Stackoverflow Project

This project a dataset of Stackoverflow questions in PySpark. Using various techniques and Natural Language Processing the dataset consisting of text and several metrics for each question is used for building a language detection model (classification) and a clustering model to group the questions.

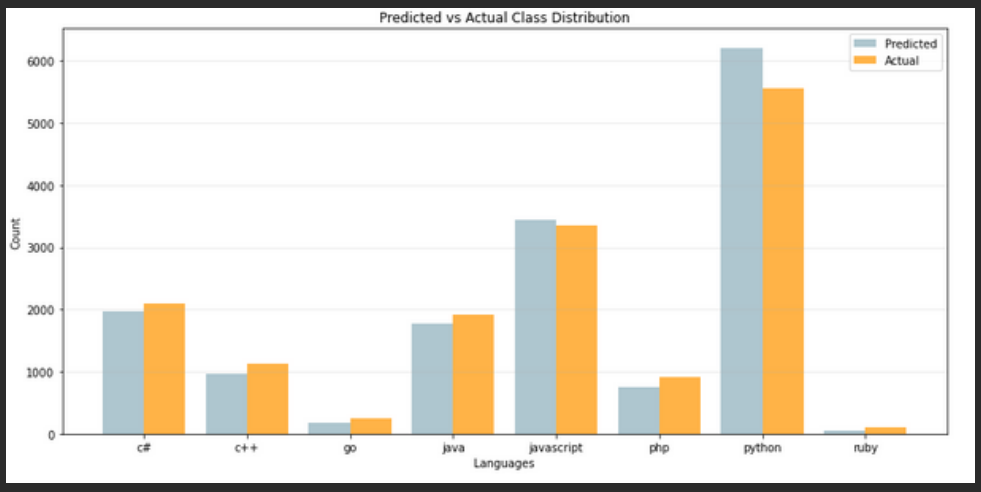
Dataset used::
Stackoverflow Datadump
https://stackoverflow.com/help/data-dumps

Steps

1. Extract the dataset from the over 100GB large Stackoverflow Dataset using PySpark and afterwards clean and preprocess it.

2. Visualize and understand the data with by creating different plots like:
   - Tags Popularity Chart: Show how the popularity of certain themes over time developed.
   - Creation Date Heatmap: Shows on whichd day and time the most posts are created

3. Creating a model to detect the primary language of a question using:
   - Logistic Regression
   - Naive Bayes
   - Random Forest



4. Cluster the questions based on 3 engagement based metrics to find groups of similar questions and analyze them afterwards