Sequential Modelling of Typing Performance Logs

1. Abstract

Sequential log data from typing sessions, including metrics such as words per minute, accuracy, and session context, forms the basis for predicting and analysing typing performance trends. In this project, a unified workflow will be developed that leverages supervised sequence learning to forecast WPM trends and incorporates unsupervised clustering analyses to reveal recurring behavioural patterns and contextual influences within the dataset.

Although the dataset consists of typing performance logs, its structure closely resembles that of large scale log data and will be treated accordingly. Data preprocessing will be distributed and performed using Apache Spark.

This approach combines predictive modelling with exploratory data analysis. Sequence learning techniques enable forecasting of future typing performance, while clustering methods offer insights into common session types, contextual factors, and the temporal stability or change in user behaviour. These analyses contribute to a deeper understanding of how typing habits develop and evolve over time.

Special attention is given to the imbalance in the dataset, since it aggregates sessions from multiple users. Such imbalance can affect model training and evaluation fairness. Strategies to mitigate data imbalance will be integrated throughout the pipeline. Distributed preprocessing via Spark will demonstrate scalable processing that reflects the demands of large scale log data applications, even though the current dataset is relatively modest in size.

The pipeline will cover stages including data ingestion and transformation, exploratory analysis, feature engineering, sequence learning, clustering, and model evaluation. This structured approach supports both predictive and descriptive insights into personalised typing behaviour.