Assignment 2 – Information Retrieval

**Task 1:**

1. Web pages are stored in search engines to display a part of the web page for the search query and to detect changes on further indexing.
2. Individual storage of web pages would cause high use of search time and the read time is minimized. Read time is way shorter and effective than the search time.

**Task 2:**

1. • Checksum technique = each character has its own value -> near-duplicates cant be detected

• CRC technique = same as checksum but consides the position of bytes -> near duplicates cant be detected but more accurate than checksum

• Shingling technique = n-grams are getting hashed, sentence gets splitted-> near duplicates can be detected

• Simhash technique = whole sentence into Vector -> near duplicates can be detected

**Task 3:**

1. Tropical fish comprise fish species that reside in tropical settings, including both freshwater and marine environments.

(1)

|  |  |
| --- | --- |
| Word | Term frequency |
| Tropical | 2 |
| Fish | 2 |
| Comprise | 1 |
| Species | 1 |
| That | 1 |
| Reside | 1 |
| In | 1 |
| Settings | 1 |
| Including | 1 |
| Both | 1 |
| Freshwater | 1 |
| And | 1 |
| Marine | 1 |
| Environments | 1 |

(2)

|  |  |
| --- | --- |
| Word | Binary representation |
| Tropical | 00000000 |
| Fish | 00000001 |
| Comprise | 00000010 |
| Species | 00000011 |
| That | 00000100 |
| Reside | 00000101 |
| In | 00000110 |
| Settings | 00000111 |
| Including | 00001000 |
| Both | 00001001 |
| Freshwater | 00001010 |
| And | 00001011 |
| Marine | 00001100 |
| Environments | 00001101 |

(3)

(Simhash V) = {-16, -16, -16, -16, -4, -4, -4, 0}T

(4)

Fingerprint(V) = {0,0,0,0,0,0,0,0}T

**Task 4:**

1. Ralf Krestel is the best.
2. Parse

2. Group

3. Select

4. Hash

5. Store

6. Compare

2. Ralf Krestel is, Krestel is the, is the best

3./4. 919, 471, 31342

5. 3-gram selection: 0 mod 2-> 31342

1. 6. Aftab is the best

Aftab is the, is the best

412, 31342

Sim(Jac(A,B)) = 1/5

D(Jac(A,B)) = 1-(1/5) = 4/5