



# TEDxSPEAKERS

Improve your public speaking skills



```
### ADD WATCH NEXT
tedx_watchnext_dataset_path = "s3://ieser-tedxspeakers-data/related_videos.csv"
watchNext_dataset = spark.read.option("header", "true").csv(tedx_watchnext_dataset_path)

columns_to_include = watchNext_dataset.columns[3:]
watchNext_dataset_toJoin = watchNext_dataset.groupBy(col("id").alias("id_ref")).agg(collect_list(struct(*columns_to_include)).alias("related"))
watchNext_dataset_toJoin.printSchema()
tedx_dataset_main = tedx_dataset_main.join(watchNext_dataset_toJoin, tedx_dataset_main.id == watchNext_dataset_toJoin.id_ref).drop("id_ref")

tedx_dataset_main = tedx_dataset_main.withColumn("_id", tedx_dataset_main["slug"])
```

## ● Aggiunta dei video correlati


Aggiunta del codice necessario per caricare nel database anche gli i video correlati.




[Vedi il codice completo](#)

# Risultato in MongoDB

Con l'aggiunta del codice della diapositiva precedente, abbiamo ottenuto all'interno di ogni oggetto, i riferimenti ai video correlati con i campi id, titolo e immagine.



```
_id: "george_zaidan_how_do_gas_masks_actually_work"
id: "526880"
slug: "george_zaidan_how_do_gas_masks_actually_work"
speakers: "George Zaidan"
title: "How do gas masks actually work?"
url: "https://www.ted.com/talks/george_zaidan_how_do_gas_masks_actually_work"
description: "You might think of gas masks as clunky military-looking devices. But i..."
duration: "254"
presenterDisplayName: "George Zaidan"
publishedAt: "2024-04-30T15:14:51Z"
url_image: "https://talkstar-assets.s3.amazonaws.com/production/talks/talk_128547/..."
tags: Array (8)
related: Array (3)
  0: Object
    slug: "stephanie_honchell_smith_whatever_happened_to_the_hole_in_the_ozone_la..."
    title: "Whatever happened to the hole in the ozone layer?"
    duration: "293"
    viewedCount: "552783"
    presenterDisplayName: "Stephanie Honchell Smith"
  1: Object
  2: Object
```



# Ricerca dei dati

Il sito scelto da cui fare scraping è Bccampus:  
<https://pressbooks.bccampus.ca/speaking/>.

A questo indirizzo è presente un corso diviso in 16 capitoli basato sul libro «Public Speaking». Il libro parla di come migliorare le proprie capacità di parlare in pubblico.

È possibile utilizzare questo sito come fonte perché **i termini e condizioni** indicano che è possibile riutilizzare e rielaborare il contenuto senza scopi commerciali.



# Scarping con BeautifulSoup

La libreria utilizzata è [BeautifulSoup](#), perché di facile implementazione.

Per poterla utilizzare in AWS Glue, va caricata su un S3 e poi aggiunto il percorso nelle impostazioni del job.



## Libraries [Info](#)

Python library path

s3://ieser-tedxspeakers-script/beautifulsoup4-4.12.3-py3-none-any.whl



# Scraper

```
# URL del sito web da cui recuperare i dati
url = "https://pressbooks.bccampus.ca/speaking/"

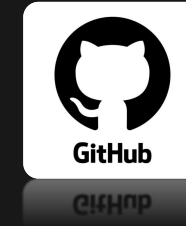
html_content = requests.get(url).content
soupBook = BeautifulSoup(html_content, 'html.parser') #Recupero il contenuto della pagina web

chapters = []
for idx, chapter in enumerate(soupBook.find_all('li', class_='toc__chapter'), start=1):
    title = link = content = "" #Resetto il tutti i contenuti

    title = chapter.find('a').get_text(strip=True).split(": ")
    title = title[1] if len(title) == 2 else "Not found"
    link = chapter.find('a')['href']

    if(link != ""):
        html_content = requests.get(link).content
        soupCapther = BeautifulSoup(html_content, 'html.parser') #Recupero il contenuto del singolo
        content = soupCapther.find('div', class_='site-content').get_text(strip=True)
        chapters.append((idx, title, link, content))

print(chapters)
transformed_df = spark.createDataFrame(chapters, ["_id", "Title", "Link", "Content"])
```



[Vedi script  
completo](#)

Dalla prima pagina recuperiamo il link di tutte le pagine interne.

Per farlo sfruttiamo le classi html che vengono messe nella pagina generale sulla lista dei capitoli.

Iteriamo ogni capitolo, per recuperare il contenuto di ogni lezione.

# Risultato nel dataset di MongoDB

```
_id: 14  
title: "Logical Reasoning"  
link: "https://pressbooks.bccampus.ca/speaking/chapter/chapter-14/"  
content: "14.1 - What is Correct Reasoning?We have seen that logos involves compos..."
```

```
_id: 9  
title: "Introductions and Conclusions"  
link: "https://pressbooks.bccampus.ca/speaking/chapter/chapter-9/"  
content: "9.1 - General Guidelines for Introductions and ConclusionsCan you imag..."
```

```
_id: 4  
title: "Audience Analysis"  
link: "https://pressbooks.bccampus.ca/speaking/chapter/chapter-4/"  
content: "4.1 - The Importance of Audience AnalysisAccording to this author, rul..."
```

```
_id: 7  
title: "Organizing and Outlining Your Speech"  
link: "https://pressbooks.bccampus.ca/speaking/chapter/chapter-7/"  
content: "7.1 - Why We Need Organization in SpeechesAs we listen, we have limits..."
```

All'interno del database ritrovo tutte le lezioni, con titolo, link alla fonte e contenuto.

L'id del campo corrisponde al numero del capitolo, così da recuperarlo più velocemente.

# Criticità



- Manca il collegamento tra i capitoli e i video di TedX
  - La struttura del sito da cui reperiamo i dati potrebbe cambiare nel tempo, e anche il codice dello Scraper andrebbe riscritto.
  - Da ogni capitolo vanno estratti i concetti chiave da riportare nella scheda della lezione.
- 