D 351 Mini Project 2

Due: October 25, 2023

Locality-Sensitive Hashing for Documents

In this project, our aim is to employ the LSH and related techniques to discover similar documents among a collection of hundreds of distinct documents. This time, every student will perform a naive Python implementation that faithfully follows the steps described in the text.

Data

You will be working on a dataset called *similarity*, which is a structured text file composed of slightly more than 1,500 paragraphs. You can find the file on the Canvas page for the assignment.

Instructions for Python Implementation

1. Download the dataset. You must have a text file named *similarity*. This will be the input to your program.

2. Your main task is to write a Python program that discovers the **similar** paragraphs hidden in the dataset by employing the ideas developed in the sections *3.3 Similarity-Preserving Summaries of Sets* and *3.4 Locality-Sensitive Hashing for Documents*. By **similar**, we mean **almost identical**. Most of the paragraphs in this corpus are generated by AI. Therefore, you must expect similarities in format and in content. And in fact, there might be some almost-identical paragraphs even hidden to us. The more you find, the better. However, we are certain that there is a pair of paragraphs -put there by us- that are almost identical in terms of the kinds and orderings of the words they are made of. We are curious to see if you will be able to find them. We urge you to review the summary of the underlying algorithm in the subsection called *3.4.3 Combining the Techniques*, which is on page 100 of the textbook.

3. The first question you are required to answer is this: how did you represent your textual data? Namely, how did you represent each document? For example, what was your choice for shingle size, and why?

4. The second question is this: what happens if you naively try to compare the signature matrices to determine the similar documents? Why do we need the *Banding Technique*? Why do we need *LSH* instead of a direct application of *minhashing*?

5. Finally, what are the five pairs of documents among the given collection that you've found to display the greatest degree of similarity? Are they indeed almost identical?