Big Data Project: Wikipedia Analysis

Project Description:

I worked on a data analysis and automatic classification project for Wikipedia content. The goal was to explore and understand the structure of Wikipedia articles across multiple thematic categories and then build a machine learning model able to automatically classify future articles into the proper category.

Key Responsibilities and Tasks:

I performed descriptive analysis for each category: number of articles, average word count, longest/shortest article lengths and word clouds to identify the most frequent terms. Then I trained a text classification model using both "summary" and "full text" fields to automatically categorize articles. The entire dataset was loaded, processed and stored using Databricks (Pandas → Spark DataFrame → Spark Table).

Outcome:

The analysis produced useful insights about Wikipedia content distribution and characteristics, and the classification model enabled automatic categorization of new articles, improving editorial efficiency and user navigation.

Technologies and Tools Used:

Python, Pandas, Spark / Databricks, NLP preprocessing, Machine Learning for text classification, word clouds, data visualization.