TriviaTED

Importazione dei dati con PySpark e AWS Glue

Paolo Olivieri

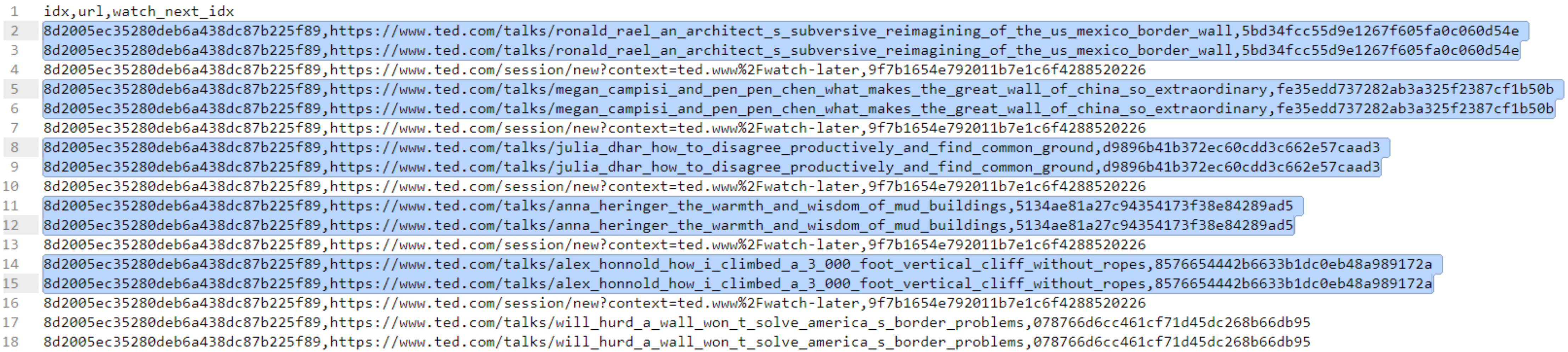
Xhorxho Papallazi

Simone Ronzoni

Dataset

Abbiamo individuato e risolto alcune criticità nei dataset

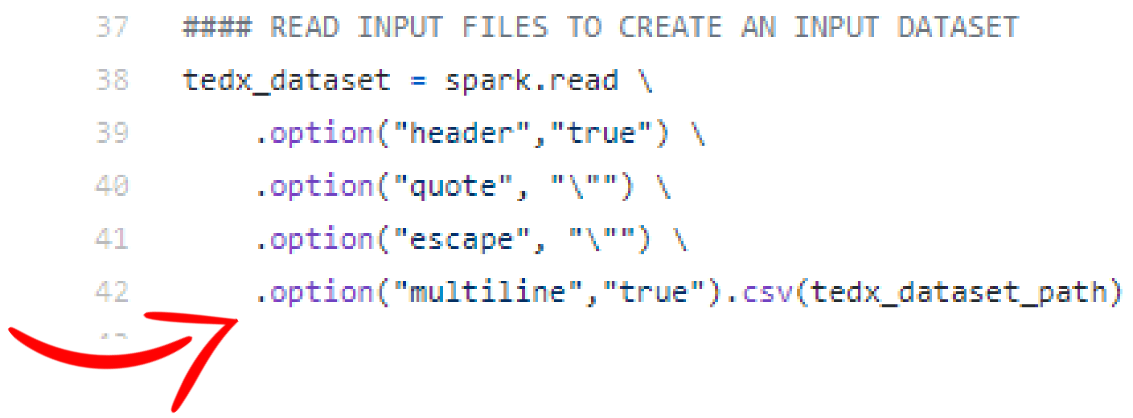
- In watch_next sono presenti numerosi duplicati
- Sono presenti record con url non valido



```
idx,url,watch_next_idx
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimagining_of_the_us_mexico_border_wall,5bd34fcc55d9e1267f605fa0c060d54e
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimagining_of_the_us_mexico_border_wall,5bd34fcc55d9e1267f605fa0c060d54e
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/session/new?context=ted.www%2Fwatch-later,9f7b1654e792011b7e1c6f4288520226
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_the_great_wall_of_china_so_extraordinary,fe35edd737282ab3a325f2387cf1b50b
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_the_great_wall_of_china_so_extraordinary,fe35edd737282ab3a325f2387cf1b50b
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/session/new?context=ted.www%2Fwatch-later,9f7b1654e792011b7e1c6f4288520226
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_find_common_ground,d9896b41b372ec60cdd3c662e57caad3
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_find_common_ground,d9896b41b372ec60cdd3c662e57caad3
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/session/new?context=ted.www%2Fwatch-later,9f7b1654e792011b7e1c6f4288520226
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_buildings,5134ae81a27c94354173f38e84289ad5
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_buildings,5134ae81a27c94354173f38e84289ad5
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/session/new?context=ted.www%2Fwatch-later,9f7b1654e792011b7e1c6f4288520226
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vertical_cliff_without_ropes,8576654442b6633b1dc0eb48a989172a
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vertical_cliff_without_ropes,8576654442b6633b1dc0eb48a989172a
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/session/new?context=ted.www%2Fwatch-later,9f7b1654e792011b7e1c6f4288520226
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_border_problems,078766d6cc461cf71d45dc268b66db95
8d2005ec35280deb6a438dc87b225f89,https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_border_problems,078766d6cc461cf71d45dc268b66db95
```

Dataset

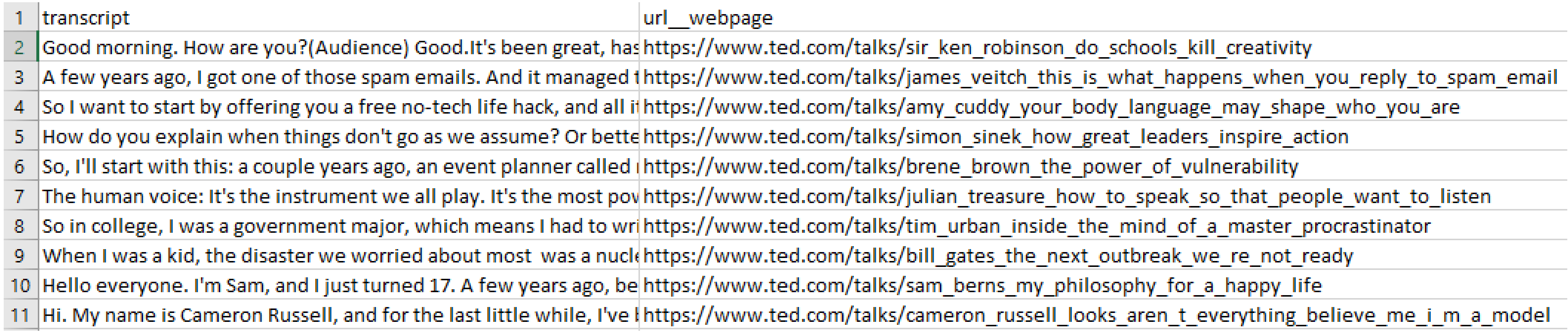
- Nei dataset sono presenti dei carriage return che non sono gestiti di default da PySpark e nemmeno nello script visto a lezione
  - Abbiamo risolto eliminandoli in fase di importazione



```
#### 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)
```

Dataset (aggiunta)

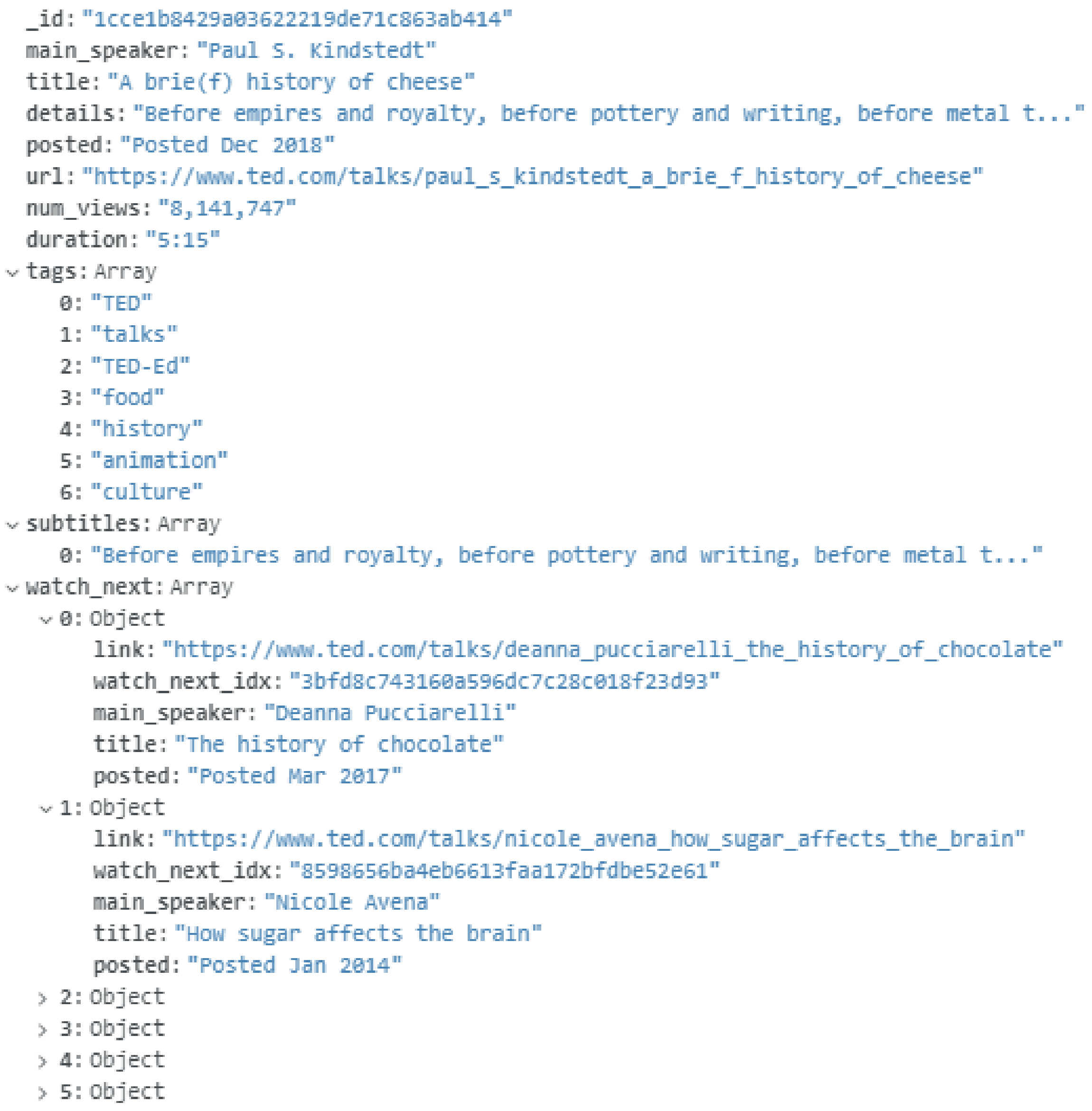
Necessari per l’implementazione della nostra applicazione abbiamo aggiunto i dati relativi alle trascrizioni dei talk in varie lingue. Per ora abbiamo utilizzato un file trovato su github generato dallo scraping delle trascrizioni (e altri dati) direttamente da TED



1	transcript	url__webpage
2	Good morning. How are you?(Audience) Good.It's been great, has	https://www.ted.com/talks/sir_ken_robinson_do_schools_kill_creativity
3	A few years ago, I got one of those spam emails. And it managed t	https://www.ted.com/talks/james_veitch_this_is_what_happens_when_you_reply_to_spam_email
4	So I want to start by offering you a free no-tech life hack, and all i	https://www.ted.com/talks/amy_cuddy_your_body_language_may_shape_who_you_are
5	How do you explain when things don't go as we assume? Or bette	https://www.ted.com/talks/simon_sinek_how_great_leaders_inspire_action
6	So, I'll start with this: a couple years ago, an event planner called	https://www.ted.com/talks/brene_brown_the_power_of_vulnerability
7	The human voice: It's the instrument we all play. It's the most pov	https://www.ted.com/talks/julian_treasure_how_to_speak_so_that_people_want_to_listen
8	So in college, I was a government major, which means I had to wri	https://www.ted.com/talks/tim_urban_inside_the_mind_of_a_master_procrastinator
9	When I was a kid, the disaster we worried about most was a nucl	https://www.ted.com/talks/bill_gates_the_next_outbreak_we_re_not_ready
10	Hello everyone. I'm Sam, and I just turned 17. A few years ago, be	https://www.ted.com/talks/sam_berns_my_philosophy_for_a_happy_life
11	Hi. My name is Cameron Russell, and for the last little while, I've	https://www.ted.com/talks/cameron_russell_looks_aren_t_everything_believe_me_i_m_a_model

Struttura dati

Dopo l’esecuzione del job la struttura di ogni documento è la seguente.



```
 _id: "1cce1b8429a03622219de71c863ab414"
 main_speaker: "Paul S. Kindstedt"
 title: "A brie(f) history of cheese"
 details: "Before empires and royalty, before pottery and writing, before metal t..."
 posted: "Posted Dec 2018"
 url: "https://www.ted.com/talks/paul_s_kindstedt_a_brie_f_history_of_cheese"
 num_views: "8,141,747"
 duration: "5:15"
⌄tags: Array
    0: "TED"
    1: "talks"
    2: "TED-Ed"
    3: "food"
    4: "history"
    5: "animation"
    6: "culture"
⌄subtitles: Array
    0: "Before empires and royalty, before pottery and writing, before metal t..."
⌄watch_next: Array
  ⌄ 0: Object
        link: "https://www.ted.com/talks/deanna_pucciarelli_the_history_of_chocolate"
        watch_next_idx: "3bfd8c743160a596dc7c28c018f23d93"
        main_speaker: "Deanna Pucciarelli"
        title: "The history of chocolate"
        posted: "Posted Mar 2017"
  ⌄ 1: Object
        link: "https://www.ted.com/talks/nicole_avena_how_sugar_affects_the_brain"
        watch_next_idx: "8598656ba4eb6613faa172bfdbe52e61"
        main_speaker: "Nicole Avena"
        title: "How sugar affects the brain"
        posted: "Posted Jan 2014"
  › 2: Object
  › 3: Object
  › 4: Object
  › 5: Object
```

Struttura dati: considerazioni

Presenza di ridondanza: abbiamo memorizzato per ogni watch next anche i dati del video oltre all’id.

Vantaggi:

- Non c’è bisogno di fare una seconda query per ottenere le informazioni dei video watch next proposti.

Svantaggi:

- Maggiore complessità e tempo richiesto per inserimento e modifica.
- Dimensione maggiore della base dati.

Nel nostro caso però andiamo a recuperare i dati facendo scraping dal sito di TedX, per cui eventuali problemi di consistenza in inserimento e modifica non ci affliggono particolarmente.

Criticità e sviluppi futuri

- Futura integrazione: reperire i dati relativi a nuovi talk mediante un job automatizzato che periodicamente effettua lo scraping.
- Future integrazione di un job che analizzi i video e in automatico possa generare i testi nelle varie lingue.
- Utilizzare i testi dei vari talk per la ricerca dettagliata.

Riferimenti

- Script PySpark: pyspark_job_tedx_population.py
- Dataset trascrizioni: TED_Talk.csv