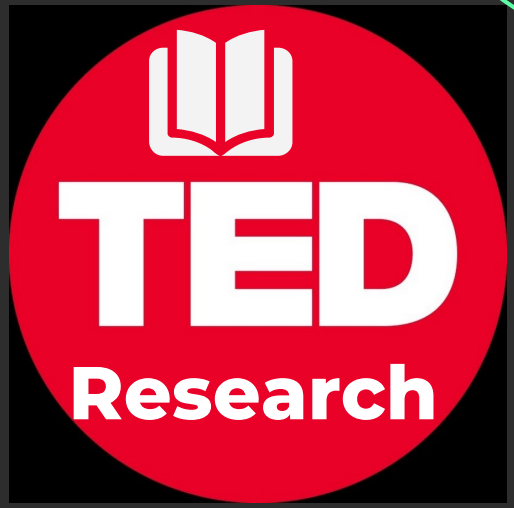
TedResearch

App Mobile a supporto della ricerca

Daniele Pendesini (1068726)
Camilla Mazzoleni (1072676)
Andrea Rota (1054128)



2 PySpark Jobs

Obiettivi

1. Aggiungere ad ogni talk la lista dei watch next suggeriti in modo che l'utente possa appofondire gli argomenti
2. Aggregare gli articoli scientifici relativi all'argomento tramite i tags del talk

Codice

Repository Github

PySpark Job

Dati trattati

- Tedx_dataset.csv
- Tags_dataset.csv
- Watch_next_dataset
- Dati da API: https://api.core.ac.uk/

Tecnologie utilizzate

- PySpark
- AWS Glue

Schema aggregato

```
_id: String,
main_speaker: String,
title: String,
url: String,
posted: String,
details: String,
main_author: String,
num_views: Number,
durations: String,
watch_nexts:[{
 url: String,
 watch_next_idx: String
}],
tags:[String],
papers: [{
 title: String,
 identifiers: [String]
}]
```

Problem vs. Solution

Necessità di aggiornare articoli

Uno degli scopi principali dell'applicazione è di allegare articoli scientifici e di attualità sotto ogni video per poter informare a aggiornare l'utente sulle tematiche

Aggiornamento automatico settimanale

Per fare ciò è necessario un job periodico che aggiorni la lista degli articoli da allegare al di sotto di ogni video, è stato deciso di schedularlo con frequenza settimanale

Problem vs. Solution

Raw data con duplicati

I dati dei dataset sono sporchi: vi sono dati duplicati e dati con chiave nulla

Filtraggio

```
tedx_dataset.filter("idx is not null")
watch_next_dataset.dropDuplicates()
```

È stato necessario implementare del codice per filtrare i dati duplicati e con chiave nulla per ottenere un dataset con cui sia più facile lavorare

Possibili evoluzioni

Job di raccomandazione personalizzata

Job di raccomandazione personalizzata che suggerisce agli utenti articoli scientifici e TEDx Talks

Raccomandazioni in base a:
Interessi
Visualizzazioni precedenti

Job di generazione automatica di abstract

Job per generare abstract sintetici per gli articoli scientifici

Abstract
Panoramica rapida del contenuto

Job di analisi delle tendenze

Job che monitora le tendenze degli articoli scientifici e dei TEDx Talks per identificare temi emergenti o popolari

Temi emergenti
Attualità
Ultime scoperte scientifiche