa (correct_id).

- **Aggregazione:** L'aggregazione finale (collect_list) ora crea un array related_videos contenente solo talk validi e verificati.

Scelta della chiave: Slug vs ID

```python
# ANALISI QUALITÀ RELATED_VIDEOS
valid_ids = tedx_dataset.select(col("id")).distinct()
valid_slugs = tedx_dataset.select(col("slug")).distinct()

# Conta ID validi nei suggerimenti
related_id_validi_count = related_videos_dataset.join(
    valid_ids,
    col("related_id") == col("valid_id"), "inner"
).count()

related_id_validi_pct = (related_id_validi_count / total_suggestions) * 100

# Conta Slug validi
slug_validi_count = related_videos_dataset.join(
    valid_slugs,
    col("slug") == col("valid_slug"), "inner"
).count()

slug_validi_pct = (slug_validi_count / total_suggestions) * 100
```

```
# Output
✓ Slug validi: 92% (8,432 suggerimenti)
✗ ID validi: 78% (molto più basso!)
→ Decisione: Usiamo gli SLUG come chiave di join!
```

Implementazione

```python
# 1 Mappa di Lookup
slug_id_map = final_list.select(
    col("slug").alias("slug_lookup"),
    col("id").alias("correct_id")
)

# 2 Risoluzione (INNER JOIN)
resolved_videos = related_videos.join(
    slug_id_map,
    related_videos.slug == slug_id_map.slug_lookup,
    "inner"  # ← Scarta non-validi
).select(col("id").alias("main_talk_id"),
         col("correct_id").alias("target_video_id"))

# 3 Aggregazione
related_videos_agg = resolved_videos.groupBy(
    "main_talk_id"
).agg(collect_list("target_video_id").alias("related_videos"))
```

Codice completo su GitHub

Criticità

Assegnazione tag manuale

L'aggiunta di nuovi tag da parte di TED richiede l'aggiornamento manuale della logica Spark. Senza questo intervento, i nuovi contenuti potrebbero finire nella materia generica "Interdisciplinare".

Dati non allineati

Avendo copiato i dati dei correlati dentro ogni documento, se un video cambia titolo o immagine, le "copie" presenti negli altri talk NON si aggiornano automaticamente, ma richiedendo un nuovo job di sincronizzazione.

Sovrapposizione delle categorie

Un video con troppi tag può apparire consigliato a molti talk, rischiando di creare "rumore" e mostrare contenuti poco pertinenti in sezioni scolastiche specifiche (o di mostrare sempre gli stessi video).

Possibili evoluzioni

Classificazione Automatica con IA

Sostituire la mappatura manuale dei tag con un modello di Natural Language Processing che assegni automaticamente i nuovi talk alle materie scolastiche, eliminando la manutenzione del codice.

Propagazione Didattica e Ranking

L’evoluzione del sistema di raccomandazione prevede l'introduzione di un meccanismo di valutazione dedicato ai docenti. Implementando una funzione di voto da 1 a 10, i professori potranno influenzare direttamente l’algoritmo del watch_next.

Classificazione per difficoltà

Permettere ai docenti di creare collezioni personalizzate su MongoDB, aggiungendo metadati come il livello di difficoltà (es. Medie vs Superiori) per adattare i contenuti all'età degli studenti in modo da trovare più velocemente video adatti alla fascia d’età desiderata.


GitHub
Trello
Rota Leonardo 1086029
Bonomelli Pietro 1087035
9