Wikipedia Analysis – Project Guidelines

Project Description

Wikidata Insights, a leading company in digital content management, has been commissioned by Wikimedia to optimize the analysis and categorization of Wikipedia content.

To support its continuous expansion and improve content organization, Wikidata Insights has decided to conduct an advanced data analysis and machine learning project.

The main objective is to better understand the vast amount of informational content provided by Wikipedia and to develop an automatic classification system capable of effectively categorizing future new articles.

Objectives

1. Descriptive Content Analysis

The first objective of the project is to conduct an Exploratory Data Analysis (EDA) to understand the characteristics of Wikipedia content divided into different thematic categories, such as:

- Culture
- Economy
- Medicine
- Technology
- Politics
- Science
- and others

The exploratory analysis should include:

- The number of articles available for each category
- The average number of words per article
- The length of the longest and shortest article in each category
- The creation of representative word clouds for each category to identify the most frequent and relevant terms

2. Development of an Automatic Classifier

The second objective is to create a machine learning model capable of automatically classifying articles according to their category.

The classification system must be trained using textual data available in the following dataset columns:

- **Summary**: Short introduction of the article
- **Documents**: Full article content

3. Identification of New Insights

The analysis should also provide valuable insights into Wikipedia content, such as:

- Article density per category
- Linguistic patterns associated with specific topics

These insights may help Wikimedia improve content organization and optimize editorial efforts.

Project Workflow

Data Loading

The dataset is stored on Amazon S3 and is available at the following link:
https://proai-datasets.s3.eu-west-3.amazonaws.com/wikipedia.csv
Using a distributed framework such as Databricks, the data should be processed efficiently. The workflow may begin with loading the dataset into a Pandas DataFrame, converting it into a Spark DataFrame, and saving it as a table named **"Wikipedia"**.

To load the dataframe and transform it into a table in a Databricks notebook, the following steps can be executed:

```
CREATE CATALOG IF NOT EXISTS my_catalog;
CREATE SCHEMA IF NOT EXISTS my_catalog.raw;
CREATE VOLUME IF NOT EXISTS my_catalog.raw.datasets;
mkdir -p /Volumes/my_catalog/raw/datasets
curl -L "https://proai-datasets.s3.eu-west-3.amazonaws.com/wikipedia.csv" \
  -o /Volumes/my_catalog/raw/datasets/wikipedia.csv
path = "/Volumes/my_catalog/raw/datasets/wikipedia.csv"
df = spark.read.csv(path, header=True, inferSchema=True)
```

Expected Results

1. Content Organization Optimization

The descriptive analysis should provide Wikimedia with a clear and detailed overview of content distribution and characteristics.
It should be possible to identify which categories require greater attention or where expansion opportunities exist.

2. Automatic Classification

The developed classification system should allow Wikimedia to automate the categorization process of new articles, improving operational efficiency and enhancing user navigation.

3. Strategic Insights

The insights obtained from exploratory analysis and classification should enable Wikimedia to optimize editorial resource allocation and better target informational initiatives.

Conclusion

This project provides Wikimedia with a powerful data analysis and automatic classification tool to improve content management.

By leveraging advanced data science and machine learning techniques, Wikimedia will be able to optimize its information infrastructure and deliver higher-quality services to users worldwide.