# DS702 Assignment 2

Release Date: 08 February 2022
Due Date: 22 February 2022

- Submit your answers as an electronic copy on Moodle (pdf, jupyter notebook).
- No unapproved extension of deadline is allowed. For emergencies and sickness, extensions must be requested as soon as possible.
- Cite your sources if you are taking help (papers, websites, students etc.).
- Plagiarism is strictly prohibited. Negative mark will be assigned for plagiarism.
- Remember to comment your code. And your answers should be detailed.

# 1 Theoretical Questions

## 1.1 Shingling

1. What are the first ten 3-shingles in the first sentence of Section 3.2? [1]
2. If we use the stop-word-based shingles of Section 3.2.4 [1], and we take the stop words to be all the words of three or fewer letters, then what are the shingles in the first sentence of Section 3.2 [1]?
3. What is the largest number of k-shingles a document of n bytes can have? You may assume that the size of the alphabet is large enough that the number of possible strings of length k is at least n.

## 1.2 Similarity-preserving summaries of sets

### 1.2.1 Hash functions

Using the data from Fig. 3.4 [1], add to the signatures of the columns the values of the following hash functions:

(a) $h_3(x) = 2x + 4 \text{ mod } 5$.

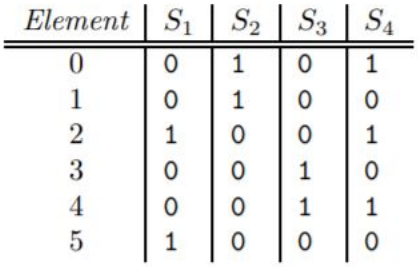
(b) $h_4(x) = 3x - 1 \text{ mod } 5$.

### 1.2.2 MinHash signature

In Fig. 1 is a matrix with six rows.

(a) Compute the minhash signature for each column if we use the following three hash functions: $h_1(x)$ = 2x + 1 mod 6; $h_2(x)$ = 3x + 2 mod 6; $h_3(x)$ = 5x + 2 mod 6.

(b) Which of these hash functions are true permutations?



| *Element* | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | 0 |
| 4 | 0 | 0 | 1 | 1 |
| 5 | 1 | 0 | 0 | 0 |

Figure 1: Matrix for Exercise 1.2.2

## 2 Distances

Write a PySpark program that implements the following distances:

(a) Jaccard Distance

(b) Cosine Distance

(c) Hamming Distance

## 3 Shingles

Find one document, and write a function that finds k-shingles of the input, and finds most common words (top 5).
k values to test are from 2 to 10.

## 4 Misleading profile selection

**Input:**

1. A textual file containing the list of movies watched by the users of a video on demand service
   - Each line of the file contains the information about one visualization : `userid, movieid, start-timestamp, end-timestamp`
   - The user with id `userid` watched the movie with id `movieid` from `start-timestamp` to `end-timestamp`
2. A second textual file containing the list of preferences for each user
   - Each line of the file contains the information about one preference : `userid, movie-genre`
   - The user with id `userid` liked the movie of type `movie-genre`
3. A third textual file containing the list of movies with the associated information
   - Each line of the file contains the information about one movie: `movieid, title, movie-genre`
   - There is only one line for each movie, i.e., each movie has one single genre

**Output:**

Select the userids of the list of users with a misleading profile

- A user has a misleading profile if more than `threshold` of the movies he/she watched are not associated with a movie genre he/she likes
- `threshold` is an argument/parameter of the application and it is specified by the user

# References

[1] Jure Leskovec et al. Mining of Massive Datasets. 2019