

# International TEDays

## HOMEWORK 2

Link utili:



Andrea Appiani - 1057683

Luca Mazzoleni - 1054271

Cristian Sacco - 1063539

Alexandr Rucodainii - 1045954

Nunzio Marco Bisceglia - 1046319



**JOB PYSPARK WATCH NEXT**

# JOB PYSPARK WATCH NEXT

## Dataset utilizzati:

- tedx
- tags
- watch\_next

## Il job inizia con:

- Import delle librerie e inizializzazione del job
- Lettura dei dati in formato .csv da piattaforma AWS S3
- Filtro dei TEDx Talks con ID nulli (4467 TEDx Talks validi trovati)

# JOB PYSPARK WATCH NEXT

**Il job prosegue con:** filtro dei watch next duplicati e degli URL non corretti

```
watch_next_dataset =  
spark.read.option("header", "true").csv(watch_next_dataset_path)  
count_items = watch_next_dataset.count()  
  
watch_next_dataset = watch_next_dataset.distinct().where('url !=  
"https://www.ted.com/session/new?context=ted.www%2Fwatch-later"')  
count_items_without_duplicates = watch_next_dataset.count()
```

Nei log è mostrato il risultato:

```
Number of items from RAW DATA 77364  
Number of items from RAW DATA without duplicates and wrong links 25788
```

# JOB PYSPARK WATCH NEXT

## Il job si conclude con:

- Raggruppamento dei tre dataset
- Connessione al database MongoDB e inserimento del dataset aggregato nella collection **tedx\_data**

```
watch_next_dataset_agg =  
watch_next_dataset.groupBy(col("idx").alias("idx_ref")).agg(collect_list("url").alias("watch_next"))  
watch_next_dataset_agg.printSchema()  
watch_next_dataset_agg = tedx_dataset.join(watch_next_dataset_agg,  
tedx_dataset.idx == watch_next_dataset_agg.idx_ref, "left") \  
    .drop("idx_ref") \  
    .select(col("idx").alias("_id"), col("*")) \  
    .drop("idx") \  
  
tags_dataset_agg =  
tags_dataset.groupBy(col("idx").alias("idx_ref")).agg(collect_list("tag").alias("tags"))  
tags_dataset_agg.printSchema()  
tedx_dataset_agg = watch_next_dataset_agg.join(tags_dataset_agg,  
watch_next_dataset_agg._id == tags_dataset_agg.idx_ref, "left") \  
    .drop("idx_ref") \  

```

# JOB PYSPARK WATCH NEXT

## Risultato:

Ad ogni TEDx Talk (idx) abbiamo associato gli array **tags** e **watch\_next**

```
_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..."
  ✓ watch_next: Array
    0: "https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_..."
    1: "https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_th..."
    2: "https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimag..."
    3: "https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_b..."
    4: "https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vert..."
    5: "https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_borde..."
  ✓ tags: Array
    0: "TED"
    1: "talks"
    2: "design"
    3: "society"
    4: "identity"
    5: "social change"
    6: "community"
    7: "humanity"
    8: "TEDx"
```

# CRITICITÀ JOB WATCH NEXT

Nonostante i filtri applicati nel dataset sono presenti alcuni record non validi dovuti a caratteri speciali come il “\n”. Si potrebbero effettuare altri controlli per rendere validi anche quei record.



# **JOB PYSPARK INTERNATIONAL DAYS**



# SCRAPER



## OBIETTIVO

Ottenere un dataset contenente la lista delle **giornate internazionali**, al fine, di mostrare all'utente i TEDx Talks relativi agli eventi celebrati e agli argomenti trattati durante le giornate internazionali.

Il match fra le giornate internazionali e i TEDx Talks verrà effettuato successivamente con una Lambda Function.



## TECNOLOGIE

Jupyter Notebook  
Selenium  
Chrome Driver

# SCRAPER



## BREVE DESCRIZIONE

- Setup dell'ambiente e installazione delle librerie
- Apertura della pagina web [List of International Days and Weeks | United Nations](#) tramite Chrome Driver
- Analisi del DOM della pagina per ottenere i dati tramite Selenium
- Creazione di un file .csv in cui salvare tutti i dati ottenuti



## FORMATO DATI

- **date:** giorno dell'anno in formato "gg-mm"
- **event:** nome della giornata internazionale
- **url:** URL della pagina web delle Nazioni Unite riguardante la giornata
- **doc:** codice documento delle Nazioni Unite
- **url\_doc:** URL del documento

# JOB PYSPARK INTERNATIONAL DAYS

**Dataset utilizzato:** `international_days` ottenuto tramite scraping

## Descrizione:

- Import librerie e inizializzazione del job
- Lettura dei dati in formato .csv da piattaforma AWS S3
- Connessione al database MongoDB e inserimento del dataset nella collection `international_days_data`

## Risultato:

```
_id: ObjectId("60b7ef98160ebe7118afbe1b")  
date: "22-03"  
event: "World Water Day"  
url: "https://www.un.org/en/observances/water-day"  
doc: "A/RES/47/193"  
url_doc: "http://undocs.org/en/A/RES/47/193"
```

## CRITICITÀ

- Alcune giornate internazionali ottenute dallo scraper non presentano alcuni campi, bisognerebbe quindi attuare alcuni accorgimenti.

## POSSIBILI EVOLUZIONI

- Come futura evoluzione, si potrebbero aggiungere anche le feste nazionali a seconda della posizione dell'utente che utilizza il servizio;
- Si potrebbero utilizzare i dati delle giornate internazionali come fonti aggiuntive per eventuali approfondimenti ai TEDx Talks.



# TEAM

Andrea Appiani 1057683

Luca Mazzoleni 1054271

Cristian Sacco 1063539

Alexandr Rucodainii 1045954

Nunzio Marco Bisceglia 1046319

