Documentation

We only have a single class in our application which is Shingler Class.

Shingler Class

This class contains functions to create shingles and assign them a shingle ID. It also creates a map that associates each shingle with its ID.

Functions

- **preProcess**: This function pre processes the texts by removing unwanted elements and tokenizing the text. In the preprocessing before we run our program we first run another program to store each sequence of the dna from dataset into a separate document, we also separate each alphabet so it is treated as a word in sequence. We remove additional numbers and endlines and blank spaces from the original document. When we run the LSH program, before shingling we read each document and add it into a list, we then tokenize each sequence to form shingles in the next step
- **makeKShingles**: This method makes shingles of length k_shingles and stores them in a set.
- **makeShingleMap**: This method maps each shingle to an ID.
- **Jaccard**: Calculate the jaccard similarity of columns c1 and c2.