IT350 - Lab 2

Annam Indhu Lekha
191IT207

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

Document Name	5 character shingle	8 character shingle	4 word shingle
13chil.txt	9056	26329	1440
3lpigs.txt	7108	19215	984
3wishes.txt	5835	15449	754
6ablemen.txt	8399	24112	1273

5 character shingle:

```
Filename : 13chil.txt
Size of 5 Character Shingle set : 9056
5 Character Shingle set : {'e rep', 'oli', 'kly j', 'ckl', 'ook o', 'y don', 'eili', 'news'

Filename : 3lpigs.txt
Size of 5 Character Shingle set : 7108
5 Character Shingle set : {'oli', 'reall', 'ckl', 'y don', 'fell', 'y ope', 'nto ', ' big',

Filename : 3wishes.txt
Size of 5 Character Shingle set : 5835
5 Character Shingle set : {'reall', 'ckl', 'ed wi', 's ba', 'nto ', ' big', 'ooks ', ' m',

Filename : 6ablemen.txt
Size of 5 Character Shingle set : 8399
5 Character Shingle set : {'lley ', 'ook o', 'ckl', 'fell', 'vell', 'ose h', 'dly i', 'ed v
```

8 character shingle:

```
Filename : 13chil.txt
Size of 8 Character Shingle set : 26329
8 Character Shingle set : {'ng pap', 'irly ', 'loving', 'reath', 'l ove', 'do snif'

Filename : 3lpigs.txt
Size of 8 Character Shingle set : 19215
8 Character Shingle set : {'i want ', 't set to', 'he la', 's he co', 'e heap o', '

Filename : 3wishes.txt
Size of 8 Character Shingle set : 15449
8 Character Shingle set : {'couple ', ' in or', ' alwa', ' stutte', 'nd si', 'e elf

Filename : 6ablemen.txt
Size of 8 Character Shingle set : 24112
8 Character Shingle set : {'couple ', 'i want ', 'n retu', 'l ove', 'nly f', 't vi'
```

4 word shingle:

```
Filename: 13chil.txt
Length of 4 word shingle list before converting to set: 1448
Length of 4 word shingle list after converting to set: 1440
Filename: 3lpigs.txt
Length of 4 word shingle list before converting to set: 994
Length of 4 word shingle list after converting to set: 984
Filename: 3wishes.txt
Length of 4 word shingle list before converting to set: 757
Length of 4 word shingle list after converting to set: 754
Filename: 6ablemen.txt
Length of 4 word shingle list before converting to set: 1276
Length of 4 word shingle list after converting to set: 1273
```

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

```
5 Character Shingles Jaccard Similarity
Comparing file: 13chil.txt
Jaccard similarity with file13chil.txt: 1
Jaccard similarity with file3lpigs.txt: 0.2660766037440276
Jaccard similarity with file3wishes.txt: 0.24882589734988259
Jaccard similarity with file6ablemen.txt: 0.2707483983692487


Comparing file: 3lpigs.txt
Jaccard similarity with file13chil.txt: 0.2660766037440276
Jaccard similarity with file3lpigs.txt: 1
Jaccard similarity with file3wishes.txt: 0.25696804894629505
Jaccard similarity with file6ablemen.txt: 0.2590938616433907


Comparing file: 3wishes.txt
Jaccard similarity with file13chil.txt: 0.24882589734988259
Jaccard similarity with file3lpigs.txt: 0.25696804894629505
Jaccard similarity with file3wishes.txt: 1
Jaccard similarity with file6ablemen.txt: 0.2439045704797693


Comparing file: 6ablemen.txt
Jaccard similarity with file13chil.txt: 0.2707483983692487
Jaccard similarity with file3lpigs.txt: 0.2590938616433907
Jaccard similarity with file3wishes.txt: 0.2439045704797693
Jaccard similarity with file6ablemen.txt: 1
```

```
8 Character Shingles Jaccard Similarity
Comparing file: 13chil.txt
Jaccard similarity with file13chil.txt: 1
Jaccard similarity with file3lpigs.txt: 0.12243690851735016
Jaccard similarity with file3wishes.txt: 0.11637228442401731
Jaccard similarity with file6ablemen.txt: 0.12646835499575684


Comparing file: 3lpigs.txt
Jaccard similarity with file13chil.txt: 0.12243690851735016
Jaccard similarity with file3lpigs.txt: 1
Jaccard similarity with file3wishes.txt: 0.12010857272110383
Jaccard similarity with file6ablemen.txt: 0.12228669118789826


Comparing file: 3wishes.txt
Jaccard similarity with file13chil.txt: 0.11637228442401731
Jaccard similarity with file3lpigs.txt: 0.12010857272110383
Jaccard similarity with file3wishes.txt: 1
Jaccard similarity with file6ablemen.txt: 0.11486543609976047


Comparing file: 6ablemen.txt
Jaccard similarity with file13chil.txt: 0.12646835499575684
Jaccard similarity with file3lpigs.txt: 0.12228669118789826
Jaccard similarity with file3wishes.txt: 0.11486543609976047
Jaccard similarity with file6ablemen.txt: 1
```

```
 4 Word Shingles Jaccard Similarity
 Comparing file: 13chil.txt
 Jaccard similarity with file13chil.txt: 1
 Jaccard similarity with file3lpigs.txt: 0.0012391573729863693
 Jaccard similarity with file3wishes.txt: 0.0013692377909630307
 Jaccard similarity with file6ablemen.txt: 0.0


 Comparing file: 3lpigs.txt
 Jaccard similarity with file13chil.txt: 0.0012391573729863693
 Jaccard similarity with file3lpigs.txt: 1
 Jaccard similarity with file3wishes.txt: 0.0005757052389176742
 Jaccard similarity with file6ablemen.txt: 0.0008869179600886918


 Comparing file: 3wishes.txt
 Jaccard similarity with file13chil.txt: 0.0013692377909630307
 Jaccard similarity with file3lpigs.txt: 0.0005757052389176742
 Jaccard similarity with file3wishes.txt: 1
 Jaccard similarity with file6ablemen.txt: 0.0014822134387351778


 Comparing file: 6ablemen.txt
 Jaccard similarity with file13chil.txt: 0.0
 Jaccard similarity with file3lpigs.txt: 0.0008869179600886918
 Jaccard similarity with file3wishes.txt: 0.0014822134387351778
 Jaccard similarity with file6ablemen.txt: 1
```

Change to any Similarity Function and check the distance - Using cosine similarity

```
5 Character Shingles Cosine Similarity
Comparing file: 13chil.txt
cosine similarity with file13chil.txt: 1
cosine similarity with file3lpigs.txt: 0.42340270705178207
cosine similarity with file3wishes.txt: 0.40815856013554747
cosine similarity with file6ablemen.txt: 0.426426495089597


Comparing file: 3lpigs.txt
cosine similarity with file13chil.txt: 0.42340270705178207
cosine similarity with file3lpigs.txt: 1
cosine similarity with file3wishes.txt: 0.4108617390898646
cosine similarity with file6ablemen.txt: 0.4129897794499623


Comparing file: 3wishes.txt
cosine similarity with file13chil.txt: 0.40815856013554747
cosine similarity with file3lpigs.txt: 0.4108617390898646
cosine similarity with file3wishes.txt: 1
cosine similarity with file6ablemen.txt: 0.3986810704556671


Comparing file: 6ablemen.txt
cosine similarity with file13chil.txt: 0.426426495089597
cosine similarity with file3lpigs.txt: 0.4129897794499623
cosine similarity with file3wishes.txt: 0.3986810704556671
cosine similarity with file6ablemen.txt: 1
```

```
8 Character Shingles Cosine Similarity
Comparing file: 13chil.txt
cosine similarity with file13chil.txt: 1
cosine similarity with file3lpigs.txt: 0.22087380381427824
cosine similarity with file3wishes.txt: 0.2159338810283226
cosine similarity with file6ablemen.txt: 0.22475675949656504


Comparing file: 3lpigs.txt
cosine similarity with file13chil.txt: 0.22087380381427824
cosine similarity with file3lpigs.txt: 1
cosine similarity with file3wishes.txt: 0.21573578042070682
cosine similarity with file6ablemen.txt: 0.21932957068880848


Comparing file: 3wishes.txt
cosine similarity with file13chil.txt: 0.2159338810283226
cosine similarity with file3lpigs.txt: 0.21573578042070682
cosine similarity with file3wishes.txt: 1
cosine similarity with file6ablemen.txt: 0.21118709256924872


Comparing file: 6ablemen.txt
cosine similarity with file13chil.txt: 0.22475675949656504
cosine similarity with file3lpigs.txt: 0.21932957068880848
cosine similarity with file3wishes.txt: 0.21118709256924872
cosine similarity with file6ablemen.txt: 1
```

```
4 Word Shingles Cosine Similarity
Comparing file: 13chil.txt
cosine similarity with file13chil.txt: 1
cosine similarity with file3lpigs.txt: 0.0025202432454547244
cosine similarity with file3wishes.txt: 0.002879083998155492
cosine similarity with file6ablemen.txt: 0.0


Comparing file: 3lpigs.txt
cosine similarity with file13chil.txt: 0.0025202432454547244
cosine similarity with file3lpigs.txt: 1
cosine similarity with file3wishes.txt: 0.0011609587199117057
cosine similarity with file6ablemen.txt: 0.0017869740450141421


Comparing file: 3wishes.txt
cosine similarity with file13chil.txt: 0.002879083998155492
cosine similarity with file3lpigs.txt: 0.0011609587199117057
cosine similarity with file3wishes.txt: 1
cosine similarity with file6ablemen.txt: 0.003062114175327031


Comparing file: 6ablemen.txt
cosine similarity with file13chil.txt: 0.0
cosine similarity with file3lpigs.txt: 0.0017869740450141421
cosine similarity with file3wishes.txt: 0.003062114175327031
cosine similarity with file6ablemen.txt: 1
```

Trying all the above for another Indian language - Hindi

```
Filename : hindi1.txt
Size of 5 Character Shingle set : 1796
5 Character Shingle set : {'एग', 'तक ह ', ' जतग', 'कह आइ', ' थ कय', ' और स', 'भ नह ', ' फन', ' पस ब

Filename : hindi2.txt
Size of 5 Character Shingle set : 3529
5 Character Shingle set : {'मल ह', 'प बस ', 'लय च', 'तक ह ', 'क आ', ' स हर', '000 ', 'षट ह ', ' और स

Filename : hindi3.txt
Size of 5 Character Shingle set : 2823
5 Character Shingle set : {'एग', 'ध कम ', ' लभ', 'क आ', 'यकत ', 'सखय क', ' पस ब', 'भकभ ', 'द करत',

Filename : hindi4.txt
Size of 5 Character Shingle set : 2093
5 Character Shingle set : {' न उस', ' न प', 'क आ', 'सक ', 'ड वह', ' चहत', 'नम', 'क हमम', 'ह गई', ' थ
```

```
Filename : hindi1.txt
Size of 8 Character Shingle set : 3965
8 Character Shingle set : {'एग', 'कषय क ओर', 'तक ह ', ' खन खन क', ' कर और', ' थ कय', ' और स

Filename : hindi2.txt
Size of 8 Character Shingle set : 8694
8 Character Shingle set : {'न नह ऐस', 'आओ मझ ', ' मर मदद', ' बरतन क', ' गव क स', 'त थ क सन',

Filename : hindi3.txt
Size of 8 Character Shingle set : 6487
8 Character Shingle set : {'एग', ' लभ', 'क आ', 'ह सकत ह', 'यकत ', ' पस ब', 'जन वल थ', 'धमकय

Filename : hindi4.txt
Size of 8 Character Shingle set : 4630
8 Character Shingle set : {'त न उस', ' मस न पछ', ' क दखन ', 'क आ', 'हलत रह', ' बत कय न', 'ड वह
```

```
Filename: hindi1.txt
Length of 4 word shingle list before converting to set: 238
Length of 4 word shingle list after converting to set: 238
Filename: hindi2.txt
Length of 4 word shingle list before converting to set: 710
Length of 4 word shingle list after converting to set: 640
Filename: hindi3.txt
Length of 4 word shingle list before converting to set: 428
Length of 4 word shingle list after converting to set: 424
Filename: hindi4.txt
Length of 4 word shingle list before converting to set: 345
Length of 4 word shingle list after converting to set: 294
```

```
5 Character Shingles Jaccard Similarity
Comparing file: hindi1.txt
Jaccard similarity with filehindi1.txt: 1
Jaccard similarity with filehindi2.txt: 0.12058080808080808
Jaccard similarity with filehindi3.txt: 0.13516834603096584
Jaccard similarity with filehindi4.txt: 0.13315850815850816


Comparing file: hindi2.txt
Jaccard similarity with filehindi1.txt: 0.12058080808080808
Jaccard similarity with filehindi2.txt: 1
Jaccard similarity with filehindi3.txt: 0.1670034907220283
Jaccard similarity with filehindi4.txt: 0.13713592233009708


Comparing file: hindi3.txt
Jaccard similarity with filehindi1.txt: 0.13516834603096584
Jaccard similarity with filehindi2.txt: 0.1670034907220283
Jaccard similarity with filehindi3.txt: 1
Jaccard similarity with filehindi4.txt: 0.1371732593106639


Comparing file: hindi4.txt
Jaccard similarity with filehindi1.txt: 0.13315850815850816
Jaccard similarity with filehindi2.txt: 0.13713592233009708
Jaccard similarity with filehindi3.txt: 0.1371732593106639
Jaccard similarity with filehindi4.txt: 1
```

```
8 Character Shingles Jaccard Similarity
Comparing file: hindi1.txt
Jaccard similarity with filehindi1.txt: 1
Jaccard similarity with filehindi2.txt: 0.05377507699991676
Jaccard similarity with filehindi3.txt: 0.06273512963904423
Jaccard similarity with filehindi4.txt: 0.061242128657858996


Comparing file: hindi2.txt
Jaccard similarity with filehindi1.txt: 0.05377507699991676
Jaccard similarity with filehindi2.txt: 1
Jaccard similarity with filehindi3.txt: 0.07628500531726339
Jaccard similarity with filehindi4.txt: 0.05897313622635511


Comparing file: hindi3.txt
Jaccard similarity with filehindi1.txt: 0.06273512963904423
Jaccard similarity with filehindi2.txt: 0.07628500531726339
Jaccard similarity with filehindi3.txt: 1
Jaccard similarity with filehindi4.txt: 0.06250597343018255


Comparing file: hindi4.txt
Jaccard similarity with filehindi1.txt: 0.061242128657858996
Jaccard similarity with filehindi2.txt: 0.05897313622635511
Jaccard similarity with filehindi3.txt: 0.06250597343018255
Jaccard similarity with filehindi4.txt: 1
```

```
4 Word Shingles Jaccard Similarity
Comparing file: hindi1.txt
Jaccard similarity with filehindi1.txt: 1
Jaccard similarity with filehindi2.txt: 0.0011402508551881414
Jaccard similarity with filehindi3.txt: 0.0
Jaccard similarity with filehindi4.txt: 0.0


Comparing file: hindi2.txt
Jaccard similarity with filehindi1.txt: 0.0011402508551881414
Jaccard similarity with filehindi2.txt: 1
Jaccard similarity with filehindi3.txt: 0.0
Jaccard similarity with filehindi4.txt: 0.0


Comparing file: hindi3.txt
Jaccard similarity with filehindi1.txt: 0.0
Jaccard similarity with filehindi2.txt: 0.0
Jaccard similarity with filehindi3.txt: 1
Jaccard similarity with filehindi4.txt: 0.0


Comparing file: hindi4.txt
Jaccard similarity with filehindi1.txt: 0.0
Jaccard similarity with filehindi2.txt: 0.0
Jaccard similarity with filehindi3.txt: 0.0
Jaccard similarity with filehindi4.txt: 1
```

```
5 Character Shingles Cosine Similarity
Comparing file: hindi1.txt
cosine similarity with filehindi1.txt: 1
cosine similarity with filehindi2.txt: 0.22760176554010386
cosine similarity with filehindi3.txt: 0.2442609738440635
cosine similarity with filehindi4.txt: 0.2357102235395571


Comparing file: hindi2.txt
cosine similarity with filehindi1.txt: 0.22760176554010386
cosine similarity with filehindi2.txt: 1
cosine similarity with filehindi3.txt: 0.2879934529874923
cosine similarity with filehindi4.txt: 0.2494705304687703


Comparing file: hindi3.txt
cosine similarity with filehindi1.txt: 0.2442609738440635
cosine similarity with filehindi2.txt: 0.2879934529874923
cosine similarity with filehindi3.txt: 1
cosine similarity with filehindi4.txt: 0.24395776172651423


Comparing file: hindi4.txt
cosine similarity with filehindi1.txt: 0.2357102235395571
cosine similarity with filehindi2.txt: 0.2494705304687703
cosine similarity with filehindi3.txt: 0.24395776172651423
cosine similarity with filehindi4.txt: 1
```

```
8 Character Shingles Cosine Similarity
Comparing file: hindi1.txt
cosine similarity with filehindi1.txt: 1
cosine similarity with filehindi2.txt: 0.1100274642859895
cosine similarity with filehindi3.txt: 0.12165826518383219
cosine similarity with filehindi4.txt: 0.11576294992758161


Comparing file: hindi2.txt
cosine similarity with filehindi1.txt: 0.1100274642859895
cosine similarity with filehindi2.txt: 1
cosine similarity with filehindi3.txt: 0.1432783278706205
cosine similarity with filehindi4.txt: 0.11695090634318579


Comparing file: hindi3.txt
cosine similarity with filehindi1.txt: 0.12165826518383219
cosine similarity with filehindi2.txt: 0.1432783278706205
cosine similarity with filehindi3.txt: 1
cosine similarity with filehindi4.txt: 0.1193343038227822


Comparing file: hindi4.txt
cosine similarity with filehindi1.txt: 0.11576294992758161
cosine similarity with filehindi2.txt: 0.11695090634318579
cosine similarity with filehindi3.txt: 0.1193343038227822
cosine similarity with filehindi4.txt: 1
```

```
4 Word Shingles Cosine Similarity
Comparing file: hindi1.txt
cosine similarity with filehindi1.txt: 1
cosine similarity with filehindi2.txt: 0.0025622501927837116
cosine similarity with filehindi3.txt: 0.0
cosine similarity with filehindi4.txt: 0.0


Comparing file: hindi2.txt
cosine similarity with filehindi1.txt: 0.0025622501927837116
cosine similarity with filehindi2.txt: 1
cosine similarity with filehindi3.txt: 0.0
cosine similarity with filehindi4.txt: 0.0


Comparing file: hindi3.txt
cosine similarity with filehindi1.txt: 0.0
cosine similarity with filehindi2.txt: 0.0
cosine similarity with filehindi3.txt: 1
cosine similarity with filehindi4.txt: 0.0


Comparing file: hindi4.txt
cosine similarity with filehindi1.txt: 0.0
cosine similarity with filehindi2.txt: 0.0
cosine similarity with filehindi3.txt: 0.0
cosine similarity with filehindi4.txt: 1
```

Using Min Hash:

```
5 Character Shingles Cosine Similarity using a different min hash function

Comparing file: 13chil.txt
Jaccard Similarity using min hash with file : 13chil.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.296875 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.296875 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.21875 after min hashing

Comparing file: 3lpigs.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.296875 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.3125 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.2421875 after min hashing

Comparing file: 3wishes.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.296875 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.3125 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.1796875 after min hashing

Comparing file: 6ablemen.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.21875 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.2421875 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.1796875 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 1.0 after min hashing
```

```
8 Character Shingles Cosine Similarity using a different min hash function

Comparing file: 13chil.txt
Jaccard Similarity using min hash with file : 13chil.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.140625 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.1171875 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.1328125 after min hashing

Comparing file: 3lpigs.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.140625 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.1875 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.1015625 after min hashing

Comparing file: 3wishes.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.1171875 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.1875 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.125 after min hashing

Comparing file: 6ablemen.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.1328125 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.1015625 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.125 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 1.0 after min hashing
```

```
4 Word Shingles Cosine Similarity using a different min hash function

Comparing file: 13chil.txt
Jaccard Similarity using min hash with file : 13chil.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.0078125 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.0 after min hashing

Comparing file: 3lpigs.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.0078125 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.0 after min hashing

Comparing file: 3wishes.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 0.0 after min hashing

Comparing file: 6ablemen.txt
Jaccard Similarity using min hash with file : 13chil.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 3lpigs.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 3wishes.txt is 0.0 after min hashing
Jaccard Similarity using min hash with file : 6ablemen.txt is 1.0 after min hashing
```

```
5 Character Shingles Cosine Similarity using a different min hash function

Comparing file: hindi1.txt
Jaccard Similarity using min hash with file : hindi1.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 0.1015625 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 0.1328125 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 0.140625 after min hashing

Comparing file: hindi2.txt
Jaccard Similarity using min hash with file : hindi1.txt is 0.1015625 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 0.171875 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 0.1171875 after min hashing

Comparing file: hindi3.txt
Jaccard Similarity using min hash with file : hindi1.txt is 0.1328125 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 0.171875 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 0.1328125 after min hashing

Comparing file: hindi4.txt
Jaccard Similarity using min hash with file : hindi1.txt is 0.140625 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 0.1171875 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 0.1328125 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 1.0 after min hashing
```

```
8 Character Shingles Cosine Similarity using a different min hash function

Comparing file: hindi1.txt
Jaccard Similarity using min hash with file : hindi1.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 0.0234375 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 0.0546875 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 0.0546875 after min hashing

Comparing file: hindi2.txt
Jaccard Similarity using min hash with file : hindi1.txt is 0.0234375 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 0.046875 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 0.0390625 after min hashing

Comparing file: hindi3.txt
Jaccard Similarity using min hash with file : hindi1.txt is 0.0546875 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 0.046875 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 1.0 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 0.0703125 after min hashing

Comparing file: hindi4.txt
Jaccard Similarity using min hash with file : hindi1.txt is 0.0546875 after min hashing
Jaccard Similarity using min hash with file : hindi2.txt is 0.0390625 after min hashing
Jaccard Similarity using min hash with file : hindi3.txt is 0.0703125 after min hashing
Jaccard Similarity using min hash with file : hindi4.txt is 1.0 after min hashing
```

```
 4 Word Shingles Cosine Similarity using a different min hash function

 Comparing file: hindi1.txt
 Jaccard Similarity using min hash with file : hindi1.txt is 1.0 after min hashing
 Jaccard Similarity using min hash with file : hindi2.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi3.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi4.txt is 0.0 after min hashing

 Comparing file: hindi2.txt
 Jaccard Similarity using min hash with file : hindi1.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi2.txt is 1.0 after min hashing
 Jaccard Similarity using min hash with file : hindi3.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi4.txt is 0.0 after min hashing

 Comparing file: hindi3.txt
 Jaccard Similarity using min hash with file : hindi1.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi2.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi3.txt is 1.0 after min hashing
 Jaccard Similarity using min hash with file : hindi4.txt is 0.0 after min hashing

 Comparing file: hindi4.txt
 Jaccard Similarity using min hash with file : hindi1.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi2.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi3.txt is 0.0 after min hashing
 Jaccard Similarity using min hash with file : hindi4.txt is 1.0 after min hashing
```

Conclusion:

1. **4 word similarity is more strict when compared to 8 character which is again more strict when compared to 5 character similarity metrics**
2. **To find the Jaccard Similarity of the documents, we need to first create the shingles of the documents and then using this, we can find out the Jaccard similarity.**
3. **Further, the process of finding the Jaccard similarity using shingles will at best case take O(n^2)**
4. **If the number of shingles is around a Million (10^6), it will take days to complete, therefore a better method is required. Hence the use of Min-hashing makes the process faster.**
5. **Tasks like Finding duplicates or Similar documents becomes more scalable because brute force solution is very slow**
6. **Right representation of document for efficient similarity comparison**
7. **Its valid for any language**