Analisi di Wikipedia

Descrizione del Progetto

Wikidata Insights, un'azienda leader nella gestione di contenuti digitali, è stata incaricata da **Wikimedia** per ottimizzare l'analisi e la categorizzazione dei contenuti di Wikipedia. Per supportare la loro continua espansione e migliorare l'organizzazione delle informazioni, Wikidata Insights ha deciso di condurre un progetto avanzato di **data analysis** e **machine learning**. L'obiettivo principale è comprendere meglio il vasto patrimonio di contenuti informativi offerti da Wikipedia e sviluppare un sistema di **classificazione automatica** che consenta di categorizzare efficacemente i nuovi articoli futuri.

Obiettivi

1. Analisi Descrittiva dei Contenuti

Il primo obiettivo del progetto è condurre un'**analisi esplorativa dei dati (EDA)** per capire le caratteristiche dei contenuti di Wikipedia suddivisi in diverse categorie tematiche, come ad esempio: - **Cultura**, **Economia**, **Medicina**, **Tecnologia**, **Politica**, **Scienza**, e altre.

L'analisi esplorativa prevede: - Il **conteggio degli articoli** presenti per ogni categoria. - Il **numero medio di parole** per articolo. - La lunghezza dell'**articolo più lungo** e di quello **più corto** per ciascuna categoria. - La creazione di **nuvole di parole** rappresentative per ogni categoria, per identificare i termini più frequenti e rilevanti.

2. Sviluppo di un Classificatore Automatico

Il secondo obiettivo è creare un **modello di machine learning** capace di classificare automaticamente gli articoli in base alla loro categoria. Il sistema di classificazione verrà addestrato utilizzando dati di testo presenti nelle seguenti colonne del dataset: - **Sommario** (`summary`): Introduzione breve dell'articolo. - **Testo Completo** (`documents`): Contenuto completo dell'articolo.

3. Identificazione di Nuovi Insights

L'analisi consentirà anche di ottenere preziosi insights sui contenuti di Wikipedia, come la densità di articoli per categoria o le tendenze linguistiche associate a determinati argomenti. Queste informazioni possono aiutare Wikimedia a migliorare l'organizzazione delle pagine e a ottimizzare i propri sforzi editoriali.

Workflow del Progetto

Caricamento dei Dati

Il dataset è salvato su S3 e reperibile al seguente link: https://proai-datasets.s3.eu-west-3.amazonaws.com/wikipedia.csv

Utilizzando un framework distribuito come **Databricks**, i dati vengono processati in modo efficiente, partendo da un **Pandas DataFrame** per essere successivamente convertiti in un **Spark DataFrame** e salvati come una tabella chiamata "Wikipedia".

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'.

Risultati Attesi

1. Ottimizzazione dell'Organizzazione dei Contenuti

L'analisi descrittiva fornirà a Wikimedia una visione chiara e dettagliata della distribuzione e delle caratteristiche dei propri contenuti. Sarà possibile identificare quali categorie necessitano di maggiore attenzione o dove sono presenti opportunità di espansione.

2. Classificazione Automatica

Il sistema di classificazione sviluppato permetterà a Wikimedia di automatizzare il processo di categorizzazione dei nuovi articoli, migliorando l'efficienza operativa e garantendo una migliore navigabilità per gli utenti.

3. Nuovi Insights Strategici

Gli insights ottenuti dall'analisi esplorativa e dalla classificazione permetteranno a Wikimedia di ottimizzare l'allocazione delle risorse editoriali, con la possibilità di orientare le proprie campagne informative in modo più mirato.

Conclusioni

Il progetto offre a **Wikimedia** un potente strumento di **analisi dati** e **classificazione automatica** per migliorare la gestione dei propri contenuti. Attraverso l'utilizzo di tecniche avanzate di **data science** e **machine learning**, Wikimedia sarà in grado di ottimizzare la propria infrastruttura informativa e offrire un servizio di qualità superiore agli utenti di tutto il mondo.

Modalità di consegna: Link a databricks