Analisi di Wikipedia

Sei stato assunto come Data Scientist da Wikimedia.

E’ necessario svolgere un’attività di EDA per analizzare e valutare statisticamente tutto il contenuto informativo offerto da Wikipedia. Il bump che ti viene fornito, possiede le seguenti categorie:

- 'culture',
- 'economics',
- 'energy',
- 'engineering',
- 'finance',
- 'humanities',
- 'medicine',
- 'pets',
- 'politics',
- 'research',
- 'science',
- 'sports',
- 'technology',
- 'trade',
- 'transport'

Nello specifico, quello che ti viene richiesto consiste nel calcolare, per ogni categoria, le seguenti informazioni.

- Numero di articoli
- Numero medio di parole utilizzate
- Numero massimo di parole presenti nell’articolo più lungo
- Numero minimo di parole presenti nell’articolo più corto

- Per ogni categoria, individuare la nuvola di parole più rappresentativa

Dopo aver svolto l'analisi richiesta, è necessario addestrare e testare un classificatore testuale capace di classificare gli articoli (secondo le categorie presenti nel dataset) che saranno in futuro inseriti,

Descrizione del dataset

Il dataset offerto è composto da 4 colonne:

- **title**: indica il titolo dell'articolo
- **summary**: contiene l'introduzione dell'articolo
- **documents**: contiene l'articolo completo
- **categoria**: contiene la categoria associata all'articolo

Per le attività precedentemente richieste, lo studente svolga il tutto considerando prima la colonna ***summary***, poi la colonna ***documents***, così da confrontare i risultati ottenuti con le due differenti colonne e verificare quale dei due classificatori ha maggiore accuratezza in termini di classificazione.

Il dataset è salvato su S3 e reperibile al seguente link:

https://proai-datasets.s3.eu-west-3.amazonaws.com/wikipedia.csv

Per poter caricare il dataframe e trasformarlo in una table basta eseguire su Notebook Databricks le seguenti righe di codice:

```
!wget https://proai-datasets.s3.eu-west-3.amazonaws.com/wikipedia.csv

import pandas as pd

dataset = pd.read_csv('/databricks/driver/wikipedia.csv')

spark_df = spark.createDataFrame(dataset)

spark_df = spark_df.drop("Unnamed: 0")

spark_df.write.saveAsTable("wikipedia")
```

N.B. Durante il loading del dataset, ci appoggiamo ad un dataframe Pandas. Questa non è una procedura comune e del tutto corretta. In questo caso ci permette di leggere correttamente (superando con poso sforzo il limite dei separator) i dati con cui definire un DataFrame Spark e una Table 'Wikipedia'.