IT 350

Assignment 2 - Similar Item set and Minhashing

Name: Niraj Nandish

Roll no.: 191IT234

Link to Colab notebook - https://colab.research.google.com/drive/1y3mRi2ByZV3XOZrnjjfAjYb0XBuZvbdR?usp=sharing

Link to Kannada document for part D - https://drive.google.com/drive/folders/1CmiirZRLgpfycEuGuj2JrF2KSn1lylNS?usp=sharing

A: How many distinct shingles are there for each document with each type of shingle?

```
# Function to get k-shingles based on characters for given k and document
def k_shingles_chars(k, filename):
    file = open(filename)
    data = file.read()
    file.close()
    duplicates = {}
    shingles = []

    for i in range(len(data) - k):
        test_shingle = data[i : i + k]
        if test_shingle not in duplicates:
            shingles.append(test_shingle)
            duplicates[test_shingle] = 1

    return shingles


# Function to get k-shingles based on words for given k and document
def k_shingles_words(k, filename):
```

```
    file = open(filename)
    data = file.read()
    file.close()
    duplicates = {}
    shingles = []
    words = data.split()

    for i in range(len(words) - k):
        test_shingle = words[i : i + k]
        if str(test_shingle) not in duplicates:
            shingles.append(str(test_shingle))
            duplicates[str(test_shingle)] = 1

    return shingles


# 5-shingles based on characters for all documents
d1_shingles_5 = k_shingles_chars(5, "document1.txt")
d2_shingles_5 = k_shingles_chars(5, "document2.txt")
d3_shingles_5 = k_shingles_chars(5, "document3.txt")
d4_shingles_5 = k_shingles_chars(5, "document4.txt")

print("Number of distinct 5-shingles based on character for document 1: ", len(d1_shingles_5))
print("Number of distinct 5-shingles based on character for document 2: ", len(d2_shingles_5))
print("Number of distinct 5-shingles based on character for document 3: ", len(d3_shingles_5))
print("Number of distinct 5-shingles based on character for document 4: ", len(d4_shingles_5))
print("-----------------------------------------------------------------------")

# 8-shingles based on characters for all documents
d1_shingles_8 = k_shingles_chars(8, "document1.txt")
d2_shingles_8 = k_shingles_chars(8, "document2.txt")
d3_shingles_8 = k_shingles_chars(8, "document3.txt")
d4_shingles_8 = k_shingles_chars(8, "document4.txt")

print("Number of distinct 8-shingles based on character for document 1: ", len(d1_shingles_8))
print("Number of distinct 8-shingles based on character for document 2: ", len(d2_shingles_8))
print("Number of distinct 8-shingles based on character for document 3: ", len(d3_shingles_8))
print("Number of distinct 8-shingles based on character for document 4: ", len(d4_shingles_8))
print("-----------------------------------------------------------------------")
```

```
# 4-shingles based on words for all documents
d1_shingles_4 = k_shingles_words(4, "document1.txt")
d2_shingles_4 = k_shingles_words(4, "document2.txt")
d3_shingles_4 = k_shingles_words(4, "document3.txt")
d4_shingles_4 = k_shingles_words(4, "document4.txt")

print("Number of distinct 4-shingles based on words for document 1: ", len(d1_shingles_4))
print("Number of distinct 4-shingles based on words for document 2: ", len(d2_shingles_4))
print("Number of distinct 4-shingles based on words for document 3: ", len(d3_shingles_4))
print("Number of distinct 4-shingles based on words for document 4: ", len(d4_shingles_4))
```

```
Number of distinct 5-shingles based on character for document 1:  5027
Number of distinct 5-shingles based on character for document 2:  3678
Number of distinct 5-shingles based on character for document 3:  3014
Number of distinct 5-shingles based on character for document 4:  4466
----------------------------------------------------------------------
Number of distinct 8-shingles based on character for document 1:  7021
Number of distinct 8-shingles based on character for document 2:  4814
Number of distinct 8-shingles based on character for document 3:  3888
Number of distinct 8-shingles based on character for document 4:  6247
----------------------------------------------------------------------
Number of distinct 4-shingles based on words for document 1:  1444
Number of distinct 4-shingles based on words for document 2:  992
Number of distinct 4-shingles based on words for document 3:  785
Number of distinct 4-shingles based on words for document 4:  1277
```

B: Compute the Jaccard distance between all pairs of documents for each type of shingling

```
# Jaccard distance betweeen pairs of documents
def jaccard_dist(shingles1, shingles2):
    union = list(set(shingles1) | set(shingles2))
    intersection = list(set(shingles1) & set(shingles2))
    dist = 1 - (len(intersection) / len(union))
    return dist
```

```
# Jaccard distance for all pairs of documents for 5-shingling based on characters
print(
    "Jaccard distance between document 1 and document 2 for 5-shingling based on characters: ",
    jaccard_dist(d1_shingles_5, d2_shingles_5),
)
print(
    "Jaccard distance between document 1 and document 3 for 5-shingling based on characters: ",
    jaccard_dist(d1_shingles_5, d3_shingles_5),
)
print(
    "Jaccard distance between document 1 and document 4 for 5-shingling based on characters: ",
    jaccard_dist(d1_shingles_5, d4_shingles_5),
)
print(
    "Jaccard distance between document 2 and document 3 for 5-shingling based on characters: ",
    jaccard_dist(d2_shingles_5, d3_shingles_5),
)
print(
    "Jaccard distance between document 2 and document 4 for 5-shingling based on characters: ",
    jaccard_dist(d2_shingles_5, d4_shingles_5),
)
print(
    "Jaccard distance between document 3 and document 4 for 5-shingling based on characters: ",
    jaccard_dist(d3_shingles_5, d4_shingles_5),
)
print("-----------------------------------------------------------------------------------------------------")

# Jaccard distance for all pairs of documents for 8-shingling based on characters
print(
    "Jaccard distance between document 1 and document 2 for 8-shingling based on characters: ",
    jaccard_dist(d1_shingles_8, d2_shingles_8),
)
print(
    "Jaccard distance between document 1 and document 3 for 8-shingling based on characters: ",
    jaccard_dist(d1_shingles_8, d3_shingles_8),
)
print(
    "Jaccard distance between document 1 and document 4 for 8-shingling based on characters: ",
    jaccard_dist(d1_shingles_8, d4_shingles_8),
```

```
)
print(
    "Jaccard distance between document 2 and document 3 for 8-shingling based on characters: ",
    jaccard_dist(d2_shingles_8, d3_shingles_8),
)
print(
    "Jaccard distance between document 2 and document 4 for 8-shingling based on characters: ",
    jaccard_dist(d2_shingles_8, d4_shingles_8),
)
print(
    "Jaccard distance between document 3 and document 4 for 8-shingling based on characters: ",
    jaccard_dist(d3_shingles_8, d4_shingles_8),
)
print("-----------------------------------------------------------------------------------------------------")

# Jaccard distance for all pairs of documents for 4-shingling based on words
print(
    "Jaccard distance between document 1 and document 2 for 4-shingling based on words: ",
    jaccard_dist(d1_shingles_4, d2_shingles_4),
)
print(
    "Jaccard distance between document 1 and document 3 for 4-shingling based on words: ",
    jaccard_dist(d1_shingles_4, d3_shingles_4),
)
print(
    "Jaccard distance between document 1 and document 4 for 4-shingling based on words: ",
    jaccard_dist(d1_shingles_4, d4_shingles_4),
)
print(
    "Jaccard distance between document 2 and document 3 for 4-shingling based on words: ",
    jaccard_dist(d2_shingles_4, d3_shingles_4),
)
print(
    "Jaccard distance between document 2 and document 4 for 4-shingling based on words: ",
    jaccard_dist(d2_shingles_4, d4_shingles_4),
)
print(
    "Jaccard distance between document 3 and document 4 for 4-shingling based on words: ",
```

```
    jaccard_dist(d3_shingles_4, d4_shingles_4),
)
```

```
Jaccard distance between document 1 and document 2 for 5-shingling based on characters:  0.8822547508988187
Jaccard distance between document 1 and document 3 for 5-shingling based on characters:  0.8892112170189252
Jaccard distance between document 1 and document 4 for 5-shingling based on characters:  0.8674540682414698
Jaccard distance between document 2 and document 3 for 5-shingling based on characters:  0.8820581356498497
Jaccard distance between document 2 and document 4 for 5-shingling based on characters:  0.8810112668315471
Jaccard distance between document 3 and document 4 for 5-shingling based on characters:  0.8788968824940048
--------------------------------------------------------------------------------------------------------
Jaccard distance between document 1 and document 2 for 8-shingling based on characters:  0.9771843401607467
Jaccard distance between document 1 and document 3 for 8-shingling based on characters:  0.9790360318203088
Jaccard distance between document 1 and document 4 for 8-shingling based on characters:  0.9741765888356271
Jaccard distance between document 2 and document 3 for 8-shingling based on characters:  0.9752708431464908
Jaccard distance between document 2 and document 4 for 8-shingling based on characters:  0.9753589624826309
Jaccard distance between document 3 and document 4 for 8-shingling based on characters:  0.9764694001211877
--------------------------------------------------------------------------------------------------------
Jaccard distance between document 1 and document 2 for 4-shingling based on words:  0.998766954377312
Jaccard distance between document 1 and document 3 for 4-shingling based on words:  0.9991019308486754
Jaccard distance between document 1 and document 4 for 4-shingling based on words:  1.0
Jaccard distance between document 2 and document 3 for 4-shingling based on words:  0.9977439368302312
Jaccard distance between document 2 and document 4 for 4-shingling based on words:  0.9995590828924162
Jaccard distance between document 3 and document 4 for 4-shingling based on words:  0.9985429820301117
```

C. Change to any Similarity Function (use any recent similarity distance) and check the distance

```
!pip install python-Levenshtein
import Levenshtein as lev
```

```
Collecting python-Levenshtein
  Downloading python-Levenshtein-0.12.2.tar.gz (50 kB)
     |████████████████████████████████| 50 kB 2.5 MB/s
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from python-Levenshtein)
Building wheels for collected packages: python-Levenshtein
  Building wheel for python-Levenshtein (setup.py) ... done
  Created wheel for python-Levenshtein: filename=python_Levenshtein-0.12.2-cp37-cp37m-linux_x86_64.whl size=1
  Stored in directory: /root/.cache/pip/wheels/05/5f/ca/7c4367734892581bb5ff896f15027a932c551080b2abd3e00d
```

```
  Successfully built python-Levenshtein
  Installing collected packages: python-Levenshtein
  Successfully installed python-Levenshtein-0.12.2
```

```
# Levenshtein distance betweeen pairs of documents
def levenshtein_dist(filename1, filename2):
    file1 = open(filename1, "r")
    file2 = open(filename2, "r")
    text1 = file1.read()
    text2 = file2.read()
    file1.close()
    file2.close()
    return lev.ratio(text1, text2)
```

```
# Levenshtein distance between all pairs of documents
print("Levenshtein Distance between document 1 and document 2 : ", levenshtein_dist("document1.txt", "document2.tx
print("Levenshtein Distance between document 1 and document 3 : ", levenshtein_dist("document1.txt", "document3.tx
print("Levenshtein Distance between document 1 and document 4 : ", levenshtein_dist("document1.txt", "document4.tx
print("Levenshtein Distance between document 2 and document 3 : ", levenshtein_dist("document2.txt", "document3.tx
print("Levenshtein Distance between document 2 and document 4 : ", levenshtein_dist("document2.txt", "document4.tx
print("Levenshtein Distance between document 3 and document 4 : ", levenshtein_dist("document1.txt", "document4.tx
```

```
  Levenshtein Distance between document 1 and document 2 :  0.40949598246895547
  Levenshtein Distance between document 1 and document 3 :  0.38708656432979904
  Levenshtein Distance between document 1 and document 4 :  0.4311125697407286
  Levenshtein Distance between document 2 and document 3 :  0.43817920852983394
  Levenshtein Distance between document 2 and document 4 :  0.43364909723908995
  Levenshtein Distance between document 3 and document 4 :  0.4311125697407286
```

D. Try the above all for anyone Indian language.

I chose kannada for this part

```
# 5-shingles based on characters for all documents
d1_shingles_5_kannada = k_shingles_chars(5, "document1-kannada.txt")
d2_shingles_5_kannada = k_shingles_chars(5, "document2-kannada.txt")
```

```
d3_shingles_5_kannada = k_shingles_chars(5, "document3-kannada.txt")
d4_shingles_5_kannada = k_shingles_chars(5, "document4-kannada.txt")

print("Number of distinct 5-shingles based on character for document 1: ", len(d1_shingles_5_kannada))
print("Number of distinct 5-shingles based on character for document 2: ", len(d2_shingles_5_kannada))
print("Number of distinct 5-shingles based on character for document 3: ", len(d3_shingles_5_kannada))
print("Number of distinct 5-shingles based on character for document 4: ", len(d4_shingles_5_kannada))
print("-----------------------------------------------------------------------")

# 8-shingles based on characters for all documents
d1_shingles_8_kannada = k_shingles_chars(8, "document1-kannada.txt")
d2_shingles_8_kannada = k_shingles_chars(8, "document2-kannada.txt")
d3_shingles_8_kannada = k_shingles_chars(8, "document3-kannada.txt")
d4_shingles_8_kannada = k_shingles_chars(8, "document4-kannada.txt")

print("Number of distinct 8-shingles based on character for document 1: ", len(d1_shingles_8_kannada))
print("Number of distinct 8-shingles based on character for document 2: ", len(d2_shingles_8_kannada))
print("Number of distinct 8-shingles based on character for document 3: ", len(d3_shingles_8_kannada))
print("Number of distinct 8-shingles based on character for document 4: ", len(d4_shingles_8_kannada))
print("-----------------------------------------------------------------------")

# 4-shingles based on words for all documents
d1_shingles_4_kannada = k_shingles_words(4, "document1-kannada.txt")
d2_shingles_4_kannada = k_shingles_words(4, "document2-kannada.txt")
d3_shingles_4_kannada = k_shingles_words(4, "document3-kannada.txt")
d4_shingles_4_kannada = k_shingles_words(4, "document4-kannada.txt")

print("Number of distinct 4-shingles based on words for document 1: ", len(d1_shingles_4_kannada))
print("Number of distinct 4-shingles based on words for document 2: ", len(d2_shingles_4_kannada))
print("Number of distinct 4-shingles based on words for document 3: ", len(d3_shingles_4_kannada))
print("Number of distinct 4-shingles based on words for document 4: ", len(d4_shingles_4_kannada))
```

```
Number of distinct 5-shingles based on character for document 1:  5213
Number of distinct 5-shingles based on character for document 2:  3671
Number of distinct 5-shingles based on character for document 3:  2970
Number of distinct 5-shingles based on character for document 4:  4583
-----------------------------------------------------------------------
Number of distinct 8-shingles based on character for document 1:  7295
Number of distinct 8-shingles based on character for document 2:  4974
```

```
Number of distinct 8-shingles based on character for document 3:  3956
Number of distinct 8-shingles based on character for document 4:  6350
-----------------------------------------------------------------------
Number of distinct 4-shingles based on words for document 1:  1062
Number of distinct 4-shingles based on words for document 2:  715
Number of distinct 4-shingles based on words for document 3:  556
Number of distinct 4-shingles based on words for document 4:  883
```

```
# Jaccard distance for all pairs of documents for 5-shingling based on characters
print(
    "Jaccard distance between document 1 and document 2 for 5-shingling based on characters: ",
    jaccard_dist(d1_shingles_5_kannada, d2_shingles_5_kannada),
)
print(
    "Jaccard distance between document 1 and document 3 for 5-shingling based on characters: ",
    jaccard_dist(d1_shingles_5_kannada, d3_shingles_5_kannada),
)
print(
    "Jaccard distance between document 1 and document 4 for 5-shingling based on characters: ",
    jaccard_dist(d1_shingles_5_kannada, d4_shingles_5_kannada),
)
print(
    "Jaccard distance between document 2 and document 3 for 5-shingling based on characters: ",
    jaccard_dist(d2_shingles_5_kannada, d3_shingles_5_kannada),
)
print(
    "Jaccard distance between document 2 and document 4 for 5-shingling based on characters: ",
    jaccard_dist(d2_shingles_5_kannada, d4_shingles_5_kannada),
)
print(
    "Jaccard distance between document 3 and document 4 for 5-shingling based on characters: ",
    jaccard_dist(d3_shingles_5_kannada, d4_shingles_5_kannada),
)
print("-----------------------------------------------------------------------------------------------------")

# Jaccard distance for all pairs of documents for 8-shingling based on characters
print(
    "Jaccard distance between document 1 and document 2 for 8-shingling based on characters: ",
```

```
    jaccard_dist(d1_shingles_8_kannada, d2_shingles_8_kannada),
)
print(
    "Jaccard distance between document 1 and document 3 for 8-shingling based on characters: ",
    jaccard_dist(d1_shingles_8_kannada, d3_shingles_8_kannada),
)
print(
    "Jaccard distance between document 1 and document 4 for 8-shingling based on characters: ",
    jaccard_dist(d1_shingles_8_kannada, d4_shingles_8_kannada),
)
print(
    "Jaccard distance between document 2 and document 3 for 8-shingling based on characters: ",
    jaccard_dist(d2_shingles_8_kannada, d3_shingles_8_kannada),
)
print(
    "Jaccard distance between document 2 and document 4 for 8-shingling based on characters: ",
    jaccard_dist(d2_shingles_8_kannada, d4_shingles_8_kannada),
)
print(
    "Jaccard distance between document 3 and document 4 for 8-shingling based on characters: ",
    jaccard_dist(d3_shingles_8_kannada, d4_shingles_8_kannada),
)
print("-----------------------------------------------------------------------------------------------------")

# Jaccard distance for all pairs of documents for 4-shingling based on words
print(
    "Jaccard distance between document 1 and document 2 for 4-shingling based on words: ",
    jaccard_dist(d1_shingles_4_kannada, d2_shingles_4_kannada),
)
print(
    "Jaccard distance between document 1 and document 3 for 4-shingling based on words: ",
    jaccard_dist(d1_shingles_4_kannada, d3_shingles_4_kannada),
)
print(
    "Jaccard distance between document 1 and document 4 for 4-shingling based on words: ",
    jaccard_dist(d1_shingles_4_kannada, d4_shingles_4_kannada),
)
print(
    "Jaccard distance between document 2 and document 3 for 4-shingling based on words: ",
```

```
    jaccard_dist(d2_shingles_4_kannada, d3_shingles_4_kannada),
)
print(
    "Jaccard distance between document 2 and document 4 for 4-shingling based on words: ",
    jaccard_dist(d2_shingles_4_kannada, d4_shingles_4_kannada),
)
print(
    "Jaccard distance between document 3 and document 4 for 4-shingling based on words: ",
    jaccard_dist(d3_shingles_4_kannada, d4_shingles_4_kannada),
)
```

```
Jaccard distance between document 1 and document 2 for 5-shingling based on characters:  0.8685685175751401
Jaccard distance between document 1 and document 3 for 5-shingling based on characters:  0.8802681992337165
Jaccard distance between document 1 and document 4 for 5-shingling based on characters:  0.8648899188876014
Jaccard distance between document 2 and document 3 for 5-shingling based on characters:  0.8698093941456774
Jaccard distance between document 2 and document 4 for 5-shingling based on characters:  0.8690052069060017
Jaccard distance between document 3 and document 4 for 5-shingling based on characters:  0.8743666169895679
--------------------------------------------------------------------------------------------------------
Jaccard distance between document 1 and document 2 for 8-shingling based on characters:  0.9599016615801966
Jaccard distance between document 1 and document 3 for 8-shingling based on characters:  0.9682714351215039
Jaccard distance between document 1 and document 4 for 8-shingling based on characters:  0.9589532310978867
Jaccard distance between document 2 and document 3 for 8-shingling based on characters:  0.9572629612330686
Jaccard distance between document 2 and document 4 for 8-shingling based on characters:  0.9598603839441536
Jaccard distance between document 3 and document 4 for 8-shingling based on characters:  0.9630747560116713
--------------------------------------------------------------------------------------------------------
Jaccard distance between document 1 and document 2 for 4-shingling based on words:  1.0
Jaccard distance between document 1 and document 3 for 4-shingling based on words:  1.0
Jaccard distance between document 1 and document 4 for 4-shingling based on words:  1.0
Jaccard distance between document 2 and document 3 for 4-shingling based on words:  0.9984239558707644
Jaccard distance between document 2 and document 4 for 4-shingling based on words:  1.0
Jaccard distance between document 3 and document 4 for 4-shingling based on words:  1.0
```

```
# Levenshtein distance between all pairs of kannada language documents
print(
    "Levenshtein distance between document 1 and document 2: ",
    levenshtein_dist("document1-kannada.txt", "document2-kannada.txt"),
)
print(
    "Levenshtein distance between document 1 and document 3: ",
    levenshtein_dist("document1-kannada.txt", "document3-kannada.txt"),
```

```
)
print(
    "Levenshtein distance between document 1 and document 4: ",
    levenshtein_dist("document1-kannada.txt", "document4-kannada.txt"),
)
print(
    "Levenshtein distance between document 2 and document 3: ",
    levenshtein_dist("document2-kannada.txt", "document3-kannada.txt"),
)
print(
    "Levenshtein distance between document 2 and document 4: ",
    levenshtein_dist("document2-kannada.txt", "document4-kannada.txt"),
)
print(
    "Levenshtein distance between document 3 and document 4: ",
    levenshtein_dist("document1-kannada.txt", "document4-kannada.txt"),
)
```

```
 Levenshtein distance between document 1 and document 2:  0.36362376100367366
 Levenshtein distance between document 1 and document 3:  0.34232600317388345
 Levenshtein distance between document 1 and document 4:  0.3777206297434611
 Levenshtein distance between document 2 and document 3:  0.3810731707317073
 Levenshtein distance between document 2 and document 4:  0.38287037037037036
 Levenshtein distance between document 3 and document 4:  0.3777206297434611
```

E. Build a min hash signature for the above experiment and provide your conclusions for the entire experiment.

```
import binascii
import random


def coeff(x):
    randlist = []
    maxshingleid = 2**32 - 1
    while x > 0:
```

```
        i = random.randint(0, maxshingleid)
        while i in randlist:
            i = random.randint(0, maxshingleid)
        randlist.append(i)
        x -= 1

    return randlist


def minhash_signatures(numhashes, shingles):
    c = 10007
    a = coeff(numhashes)
    b = coeff(numhashes)
    signature = []
    for i in range(0, numhashes):
        minhashcode = c + 1
        for id in shingles:
            id = binascii.crc32(id.encode()) & 0xFFFFFFFF
            hashcode = (a[i] * id + b[i]) % c
            if hashcode < minhashcode:
                minhashcode = hashcode
        signature.append(minhashcode)
    return signature


# signatures for 5-shingles based on characters
d1_sign_5 = minhash_signatures(10, d1_shingles_5)
d2_sign_5 = minhash_signatures(10, d2_shingles_5)
d3_sign_5 = minhash_signatures(10, d3_shingles_5)
d4_sign_5 = minhash_signatures(10, d4_shingles_5)

print("Min hash signature of 5-shingle construct based on character of document 1: ", d1_sign_5)
print("Min hash signature of 5-shingle construct based on character of document 2: ", d2_sign_5)
print("Min hash signature of 5-shingle construct based on character of document 3: ", d3_sign_5)
print("Min hash signature of 5-shingle construct based on character of document 4: ", d4_sign_5)
print("-------------------------------------------------------------------------------------------------")

# signatures for 8-shingles based on characters
d1_sign_8 = minhash_signatures(10, d1_shingles_8)
```

```
d2_sign_8 = minhash_signatures(10, d2_shingles_8)
d3_sign_8 = minhash_signatures(10, d3_shingles_8)
d4_sign_8 = minhash_signatures(10, d4_shingles_8)

print("Min hash signature of 8-shingle construct based on character of document 1: ", d1_sign_8)
print("Min hash signature of 8-shingle construct based on character of document 2: ", d2_sign_8)
print("Min hash signature of 8-shingle construct based on character of document 3: ", d3_sign_8)
print("Min hash signature of 8-shingle construct based on character of document 4: ", d4_sign_8)
print("-----------------------------------------------------------------------------------------------------")

# signatures for 4-shingles based on words
d1_sign_4 = minhash_signatures(10, d1_shingles_4)
d2_sign_4 = minhash_signatures(10, d2_shingles_4)
d3_sign_4 = minhash_signatures(10, d3_shingles_4)
d4_sign_4 = minhash_signatures(10, d4_shingles_4)

print("Min hash signature of 4-shingle construct based on words of document 1: ", d1_sign_4)
print("Min hash signature of 4-shingle construct based on words of document 2: ", d2_sign_4)
print("Min hash signature of 4-shingle construct based on words of document 3: ", d3_sign_4)
print("Min hash signature of 4-shingle construct based on words of document 4: ", d4_sign_4)
```

```
Min hash signature of 5-shingle construct based on character of document 1:  [0, 0, 1, 0, 2, 0, 1, 7, 0, 5]
Min hash signature of 5-shingle construct based on character of document 2:  [0, 0, 2, 3, 1, 0, 0, 2, 6, 7]
Min hash signature of 5-shingle construct based on character of document 3:  [2, 1, 3, 2, 2, 0, 1, 7, 9, 0]
Min hash signature of 5-shingle construct based on character of document 4:  [1, 1, 2, 1, 0, 0, 2, 3, 1, 0]
-----------------------------------------------------------------------------------------------------
Min hash signature of 8-shingle construct based on character of document 1:  [1, 1, 1, 0, 0, 2, 2, 0, 0, 0]
Min hash signature of 8-shingle construct based on character of document 2:  [4, 1, 2, 1, 0, 0, 0, 6, 2, 1]
Min hash signature of 8-shingle construct based on character of document 3:  [3, 2, 2, 8, 6, 0, 8, 0, 2, 1]
Min hash signature of 8-shingle construct based on character of document 4:  [0, 2, 8, 1, 0, 0, 0, 5, 1, 1]
-----------------------------------------------------------------------------------------------------
Min hash signature of 4-shingle construct based on words of document 1:  [0, 11, 4, 4, 4, 7, 12, 0, 1, 3]
Min hash signature of 4-shingle construct based on words of document 2:  [5, 8, 2, 2, 7, 8, 13, 4, 3, 7]
Min hash signature of 4-shingle construct based on words of document 3:  [1, 8, 5, 6, 1, 1, 33, 11, 34, 26]
Min hash signature of 4-shingle construct based on words of document 4:  [2, 1, 1, 0, 7, 7, 10, 5, 2, 4]
```

```
# Jaccard distance for all pairs 5-shingles based on characters min-hash signatures
print(
    "Jaccard distance between signature 1 and signature 2 for 5-shingling based on characters: ",
```

```
    jaccard_dist(d1_sign_5, d2_sign_5),
)
print(
    "Jaccard distance between signature 1 and signature 3 for 5-shingling based on characters: ",
    jaccard_dist(d1_sign_5, d3_sign_5),
)
print(
    "Jaccard distance between signature 1 and signature 4 for 5-shingling based on characters: ",
    jaccard_dist(d1_sign_5, d4_sign_5),
)
print(
    "Jaccard distance between signature 2 and signature 3 for 5-shingling based on characters: ",
    jaccard_dist(d2_sign_5, d3_sign_5),
)
print(
    "Jaccard distance between signature 2 and signature 4 for 5-shingling based on characters: ",
    jaccard_dist(d2_sign_5, d4_sign_5),
)
print(
    "Jaccard distance between signature 3 and signature 4 for 5-shingling based on characters: ",
    jaccard_dist(d3_sign_5, d4_sign_5),
)
print(
    "-----------------------------------------------------------------------------------------------------"
)

# Jaccard distance for all pairs 8-shingles based on characters min-hash signatures
print(
    "Jaccard distance between signature 1 and signature 2 for 8-shingling based on characters: ",
    jaccard_dist(d1_sign_8, d2_sign_8),
)
print(
    "Jaccard distance between signature 1 and signature 3 for 8-shingling based on characters: ",
    jaccard_dist(d1_sign_8, d3_sign_8),
)
print(
    "Jaccard distance between signature 1 and signature 4 for 8-shingling based on characters: ",
    jaccard_dist(d1_sign_8, d4_sign_8),
)
```

```
print(
    "Jaccard distance between signature 2 and signature 3 for 8-shingling based on characters: ",
    jaccard_dist(d2_sign_8, d3_sign_8),
)
print(
    "Jaccard distance between signature 2 and signature 4 for 8-shingling based on characters: ",
    jaccard_dist(d2_sign_8, d4_sign_8),
)
print(
    "Jaccard distance between signature 3 and signature 4 for 8-shingling based on characters: ",
    jaccard_dist(d3_sign_8, d4_sign_8),
)
print(
    "-------------------------------------------------------------------------------------------------------"
)

# Jaccard distance for all pairs 4-shingles based on words min-hash signatures
print(
    "Jaccard distance between signature 1 and signature 2 for 4-shingling based on words: ",
    jaccard_dist(d1_sign_4, d2_sign_4),
)
print(
    "Jaccard distance between signature 1 and signature 3 for 4-shingling based on words: ",
    jaccard_dist(d1_sign_4, d3_sign_4),
)
print(
    "Jaccard distance between signature 1 and signature 4 for 4-shingling based on words: ",
    jaccard_dist(d1_sign_4, d4_sign_4),
)
print(
    "Jaccard distance between signature 2 and signature 3 for 4-shingling based on words: ",
    jaccard_dist(d2_sign_4, d3_sign_4),
)
print(
    "Jaccard distance between signature 2 and signature 4 for 4-shingling based on words: ",
    jaccard_dist(d2_sign_4, d4_sign_4),
)
print(
    "Jaccard distance between signature 3 and signature 4 for 4-shingling based on words: ",
```

```
    jaccard_dist(d3_sign_4, d4_sign_4),
)
```

```
Jaccard distance between signature 1 and signature 2 for 5-shingling based on characters:  0.4285714285714286
Jaccard distance between signature 1 and signature 3 for 5-shingling based on characters:  0.4285714285714286
Jaccard distance between signature 1 and signature 4 for 5-shingling based on characters:  0.5
Jaccard distance between signature 2 and signature 3 for 5-shingling based on characters:  0.2857142857142857
Jaccard distance between signature 2 and signature 4 for 5-shingling based on characters:  0.3333333333333333
Jaccard distance between signature 3 and signature 4 for 5-shingling based on characters:  0.3333333333333333
--------------------------------------------------------------------------------------------------------
Jaccard distance between signature 1 and signature 2 for 8-shingling based on characters:  0.4
Jaccard distance between signature 1 and signature 3 for 8-shingling based on characters:  0.5
Jaccard distance between signature 1 and signature 4 for 8-shingling based on characters:  0.4
Jaccard distance between signature 2 and signature 3 for 8-shingling based on characters:  0.4285714285714286
Jaccard distance between signature 2 and signature 4 for 8-shingling based on characters:  0.5714285714285714
Jaccard distance between signature 3 and signature 4 for 8-shingling based on characters:  0.4285714285714286
--------------------------------------------------------------------------------------------------------
Jaccard distance between signature 1 and signature 2 for 4-shingling based on words:  0.7272727272727273
Jaccard distance between signature 1 and signature 3 for 4-shingling based on words:  0.8461538461538461
Jaccard distance between signature 1 and signature 4 for 4-shingling based on words:  0.6
Jaccard distance between signature 2 and signature 3 for 4-shingling based on words:  0.8461538461538461
Jaccard distance between signature 2 and signature 4 for 4-shingling based on words:  0.6
Jaccard distance between signature 3 and signature 4 for 4-shingling based on words:  0.8461538461538461
```

Conclusion

I generated the different k-shingles specified in the tasks & calculated the Jaccard and Levenshtein distance between the generated k-shingles. The above experiment was also performed for the Indian Language - Kannada. Then I made a function to generate the min-hash signature of the above generated shingles. Min-hashing is the process of compressing generated sets of unique shingles into smaller representations called 'signatures'. Larger the value of hashes(number of permutations), larger the size of the signature and more accurate representation of the shingles. The purpose of using min-hashing is to get a faster approximation of the jaccard distance but it is not completely accurate.

✓ 0s completed at 12:48 AM