Design Document

Team Members:

- Surinderpal Singh Virk - 2018A7PS0234H
- Arpit Adlakha – 2018A7PS250H
- Saksham Pandey – 2018A7PS0259H

In this Design Document we are going to tell about the data structures used for making this application and report on the different distance measures used. First we are going to discuss about the data structures used.

Different Data Structures Used

- **Shingle set**: In this we take consecutive words and bind them together as a single object. If k words are taken and combined then we call it as k-shingle. In our application we are taking k=5.
- **Hash function**: A list of N_HASHES hash functions which contain a unique permutation of the document IDs in an array in each row
- **Signature Matrix**: A 2D array of size N_HASHES * size of docs which is filled during the LSH algorithm.

We choose a smaller length of k shingle because for a large dataset of over 1600 documents making an index table took 15 seconds while for length 6 it took over 60 seconds. However more similarity was calculated more accurately when shingle length was chosen to be 6. We see that the time complexity grows exponentially. We therefore had to sacrifice accuracy for speed.

After Data Structures now comes distance measured used. We have used Jaccard Similarity as a distance measured.

- **Jaccard Similarity**

  $J(A,B)=|A \cap B| / |AUB|$

Pre Processing Done

In the pre-processing 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 end lines 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

We choose the bucket size and size of bands equal to 2 and 5 respectively (for convenient output).

Runtime

```
Import:  0.017015399999999792
Shingling and Shingle mapping:  0.4665178000000001
Hashing:  0.009903800000000018
Signature Matrix:  3.3465115999999995
Buckets and Bands:  0.0033086999999998312
Enter a document ID 64
The candidate pairs are-
64,6
64,29
64,45
64,51
64,55
64,65
64,66
64,109
64,112
64,114
64,133

Jaccard Similarity
Doc65 0.9702602230483272
Doc66 0.8996282527881041
Doc55 0.2936802973977695
Doc112 0.2862453531598513
Doc29 0.275092936802974
Doc51 0.25650557620817843
Doc6 0.241635687732342
Doc114 0.21189591078066913
Doc45 0.1895910780669145
Doc109 0.18587360594795538
Doc133 0.17472118959107807

Retrieval:  0.040275400000000516
```

- Import : 0.01702s
- Pre Processing(Shingling and Shingle mapping): 0.46652s
- Hashing: 0.0099s
- Signature Matrix: 3.34651s
- Buckets and Bands: 0.00331s
- Retrieval: 0.04027s