Report 1: Similarity Detection

Group 150: Lorenzo Deflorian, Riccardo Fragale

November 9, 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. In particular, every time the test is running we save all the data scraped from the dataset looking only for the author we decide to specify. We collect all the data regarding title of book, book number and the content of the short story.

2 Methods

3 Results

We decided to focus on the short stories by Charles Dickens and Edgar Allan Poe. In both cases we have a reasonable number of contents; 6 stories for the first and seven for the latter. Since their writing style is different, and also the contents and the themes of their short stories, we decided to compare the writing of a single author to the ones of the same writer. We expected to find good similarity as generally authors tend to use similar words and adopt a certain syntactical set of structures to build phrases. Our result were not as simple as previously thought. In particular, there is a high relationship (we don't know whether causality or implication) between the number of rows selected per band and the ability of our code to find a good similarity rate. The higher is the number of rows per band, the lower are the candidate pairs found y our mining pipeline. This is in our opinion reasonable as increasing the number of rows per band implies that a longer portion of the text of the short story should be almost identical to another one. This is less probable

as the authors tend to use similar phrase structures but they do not entirely copy a list of words inside their texts.

We decided to add logging messages regarding the length of all the steps of our pipeline. We found out that the longest parts are shingling and minhashing while LSH and a the final comparison of the signatures are quite fast. This implies that it is crucial to reduce as much as possible the complexity of the first two procedures (also by using Spark) while the other two are less time consuming and can be done directly on our machine without parallelizing.

4 Conclusions

This homework let us correctly understand all the steps needed to correctly identify similarities inside similar texts. We also realized that they are very important tools as they can be used to identify possible plagiarism which is a key aspect in the literature markets. Moreover, they can be employed to find stylical aspects of authors (such as we in part did) since we could now how much parts of a text are always used by a writer inside their operas.