Report 1: Similarity Detection

Group 150: Lorenzo Deflorian, Riccardo Fragale

November 5, 2025

1 Data

We decided to employ a dataset taken from Kaggle (`https://www.kaggle.com/datasets/shubchat/1002-short-stories-from-project-guttenberg`). It contains a a set of short stories extracted from the wonderful portal of Project Guttenberg. They are very well known short stories from famoous writers in history. In order to extract this files we employed a script where we use kagglehub API to download the files and store them inside the repository for this lab. For the sake of our project, we decided to analyze and compare texts from a single author as there is a higher chance that dococuments are similar.

2 Methods

3 Results

4 Conclusions