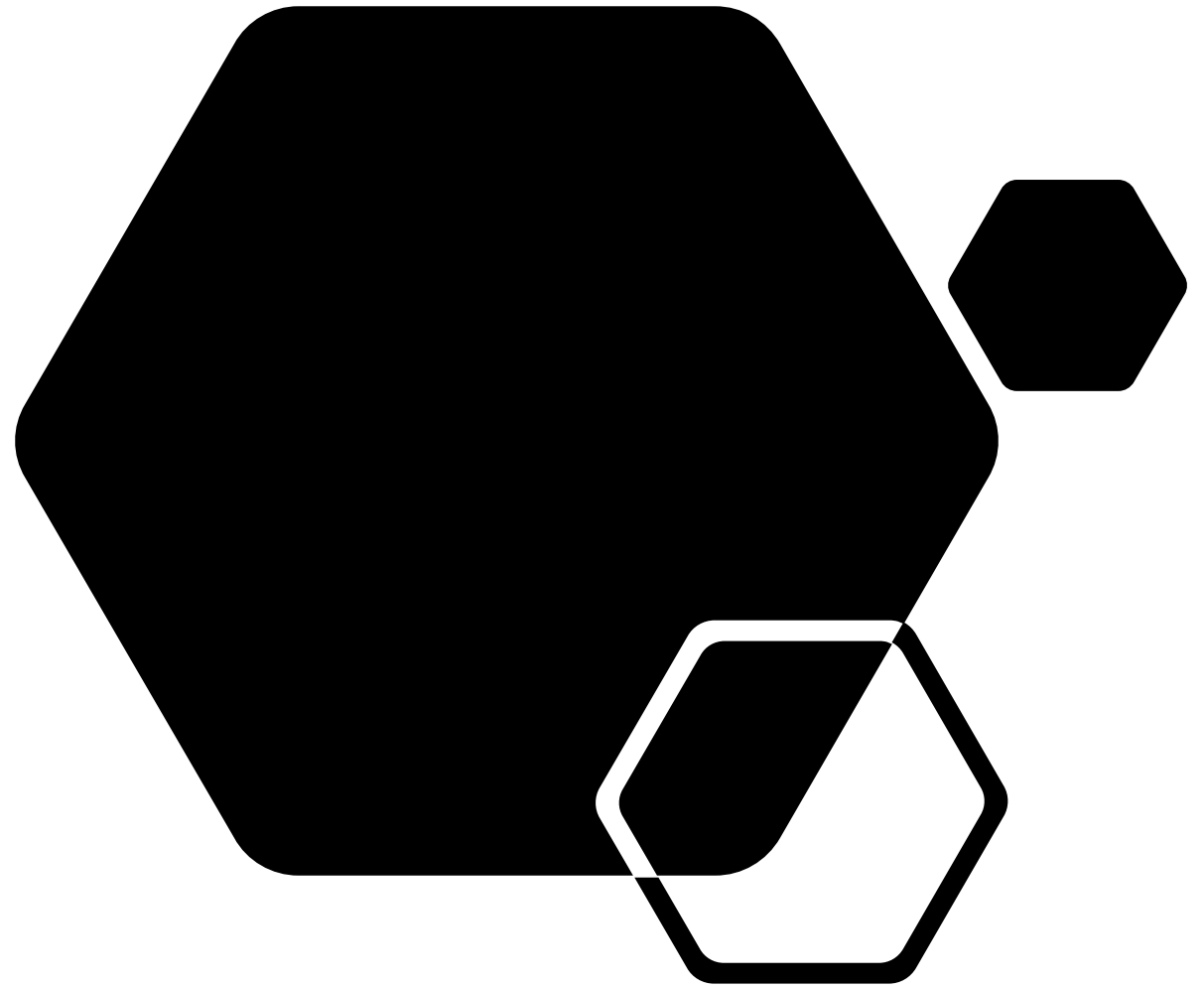
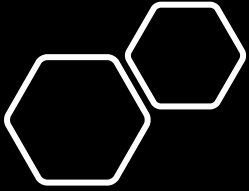


Gerry

AWS Glue & PySpark

Carne Federico – 1059865





Watch Next - Script

- Aggiunge a ogni talk la lista dei watch next suggeriti da TED
- Alcuni talk hanno un detail multilinea, spark durante l'import di `tedx_dataset` considera erroneamente ogni riga del detail come se fosse un nuovo record. Per correggere la lettura basta aggiungere `option("multiline", "true")`
- `watch_next_dataset` conteneva almeno due record uguali per ogni tupla (`idx`, `watch_next_idx`) e per ogni talk veniva indicato tra i `watch_next` anche l'url <https://www.ted.com/session/new?context=ted.www%2Fwatch-later> che non è un talk valido. Dopo aver effettuato il drop dei duplicati e rimosso i record che puntano all'url soprastante, il dataset è passato da 77364 righe a 25788

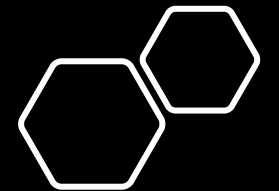
```
# READ WATCH_NEXT DATASET
watch_next_dataset_path = "s3://unibg-data-2021-1059865/watch_next_dataset.csv"
watch_next_dataset_raw = spark.read.option("header", "true").csv(watch_next_dataset_path)

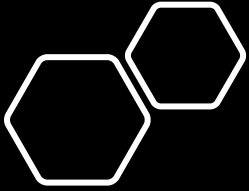
watch_next_dataset = watch_next_dataset_raw.drop_duplicates() \
    .where('url != "https://www.ted.com/session/new?context=tet.www%2Fwatch-later"')

# ADD WATCH_NEXT TO AGGREGATE MODEL
watch_next_dataset_agg = watch_next_dataset.groupBy(col("idx").alias("idx_ref_watch_next")) \
    .agg(collect_list("watch_next_idx").alias("watch_next"))

tedx_dataset_agg = tedx_dataset_agg.join(watch_next_dataset_agg,
tedx_dataset_agg._id == watch_next_dataset_agg.idx_ref_watch_next,
"left").drop("idx_ref_watch_next")
```

Watch Next - Script

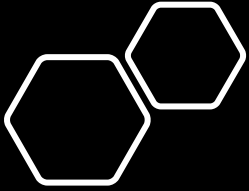




Watch Next - Collection

Dopo l'esecuzione dello script precedente, un documento della collection `tedx_data` ha la seguente struttura:

```
_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..."  
posted: "Posted Apr 2020"  
url: "https://www.ted.com/talks/alexandra_auer_the_intangible..."  
tags: Array  
  0: "TED"  
  1: "talks"  
  2: "design"  
  ...  
watch_next: Array  
  0: "8576654442b6633b1dc0eb48a989172a"  
  1: "078766d6cc461cf71d45dc268b66db95"  
  2: "d9896b41b372ec60cdd3c662e57caad3"  
  ...
```



Gerry

Transcript - Scraper & Script

Scraper - [Transcript Scraper.ipynb](#)

- Importa il file `tedx_dataset.csv`
- Per ogni talk apre la pagina `talk['url'] + '/transcript'`
- Se esiste ricava la trascrizione, formata da una o più frasi. Ogni frase ha associato l'istante in cui lo speaker inizia a pronunciarla.
- Ignora frasi come (Applause), (Laughs), (Inaudible), ...

Script

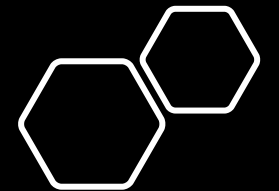
- Aggiunge ai talk la loro trascrizione, rimuovendo i talk per i quali non è stata trovata
- Mantiene la struttura (`timestamp`, `sentence`) perché una possibile evoluzione del servizio è indicare, per le risposte sbagliate, in quale punto del talk viene detta la risposta

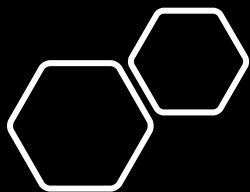
```
# READ TRANSCRIPT DATASET
transcript_dataset_path = "s3://unibg-data-2021-1059865/transcript_dataset.csv"
transcript_dataset = spark.read.option("header", "true").csv(transcript_dataset_path)

# ADD TRANSCRIPT TO AGGREGATE MODEL
transcript_dataset_agg = transcript_dataset.groupBy(col("idx").alias("idx_ref_transcript"))\
.agg(collect_list(struct("timestamp", "sentence")).alias("transcript"))

gerry_dataset = tedx_dataset_agg.join(transcript_dataset_agg,
tedx_dataset_agg._id == transcript_dataset_agg.idx_ref_transcript).drop("idx_ref_transcript")
```

Gerry Transcript - Script

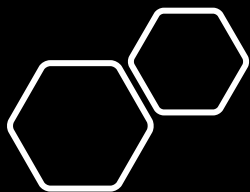




Gerry Transcript - Collection

Dopo l'esecuzione dello script precedente, un documento della collection gerry_data ha la seguente struttura:

```
_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..."  
posted: "Posted Apr 2020"  
url: "https://www.ted.com/talks/alexandra_auer_the_intangible..."  
tags: Array  
watch_next: Array  
transcript: Array  
  0: Object  
    timestamp: "00:01"  
    sentence: "Humankind loves to build walls. Have you ever not..."  
  1: Object  
    timestamp: "00:26"  
    sentence: "Growing up in Germany, the fall of the Berlin Wall..."  
  2: Object  
  3: Object  
  ...
```



Criticità e possibili evoluzioni

Criticità

- La pagina `ted.com/talks/:title` ha almeno 3 versioni differenti, rendendo complesso lo scraping
- Nella pagina web il timestamp di una frase viene prima impostato a 0 e poi incrementato fino al valore corretto attraverso uno script, bisogna attendere che il talk venga caricato (ma l'incremento può comunque avvenire in ritardo)
- Nelle trascrizioni di TED sono presenti rappresentazioni di suoni non utili per generare un quiz e quindi da rimuovere

Possibili evoluzioni

- Includere lo scraper all'interno di AWS, per automatizzare il processo di scraping e integrare Amazon Elastic Transcoder per generare la trascrizione di quei talk che non ne hanno una