IT350 Assignment 2

NAME: SUYASH CHINTAWAR
ROLL NO.: 191IT109
TOPIC: SHINGLING AND
MIN-HASHING

Note:
1) The colab link has been attached below. After opening the link, if it opens in drive, click on “Open with Google Colaboratory” to view the complete code.
2) Only output screenshots have been attached along with the explanation. Code for the same can be found in the colab notebook.

Colab notebook link:
https://colab.research.google.com/drive/1xlxEck-kVEnQD6pGo-ebx7IYxA9iXOwJ

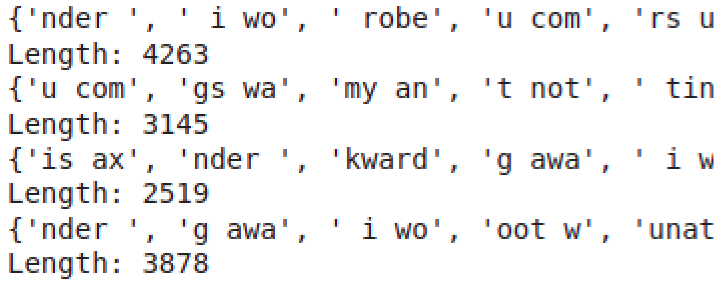
The english text files are downloaded from the links below,
1. http://www.textfiles.com/stories/13chil.txt
2. http://www.textfiles.com/stories/3wishes.txt
3. http://www.textfiles.com/stories/3lpigs.txt
4. http://www.textfiles.com/stories/6ablemen.txt

Q1. How many distinct shingles are there for each document with each type of shingle?

(a) 5-character shingles

'13chil.txt'	- 4263
'3lpigs.txt'	- 3145
'3wishes.txt'	- 2519
'6ablemen.txt'	- 3878

Output Screenshot:

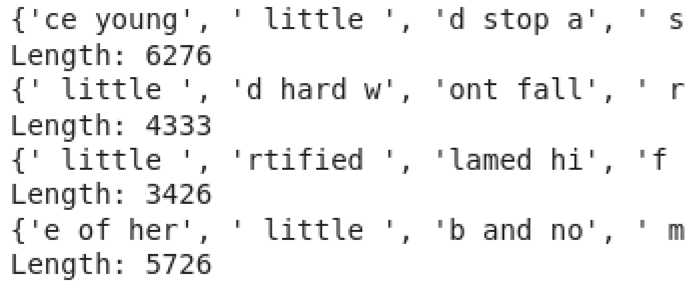


```
{'nder ', ' i wo', ' robe', 'u com', 'rs u
Length: 4263
{'u com', 'gs wa', 'my an', 't not', ' tin
Length: 3145
{'is ax', 'nder ', 'kward', 'g awa', ' i w
Length: 2519
{'nder ', 'g awa', ' i wo', 'oot w', 'unat
Length: 3878
```

(b) 8-character shingles

'13chil.txt'	- 6276
'3lpigs.txt'	- 4333
'3wishes.txt'	- 3426
'6ablemen.txt'	- 5276

Output Screenshot:

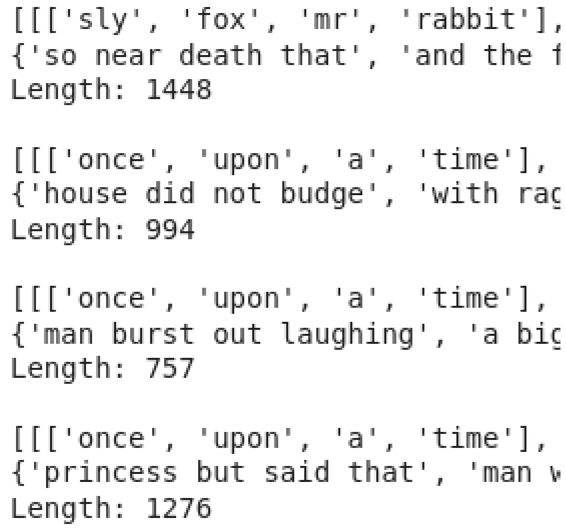


```
{'ce young', ' little ', 'd stop a', ' s
Length: 6276
{' little ', 'd hard w', 'ont fall', ' r
Length: 4333
{' little ', 'rtified ', 'lamed hi', 'f
Length: 3426
{'e of her', ' little ', 'b and no', ' m
Length: 5726
```

(c) 4-word shingles

'13chil.txt'	- 1448
'3lpigs.txt'	- 994
'3wishes.txt'	- 757
'6ablemen.txt'	- 1276

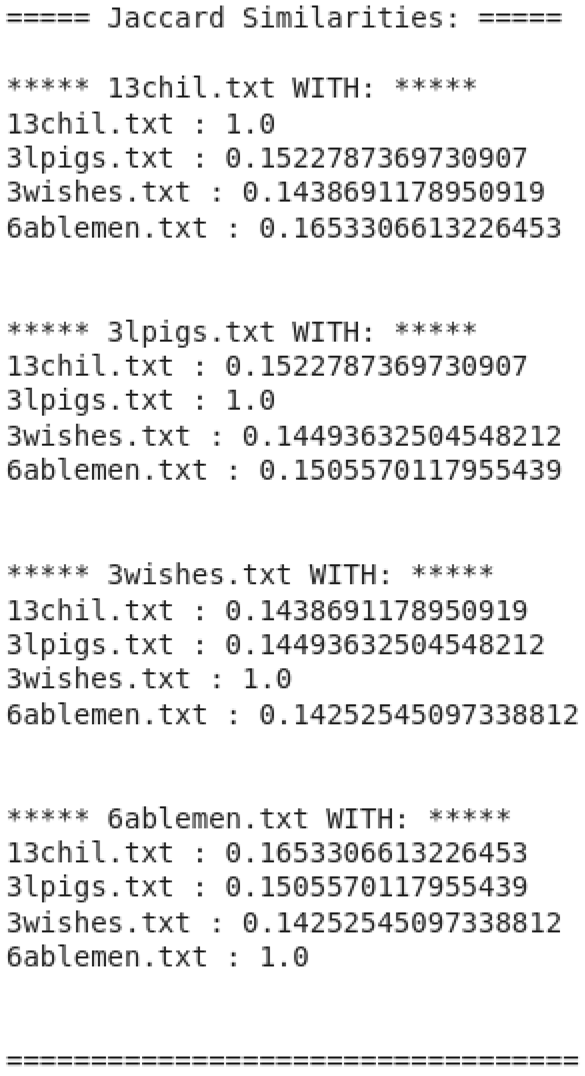
Output Screenshot:



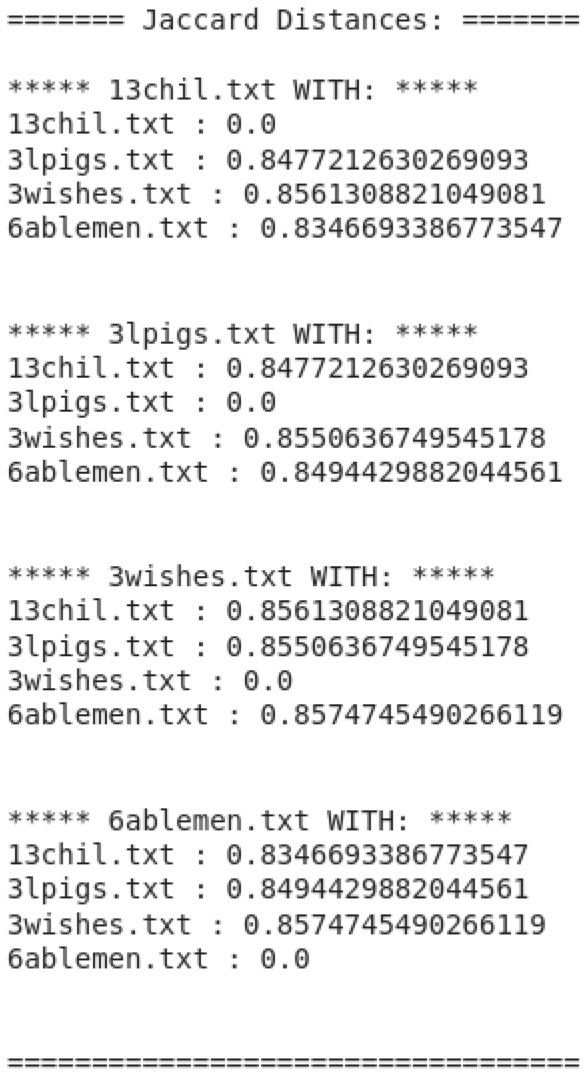
```
[[['sly', 'fox', 'mr', 'rabbit'],
{'so near death that', 'and the f
Length: 1448

[[['once', 'upon', 'a', 'time'],
{'house did not budge', 'with rag
Length: 994

[[['once', 'upon', 'a', 'time'],
{'man burst out laughing', 'a big
Length: 757

[[['once', 'upon', 'a', 'time'],
{'princess but said that', 'man w
Length: 1276
```

Q2. Compute the Jaccard distance between all pairs of documents for each type of shingling.

(a) 5-character shingles

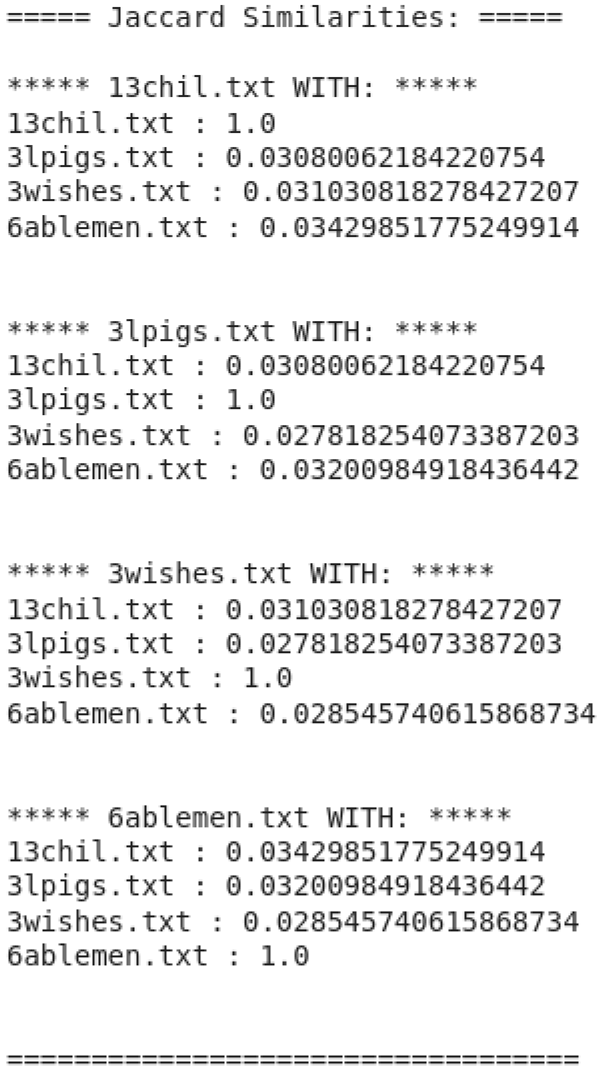


```
===== Jaccard Similarities: =====

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.1522787369730907
3wishes.txt : 0.1438691178950919
6ablemen.txt : 0.1653306613226453


***** 3lpigs.txt WITH: *****
13chil.txt : 0.1522787369730907
3lpigs.txt : 1.0
3wishes.txt : 0.14493632504548212
6ablemen.txt : 0.1505570117955439


***** 3wishes.txt WITH: *****
13chil.txt : 0.1438691178950919
3lpigs.txt : 0.14493632504548212
3wishes.txt : 1.0
6ablemen.txt : 0.14252545097338812


***** 6ablemen.txt WITH: *****
13chil.txt : 0.1653306613226453
3lpigs.txt : 0.1505570117955439
3wishes.txt : 0.14252545097338812
6ablemen.txt : 1.0


==================================
```

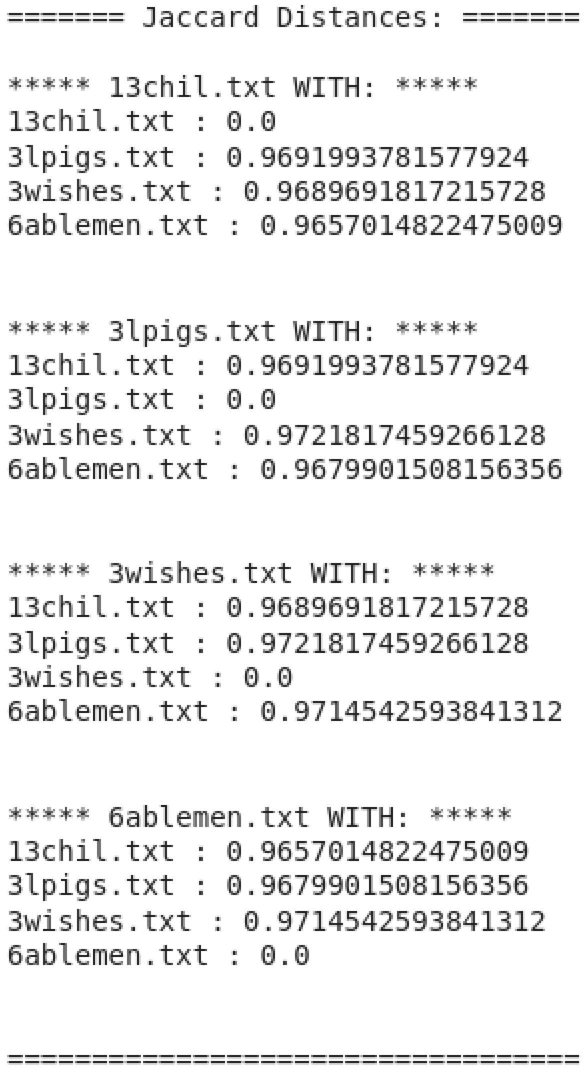


```
======= Jaccard Distances: =======

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.8477212630269093
3wishes.txt : 0.8561308821049081
6ablemen.txt : 0.8346693386773547


***** 3lpigs.txt WITH: *****
13chil.txt : 0.8477212630269093
3lpigs.txt : 0.0
3wishes.txt : 0.8550636749545178
6ablemen.txt : 0.8494429882044561


***** 3wishes.txt WITH: *****
13chil.txt : 0.8561308821049081
3lpigs.txt : 0.8550636749545178
3wishes.txt : 0.0
6ablemen.txt : 0.8574745490266119


***** 6ablemen.txt WITH: *****
13chil.txt : 0.8346693386773547
3lpigs.txt : 0.8494429882044561
3wishes.txt : 0.8574745490266119
6ablemen.txt : 0.0


==================================
```

(b) 8-character shingles

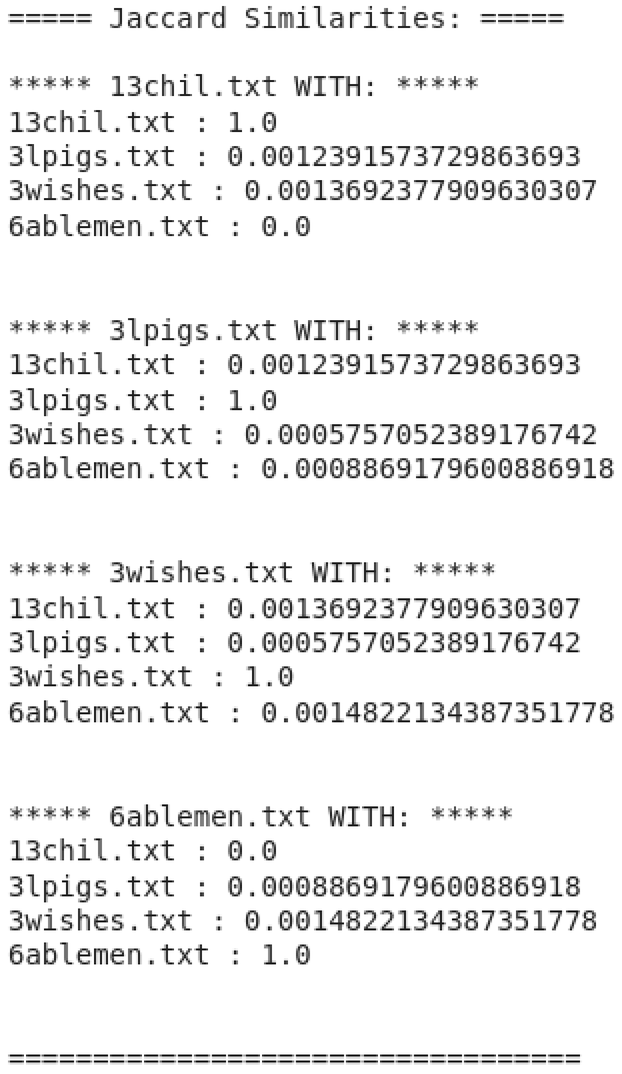


```
===== Jaccard Similarities: =====

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.03080062184220754
3wishes.txt : 0.031030818278427207
6ablemen.txt : 0.03429851775249914


***** 3lpigs.txt WITH: *****
13chil.txt : 0.03080062184220754
3lpigs.txt : 1.0
3wishes.txt : 0.027818254073387203
6ablemen.txt : 0.03200984918436442


***** 3wishes.txt WITH: *****
13chil.txt : 0.031030818278427207
3lpigs.txt : 0.027818254073387203
3wishes.txt : 1.0
6ablemen.txt : 0.028545740615868734


***** 6ablemen.txt WITH: *****
13chil.txt : 0.03429851775249914
3lpigs.txt : 0.03200984918436442
3wishes.txt : 0.028545740615868734
6ablemen.txt : 1.0

==================================
```

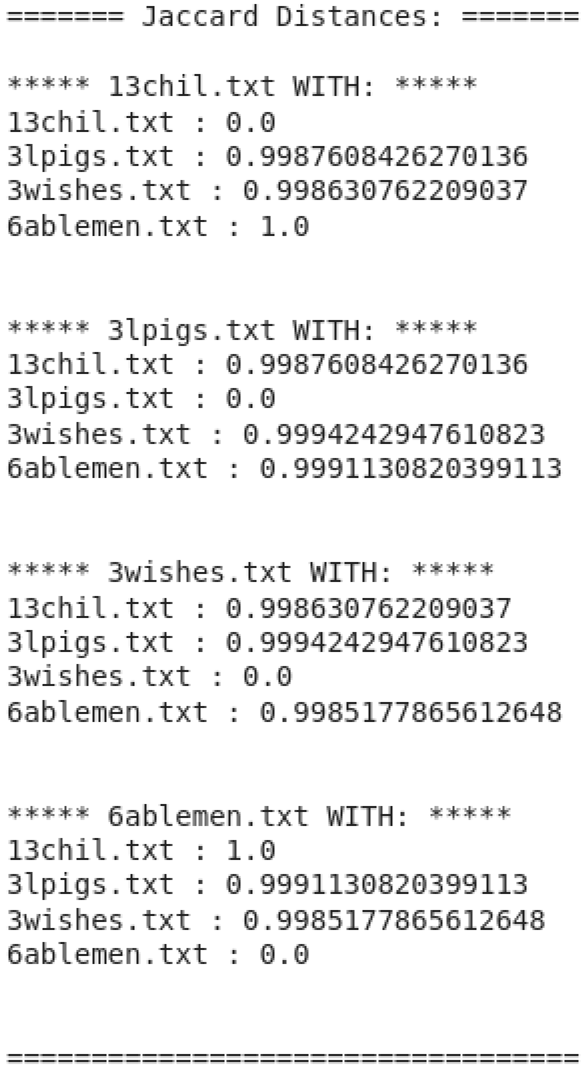


```
======= Jaccard Distances: =======

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.9691993781577924
3wishes.txt : 0.9689691817215728
6ablemen.txt : 0.9657014822475009


***** 3lpigs.txt WITH: *****
13chil.txt : 0.9691993781577924
3lpigs.txt : 0.0
3wishes.txt : 0.9721817459266128
6ablemen.txt : 0.9679901508156356


***** 3wishes.txt WITH: *****
13chil.txt : 0.9689691817215728
3lpigs.txt : 0.9721817459266128
3wishes.txt : 0.0
6ablemen.txt : 0.9714542593841312


***** 6ablemen.txt WITH: *****
13chil.txt : 0.9657014822475009
3lpigs.txt : 0.9679901508156356
3wishes.txt : 0.9714542593841312
6ablemen.txt : 0.0

==================================
```

(c) 4-word shingles



```
===== Jaccard Similarities: =====

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.0012391573729863693
3wishes.txt : 0.0013692377909630307
6ablemen.txt : 0.0


***** 3lpigs.txt WITH: *****
13chil.txt : 0.0012391573729863693
3lpigs.txt : 1.0
3wishes.txt : 0.0005757052389176742
6ablemen.txt : 0.0008869179600886918


***** 3wishes.txt WITH: *****
13chil.txt : 0.0013692377909630307
3lpigs.txt : 0.0005757052389176742
3wishes.txt : 1.0
6ablemen.txt : 0.0014822134387351778


***** 6ablemen.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.0008869179600886918
3wishes.txt : 0.0014822134387351778
6ablemen.txt : 1.0

=================================
```

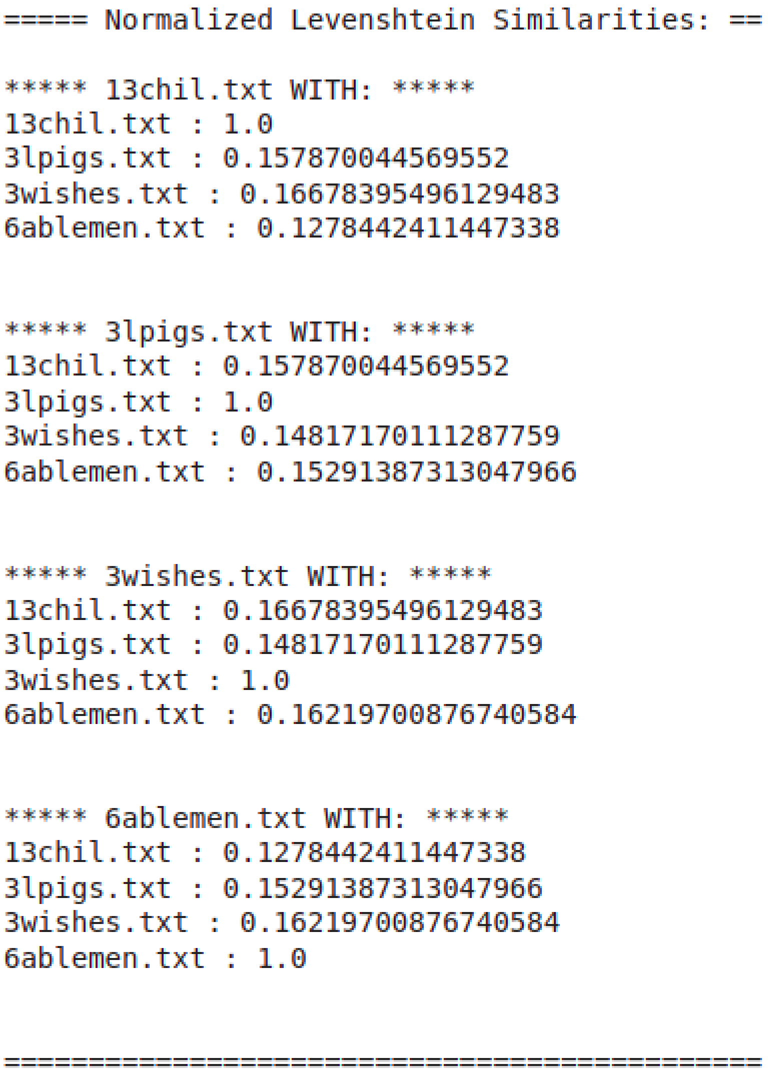


```
======= Jaccard Distances: =======

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.9987608426270136
3wishes.txt : 0.998630762209037
6ablemen.txt : 1.0


***** 3lpigs.txt WITH: *****
13chil.txt : 0.9987608426270136
3lpigs.txt : 0.0
3wishes.txt : 0.9994242947610823
6ablemen.txt : 0.9991130820399113


***** 3wishes.txt WITH: *****
13chil.txt : 0.998630762209037
3lpigs.txt : 0.9994242947610823
3wishes.txt : 0.0
6ablemen.txt : 0.9985177865612648


***** 6ablemen.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.9991130820399113
3wishes.txt : 0.9985177865612648
6ablemen.txt : 0.0

=================================
```

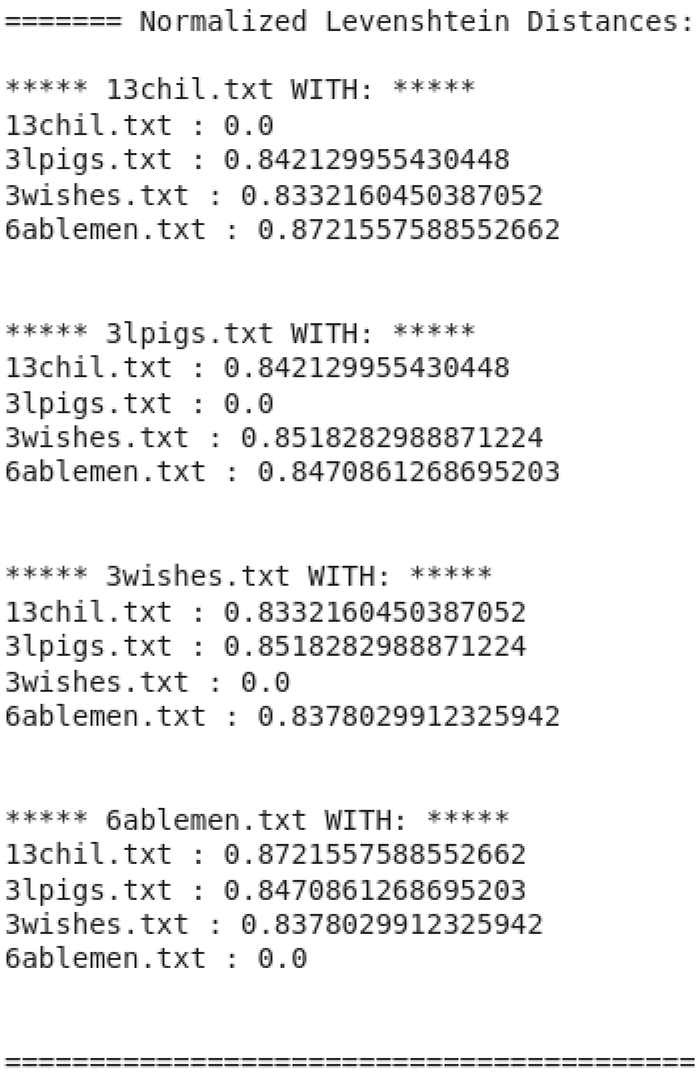
Q3. Change to any Similarity Function (use any recent similarity distance) and check the distance.

Normalized Levenshtein distance has been used here to compute similarity and distances between shingles. The library ‘`strsimpy`’ has been used for this purpose.

<u>(a) 5-character shingles</u>

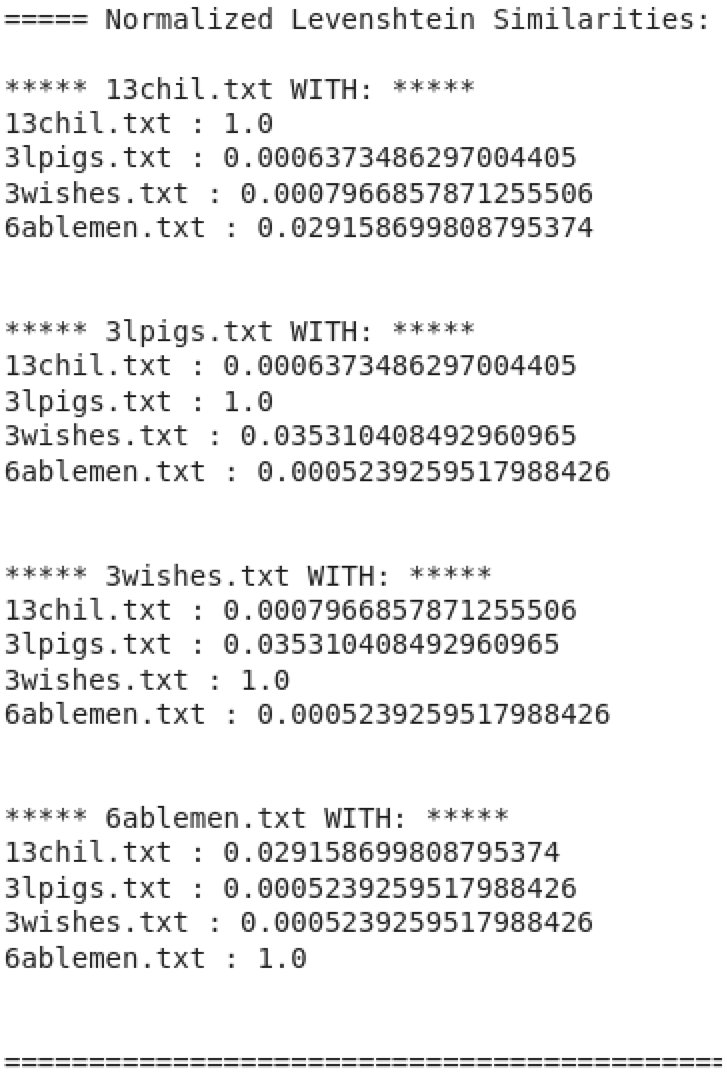


```
===== Normalized Levenshtein Similarities: ==

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.157870044569552
3wishes.txt : 0.16678395496129483
6ablemen.txt : 0.1278442411447338


***** 3lpigs.txt WITH: *****
13chil.txt : 0.157870044569552
3lpigs.txt : 1.0
3wishes.txt : 0.14817170111287759
6ablemen.txt : 0.15291387313047966


***** 3wishes.txt WITH: *****
13chil.txt : 0.16678395496129483
3lpigs.txt : 0.14817170111287759
3wishes.txt : 1.0
6ablemen.txt : 0.16219700876740584


***** 6ablemen.txt WITH: *****
13chil.txt : 0.1278442411447338
3lpigs.txt : 0.15291387313047966
3wishes.txt : 0.16219700876740584
6ablemen.txt : 1.0


=============================================
```

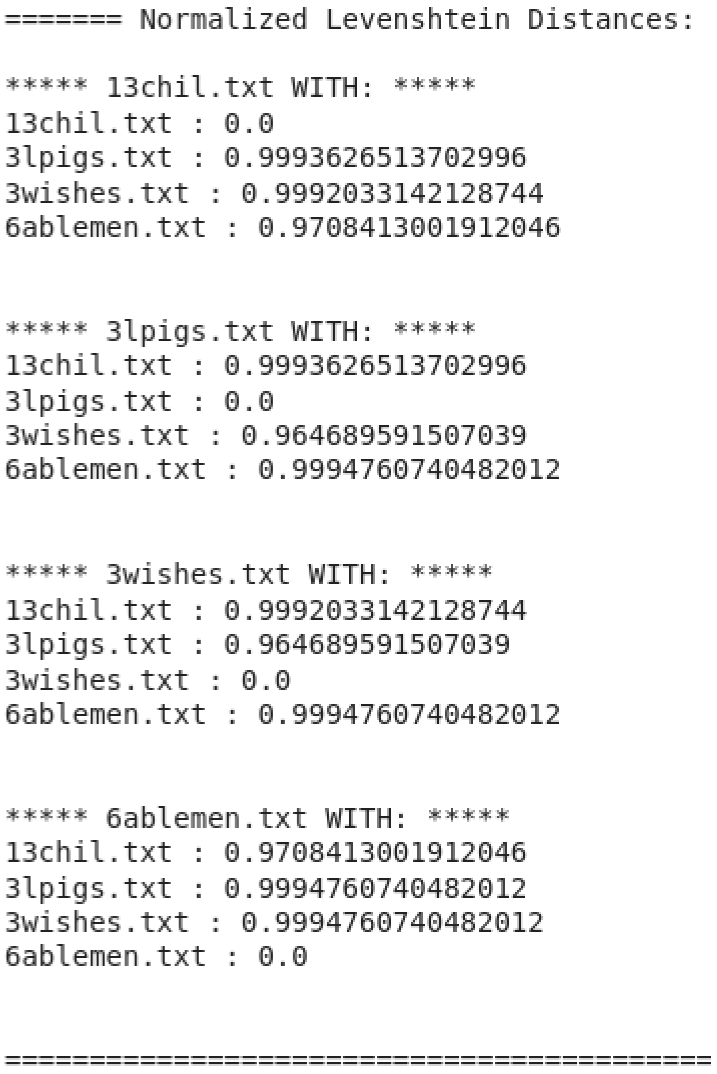


```
======= Normalized Levenshtein Distances:

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.842129955430448
3wishes.txt : 0.8332160450387052
6ablemen.txt : 0.8721557588552662


***** 3lpigs.txt WITH: *****
13chil.txt : 0.842129955430448
3lpigs.txt : 0.0
3wishes.txt : 0.8518282988871224
6ablemen.txt : 0.8470861268695203


***** 3wishes.txt WITH: *****
13chil.txt : 0.8332160450387052
3lpigs.txt : 0.8518282988871224
3wishes.txt : 0.0
6ablemen.txt : 0.8378029912325942


***** 6ablemen.txt WITH: *****
13chil.txt : 0.8721557588552662
3lpigs.txt : 0.8470861268695203
3wishes.txt : 0.8378029912325942
6ablemen.txt : 0.0


=========================================
```

(b) 8-character shingles

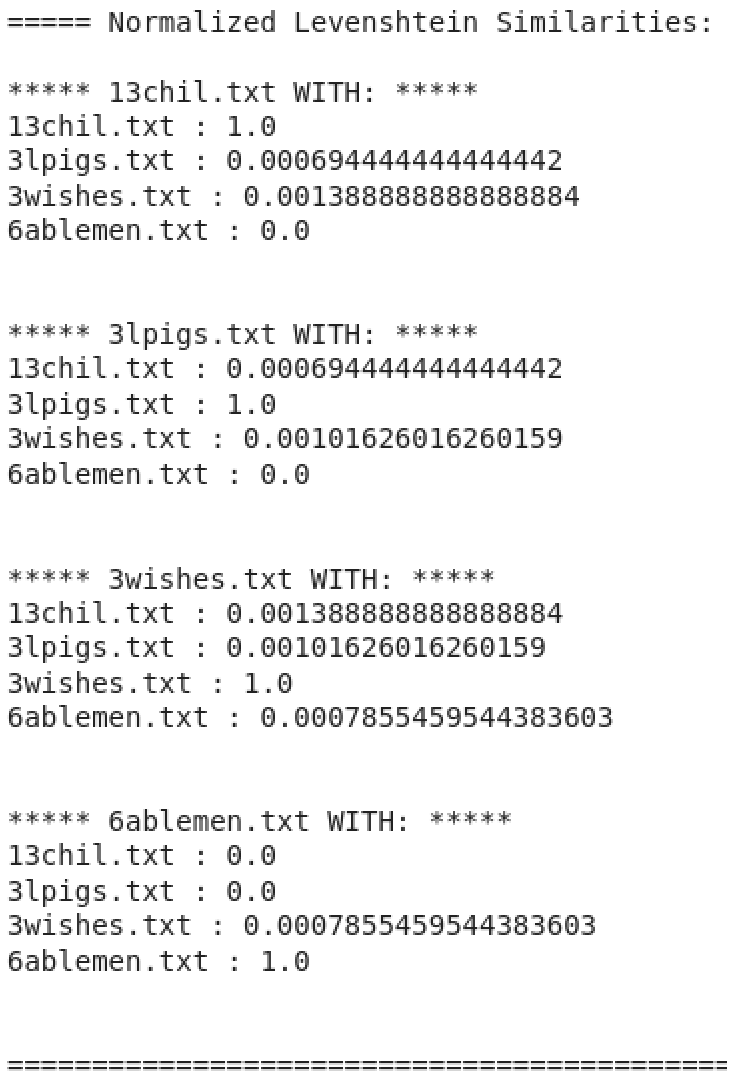


```
===== Normalized Levenshtein Similarities:

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.0006373486297004405
3wishes.txt : 0.0007966857871255506
6ablemen.txt : 0.029158699808795374


***** 3lpigs.txt WITH: *****
13chil.txt : 0.0006373486297004405
3lpigs.txt : 1.0
3wishes.txt : 0.035310408492960965
6ablemen.txt : 0.0005239259517988426


***** 3wishes.txt WITH: *****
13chil.txt : 0.0007966857871255506
3lpigs.txt : 0.035310408492960965
3wishes.txt : 1.0
6ablemen.txt : 0.0005239259517988426


***** 6ablemen.txt WITH: *****
13chil.txt : 0.029158699808795374
3lpigs.txt : 0.0005239259517988426
3wishes.txt : 0.0005239259517988426
6ablemen.txt : 1.0

===========================================
```

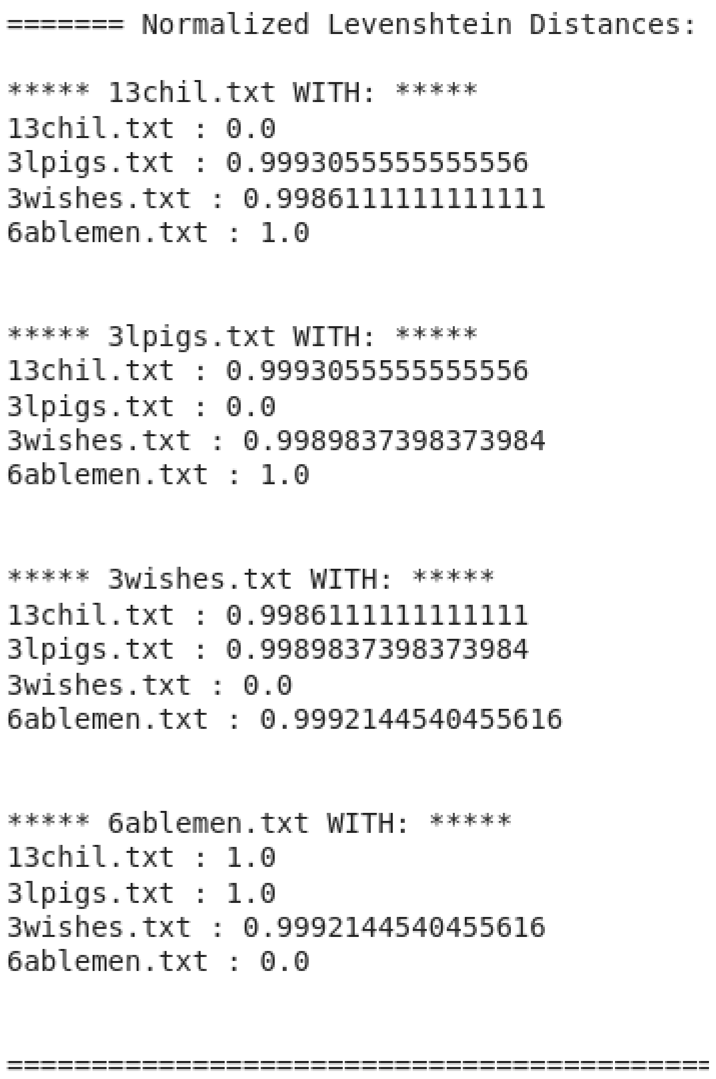


```
======= Normalized Levenshtein Distances:

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.9993626513702996
3wishes.txt : 0.9992033142128744
6ablemen.txt : 0.9708413001912046


***** 3lpigs.txt WITH: *****
13chil.txt : 0.9993626513702996
3lpigs.txt : 0.0
3wishes.txt : 0.964689591507039
6ablemen.txt : 0.9994760740482012


***** 3wishes.txt WITH: *****
13chil.txt : 0.9992033142128744
3lpigs.txt : 0.964689591507039
3wishes.txt : 0.0
6ablemen.txt : 0.9994760740482012


***** 6ablemen.txt WITH: *****
13chil.txt : 0.9708413001912046
3lpigs.txt : 0.9994760740482012
3wishes.txt : 0.9994760740482012
6ablemen.txt : 0.0

==========================================
```

(c) 4-word shingles



```
===== Normalized Levenshtein Similarities:

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.000694444444444442
3wishes.txt : 0.001388888888888884
6ablemen.txt : 0.0


***** 3lpigs.txt WITH: *****
13chil.txt : 0.000694444444444442
3lpigs.txt : 1.0
3wishes.txt : 0.00101626016260159
6ablemen.txt : 0.0


***** 3wishes.txt WITH: *****
13chil.txt : 0.001388888888888884
3lpigs.txt : 0.00101626016260159
3wishes.txt : 1.0
6ablemen.txt : 0.0007855459544383603


***** 6ablemen.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.0
3wishes.txt : 0.0007855459544383603
6ablemen.txt : 1.0

===========================================
```



```
======= Normalized Levenshtein Distances:

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.9993055555555556
3wishes.txt : 0.9986111111111111
6ablemen.txt : 1.0


***** 3lpigs.txt WITH: *****
13chil.txt : 0.9993055555555556
3lpigs.txt : 0.0
3wishes.txt : 0.9989837398373984
6ablemen.txt : 1.0


***** 3wishes.txt WITH: *****
13chil.txt : 0.9986111111111111
3lpigs.txt : 0.9989837398373984
3wishes.txt : 0.0
6ablemen.txt : 0.9992144540455616


***** 6ablemen.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 1.0
3wishes.txt : 0.9992144540455616
6ablemen.txt : 0.0

==========================================
```

Q4. Try the above all for anyone Indian language.

Hindi text files have been used which were downloaded from the below links,

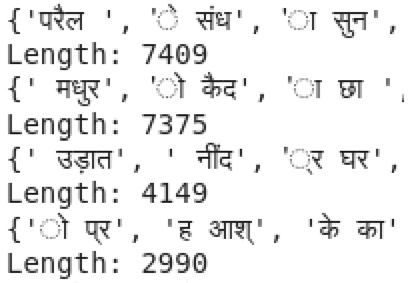
1. https://raw.githubusercontent.com/gayatrivenugopal/hindi-corpus-stoplemmas/master/aesthetics%20corpus/atmaram.txt
2. https://raw.githubusercontent.com/gayatrivenugopal/hindi-corpus-stoplemmas/master/aesthetics%20corpus/dhikkar.txt
3. https://raw.githubusercontent.com/gayatrivenugopal/hindi-corpus-stoplemmas/master/aesthetics%20corpus/pariksha.txt
4. https://raw.githubusercontent.com/gayatrivenugopal/hindi-corpus-stoplemmas/master/aesthetics%20corpus/kshama.txt

Q4(A). Distinct shingles for each document (Hindi Language)

(a) 5-character shingles

‘hindi1.txt’	-	7409
‘hindi2.txt’	-	7375
‘hindi3.txt’	-	4149
‘hindi4.txt’	-	2990

Output Screenshot:

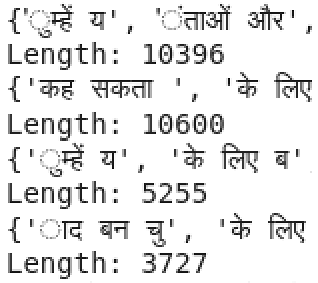


```
{'परैल ', 'ै संध', 'ा सुन',
Length: 7409
{' मधुर', 'ो कैद', 'ा छा ',
Length: 7375
{' उड़ात', ' नींद', '्र घर',
Length: 4149
{'ो प्र', 'ह आश्', 'के का'
Length: 2990
```

(b) 8-character shingles

‘hindi1.txt’	-	10396
‘hindi2.txt’	-	10600
‘hindi3.txt’	-	5255
‘hindi4.txt’	-	3727

Output Screenshot:



```
{'◌ुम्हें य', '◌ंताओं और',
Length: 10396
{'कह सकता ', 'के लिए
Length: 10600
{'◌ुम्हें य', 'के लिए ब'
Length: 5255
{'◌ाद बन चु', 'के लिए
Length: 3727
```

(c) 4-word shingles

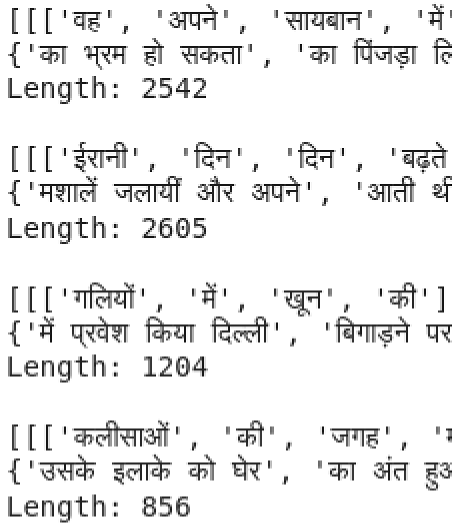
‘hindi1.txt’ - 2542

‘hindi2.txt’ - 2605

‘hindi3.txt’ - 1204

‘hindi4.txt’ - 856

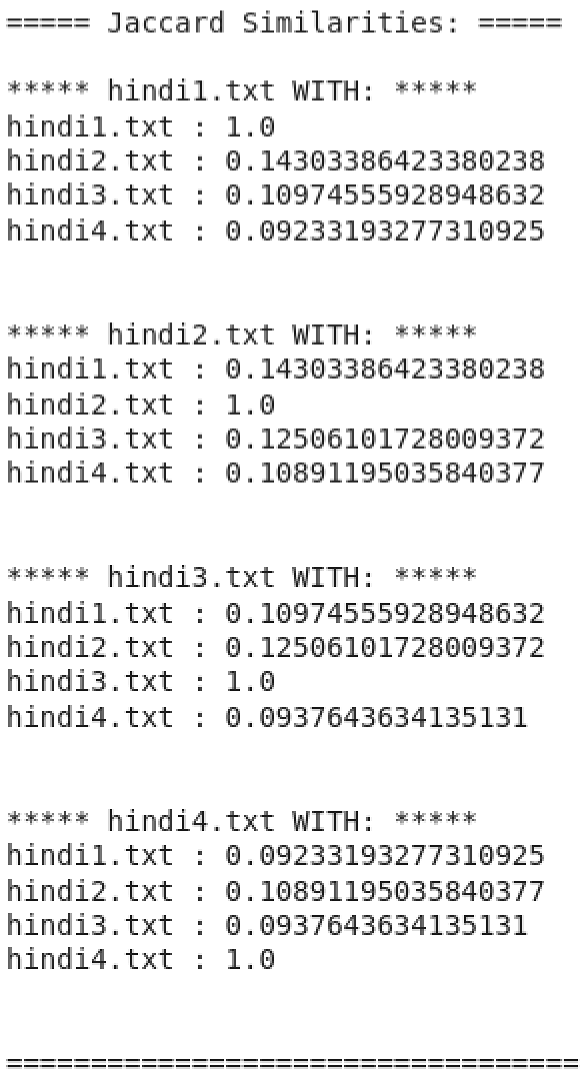
Output Screenshot:



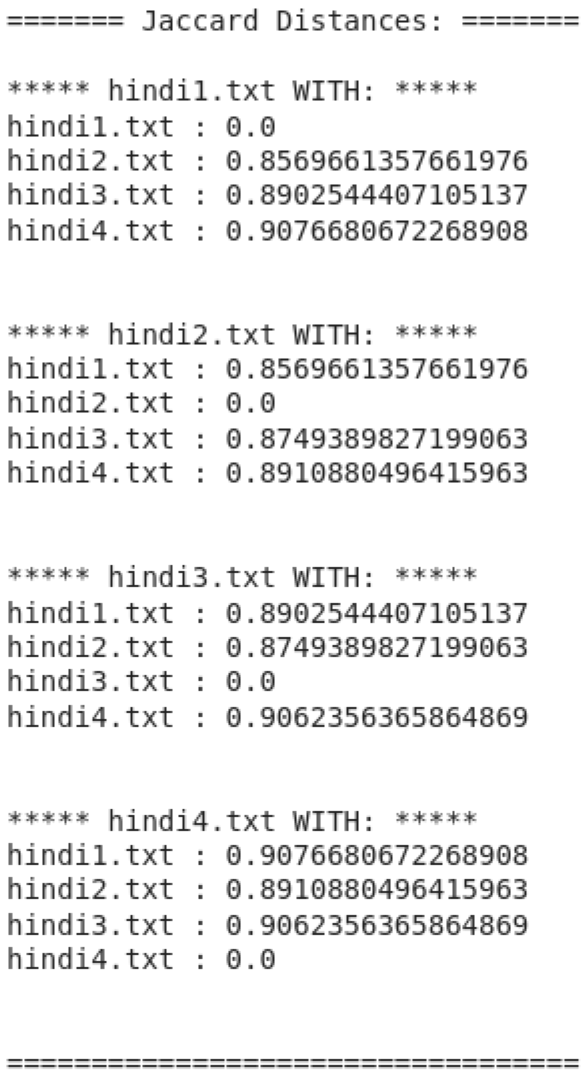
```
[[['वह', 'अपने', 'सायबान', 'में'
{'का भ्रम हो सकता', 'का पिंजड़ा ि
Length: 2542

[[['ईरानी', 'दिन', 'दिन', 'बढ़ते
{'मशालें जलायीं और अपने', 'आती थ
Length: 2605

[[['गलियों', 'में', 'खून', 'की']
{'में प्रवेश किया दिल्ली', 'बिगाड़ने पर
Length: 1204

[[['कलीसाओं', 'की', 'जगह', '
{'उसके इलाके को घेर', 'का अंत हुअ
Length: 856
```

Q4(B). Jaccard Distances of the Hindi documents.

(a) 5-character shingles

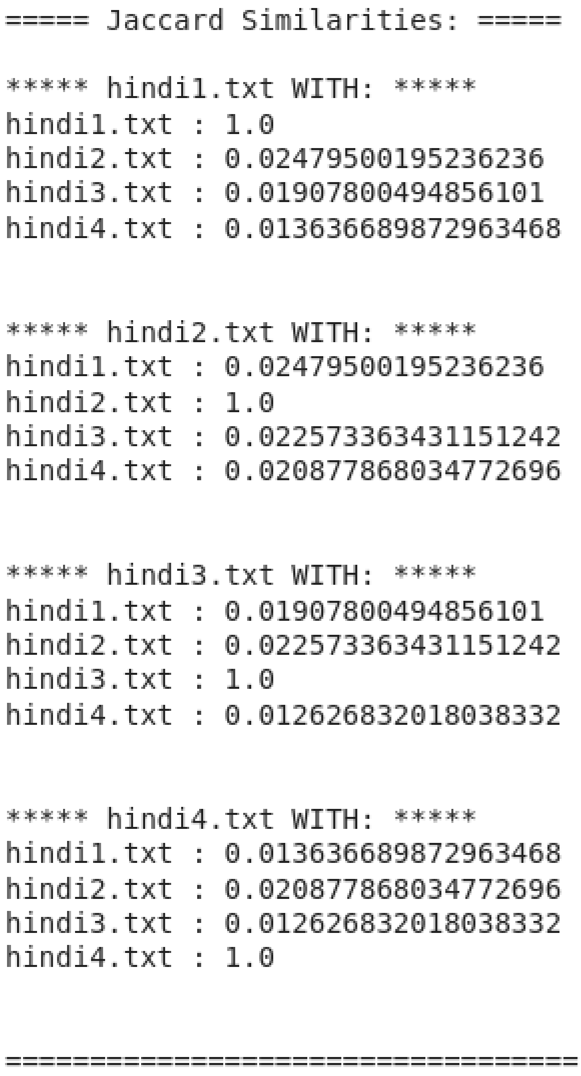


```
===== Jaccard Similarities: =====

***** hindi1.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.14303386423380238
hindi3.txt : 0.10974555928948632
hindi4.txt : 0.09233193277310925


***** hindi2.txt WITH: *****
hindi1.txt : 0.14303386423380238
hindi2.txt : 1.0
hindi3.txt : 0.12506101728009372
hindi4.txt : 0.10891195035840377


***** hindi3.txt WITH: *****
hindi1.txt : 0.10974555928948632
hindi2.txt : 0.12506101728009372
hindi3.txt : 1.0
hindi4.txt : 0.0937643634135131


***** hindi4.txt WITH: *****
hindi1.txt : 0.09233193277310925
hindi2.txt : 0.10891195035840377
hindi3.txt : 0.0937643634135131
hindi4.txt : 1.0

==================================
```

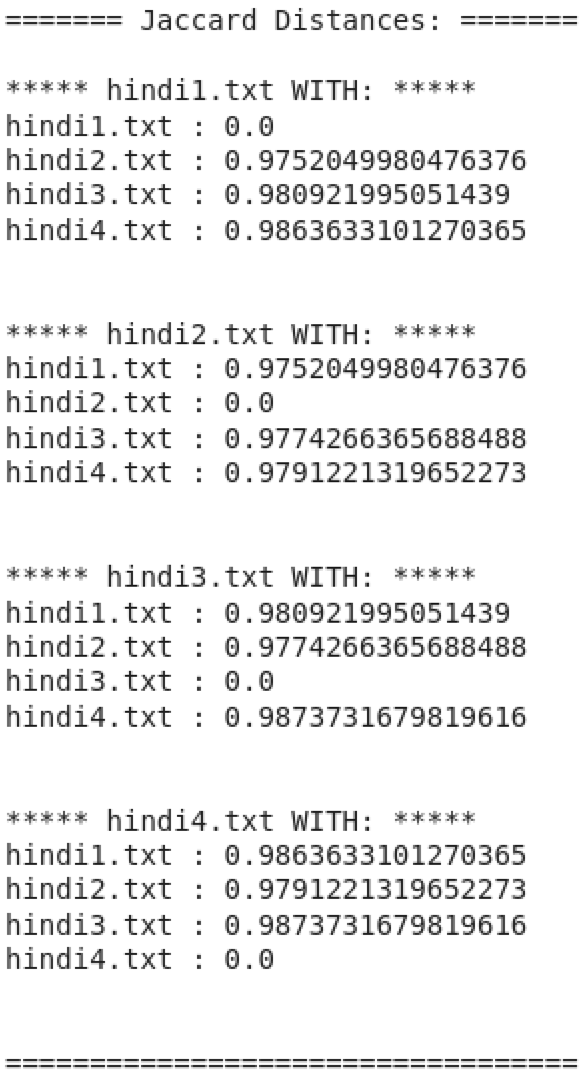


```
======= Jaccard Distances: =======

***** hindi1.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.8569661357661976
hindi3.txt : 0.8902544407105137
hindi4.txt : 0.9076680672268908


***** hindi2.txt WITH: *****
hindi1.txt : 0.8569661357661976
hindi2.txt : 0.0
hindi3.txt : 0.8749389827199063
hindi4.txt : 0.8910880496415963


***** hindi3.txt WITH: *****
hindi1.txt : 0.8902544407105137
hindi2.txt : 0.8749389827199063
hindi3.txt : 0.0
hindi4.txt : 0.9062356365864869


***** hindi4.txt WITH: *****
hindi1.txt : 0.9076680672268908
hindi2.txt : 0.8910880496415963
hindi3.txt : 0.9062356365864869
hindi4.txt : 0.0

==================================
```

(b) 8-character shingles

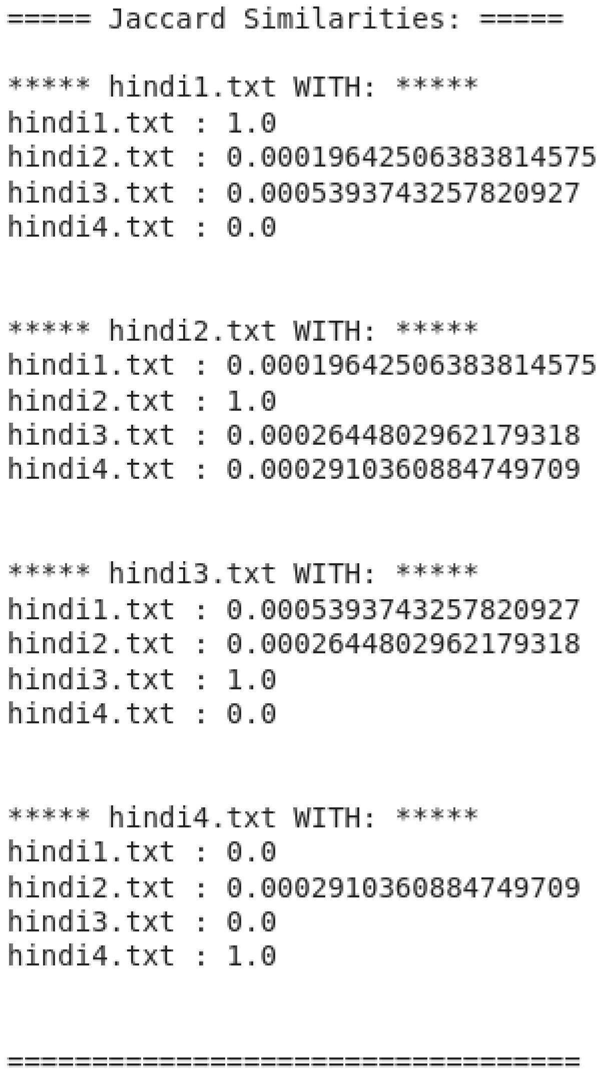


```
===== Jaccard Similarities: =====

***** hindi1.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.02479500195236236
hindi3.txt : 0.01907800494856101
hindi4.txt : 0.013636689872963468


***** hindi2.txt WITH: *****
hindi1.txt : 0.02479500195236236
hindi2.txt : 1.0
hindi3.txt : 0.022573363431151242
hindi4.txt : 0.020877868034772696


***** hindi3.txt WITH: *****
hindi1.txt : 0.01907800494856101
hindi2.txt : 0.022573363431151242
hindi3.txt : 1.0
hindi4.txt : 0.012626832018038332


***** hindi4.txt WITH: *****
hindi1.txt : 0.013636689872963468
hindi2.txt : 0.020877868034772696
hindi3.txt : 0.012626832018038332
hindi4.txt : 1.0


===================================
```

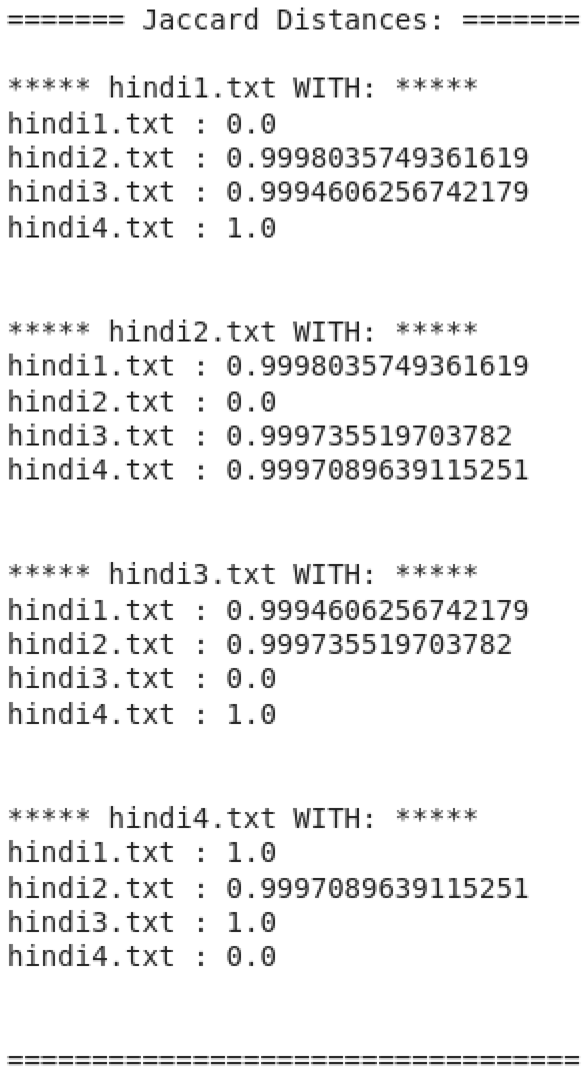


```
======= Jaccard Distances: =======

***** hindi1.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.9752049980476376
hindi3.txt : 0.980921995051439
hindi4.txt : 0.9863633101270365


***** hindi2.txt WITH: *****
hindi1.txt : 0.9752049980476376
hindi2.txt : 0.0
hindi3.txt : 0.9774266365688488
hindi4.txt : 0.9791221319652273


***** hindi3.txt WITH: *****
hindi1.txt : 0.980921995051439
hindi2.txt : 0.9774266365688488
hindi3.txt : 0.0
hindi4.txt : 0.9873731679819616


***** hindi4.txt WITH: *****
hindi1.txt : 0.9863633101270365
hindi2.txt : 0.9791221319652273
hindi3.txt : 0.9873731679819616
hindi4.txt : 0.0


===================================
```

(c) 4-word shingles



```
===== Jaccard Similarities: =====

***** hindi1.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.00019642506383814575
hindi3.txt : 0.0005393743257820927
hindi4.txt : 0.0


***** hindi2.txt WITH: *****
hindi1.txt : 0.00019642506383814575
hindi2.txt : 1.0
hindi3.txt : 0.0002644802962179318
hindi4.txt : 0.0002910360884749709


***** hindi3.txt WITH: *****
hindi1.txt : 0.0005393743257820927
hindi2.txt : 0.0002644802962179318
hindi3.txt : 1.0
hindi4.txt : 0.0


***** hindi4.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.0002910360884749709
hindi3.txt : 0.0
hindi4.txt : 1.0


===================================
```

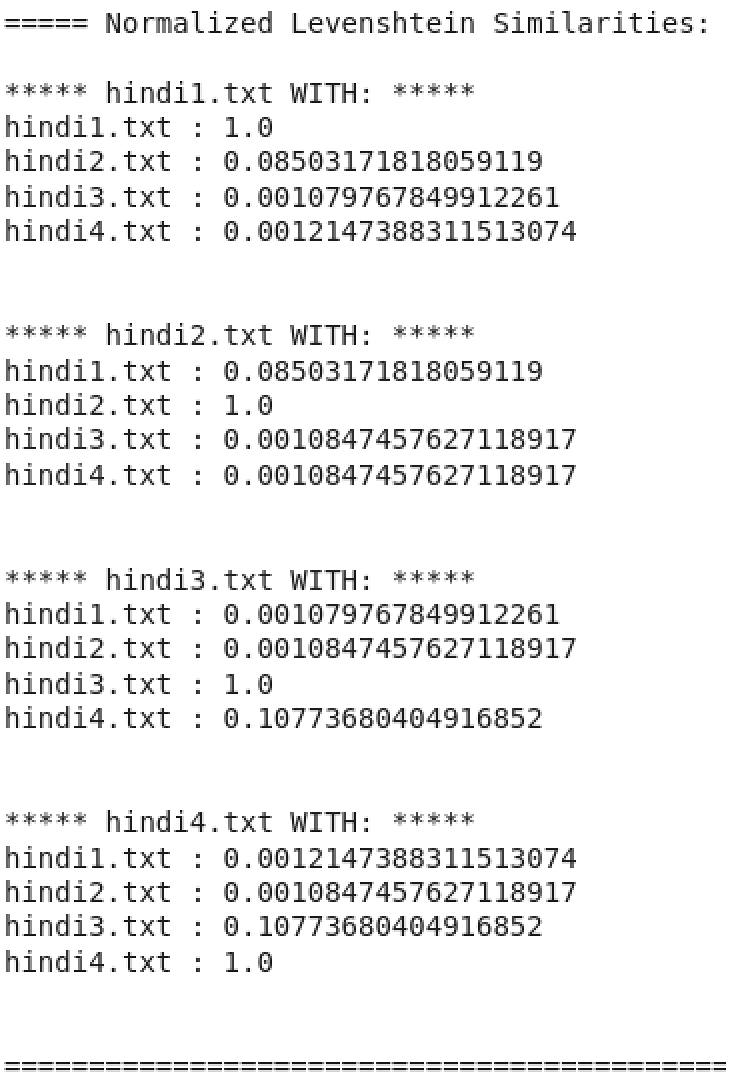


```
======= Jaccard Distances: =======

***** hindi1.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.9998035749361619
hindi3.txt : 0.9994606256742179
hindi4.txt : 1.0


***** hindi2.txt WITH: *****
hindi1.txt : 0.9998035749361619
hindi2.txt : 0.0
hindi3.txt : 0.999735519703782
hindi4.txt : 0.9997089639115251


***** hindi3.txt WITH: *****
hindi1.txt : 0.9994606256742179
hindi2.txt : 0.999735519703782
hindi3.txt : 0.0
hindi4.txt : 1.0


***** hindi4.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.9997089639115251
hindi3.txt : 1.0
hindi4.txt : 0.0


===================================
```

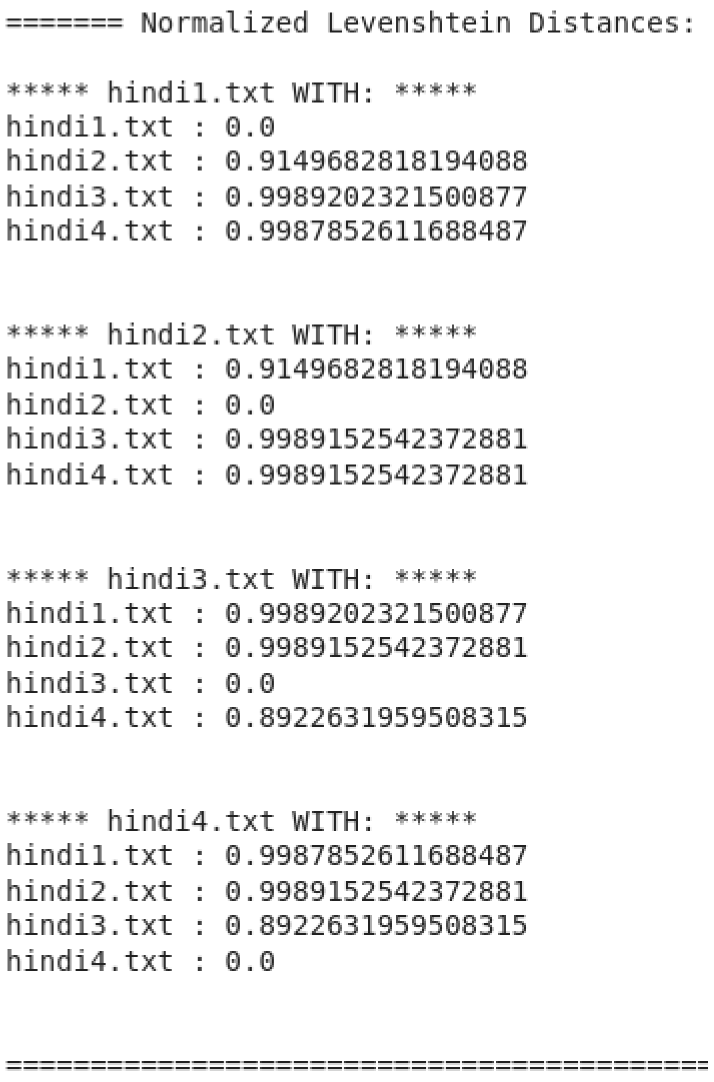
Q4(C). Using different similarity function (Hindi Language).

Here also, normalized levenshtein similarity/distance has been used as for English documents.

(a) 5-character shingles

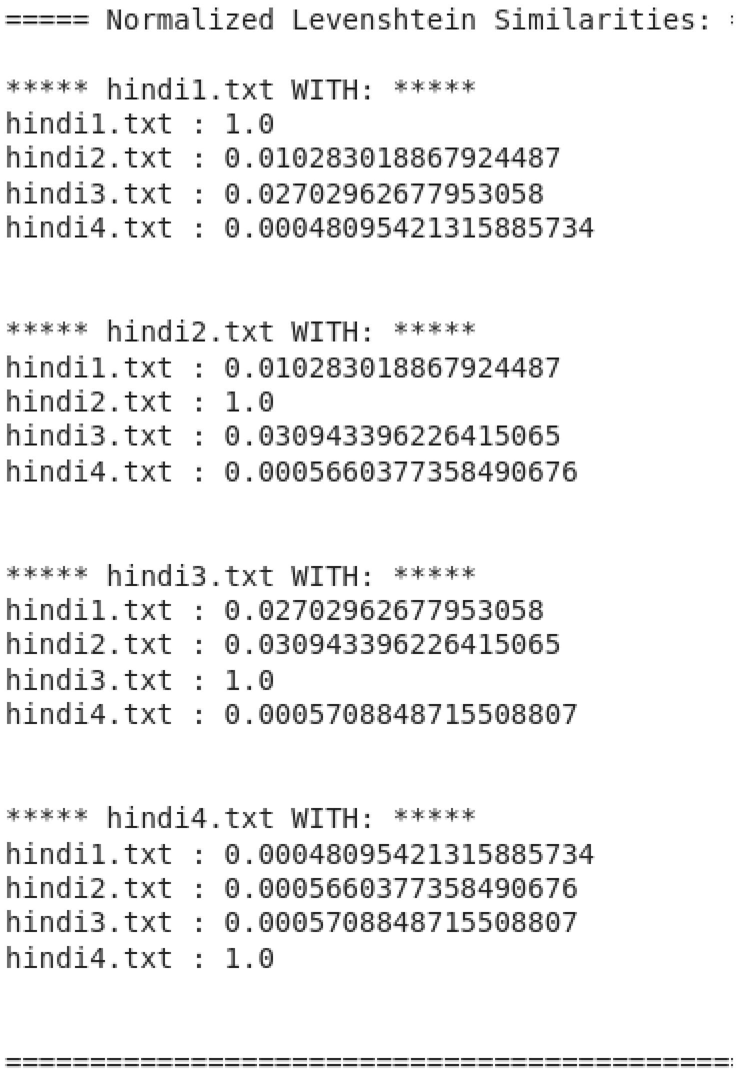


```
===== Normalized Levenshtein Similarities:

***** hindi1.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.08503171818059119
hindi3.txt : 0.001079767849912261
hindi4.txt : 0.0012147388311513074


***** hindi2.txt WITH: *****
hindi1.txt : 0.08503171818059119
hindi2.txt : 1.0
hindi3.txt : 0.0010847457627118917
hindi4.txt : 0.0010847457627118917


***** hindi3.txt WITH: *****
hindi1.txt : 0.001079767849912261
hindi2.txt : 0.0010847457627118917
hindi3.txt : 1.0
hindi4.txt : 0.10773680404916852


***** hindi4.txt WITH: *****
hindi1.txt : 0.0012147388311513074
hindi2.txt : 0.0010847457627118917
hindi3.txt : 0.10773680404916852
hindi4.txt : 1.0

==========================================
```

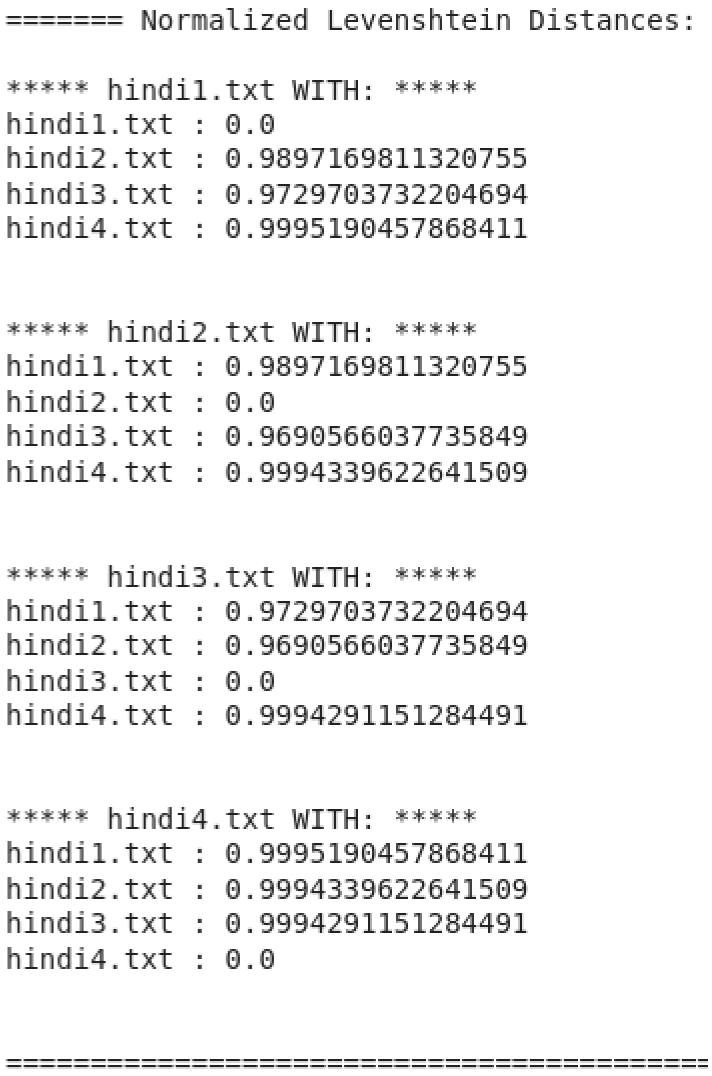


```
======= Normalized Levenshtein Distances:

***** hindi1.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.9149682818194088
hindi3.txt : 0.9989202321500877
hindi4.txt : 0.9987852611688487


***** hindi2.txt WITH: *****
hindi1.txt : 0.9149682818194088
hindi2.txt : 0.0
hindi3.txt : 0.9989152542372881
hindi4.txt : 0.9989152542372881


***** hindi3.txt WITH: *****
hindi1.txt : 0.9989202321500877
hindi2.txt : 0.9989152542372881
hindi3.txt : 0.0
hindi4.txt : 0.8922631959508315


***** hindi4.txt WITH: *****
hindi1.txt : 0.9987852611688487
hindi2.txt : 0.9989152542372881
hindi3.txt : 0.8922631959508315
hindi4.txt : 0.0

==========================================
```

(b) 8-character shingles

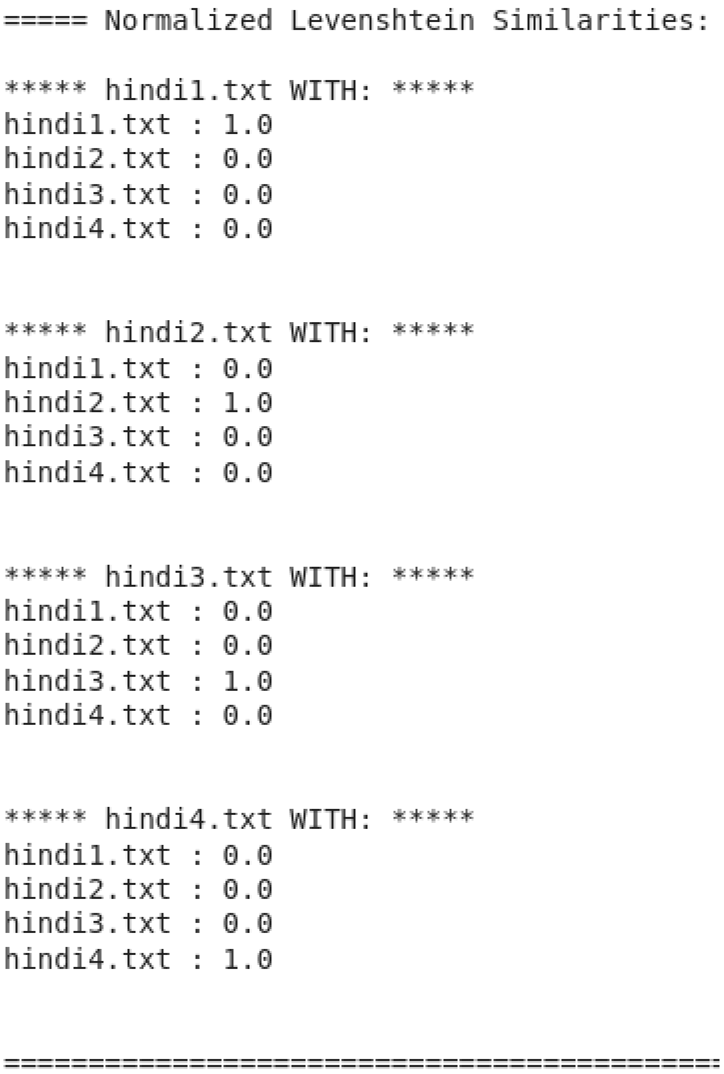


```
===== Normalized Levenshtein Similarities: =

***** hindi1.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.010283018867924487
hindi3.txt : 0.02702962677953058
hindi4.txt : 0.00048095421315885734


***** hindi2.txt WITH: *****
hindi1.txt : 0.010283018867924487
hindi2.txt : 1.0
hindi3.txt : 0.030943396226415065
hindi4.txt : 0.0005660377358490676


***** hindi3.txt WITH: *****
hindi1.txt : 0.02702962677953058
hindi2.txt : 0.030943396226415065
hindi3.txt : 1.0
hindi4.txt : 0.0005708848715508807


***** hindi4.txt WITH: *****
hindi1.txt : 0.00048095421315885734
hindi2.txt : 0.0005660377358490676
hindi3.txt : 0.0005708848715508807
hindi4.txt : 1.0

===========================================
```

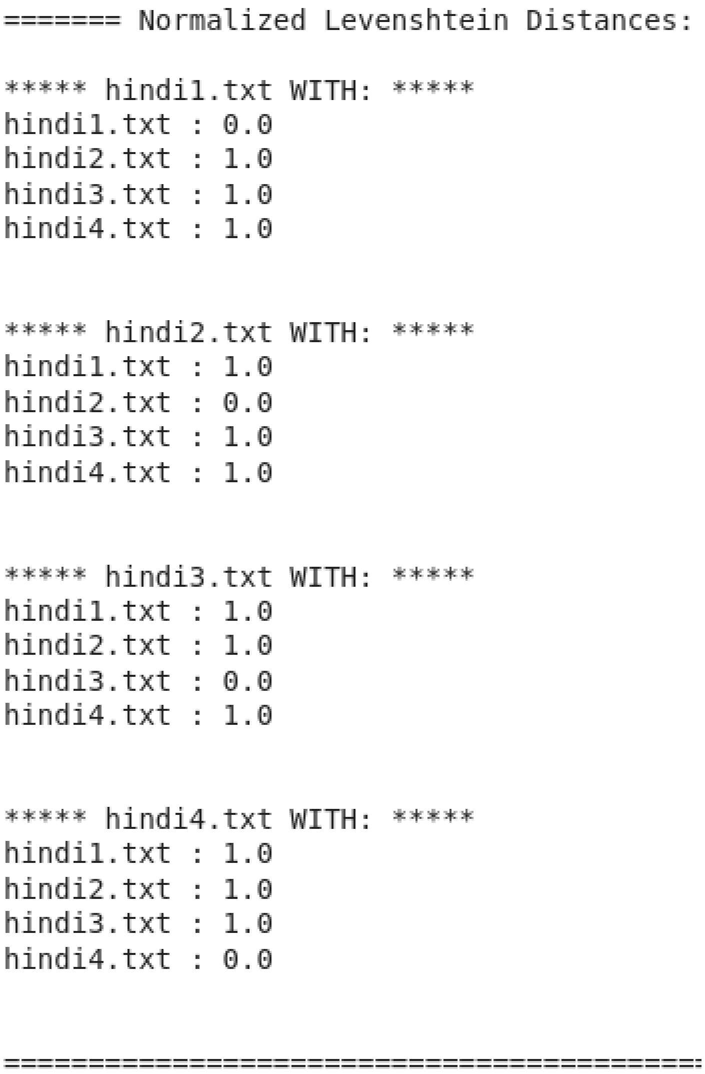


```
======= Normalized Levenshtein Distances:

***** hindi1.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.9897169811320755
hindi3.txt : 0.9729703732204694
hindi4.txt : 0.9995190457868411


***** hindi2.txt WITH: *****
hindi1.txt : 0.9897169811320755
hindi2.txt : 0.0
hindi3.txt : 0.9690566037735849
hindi4.txt : 0.9994339622641509


***** hindi3.txt WITH: *****
hindi1.txt : 0.9729703732204694
hindi2.txt : 0.9690566037735849
hindi3.txt : 0.0
hindi4.txt : 0.9994291151284491


***** hindi4.txt WITH: *****
hindi1.txt : 0.9995190457868411
hindi2.txt : 0.9994339622641509
hindi3.txt : 0.9994291151284491
hindi4.txt : 0.0

==========================================
```

(c) 4-word shingles



```
===== Normalized Levenshtein Similarities:

***** hindi1.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.0
hindi3.txt : 0.0
hindi4.txt : 0.0


***** hindi2.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 1.0
hindi3.txt : 0.0
hindi4.txt : 0.0


***** hindi3.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.0
hindi3.txt : 1.0
hindi4.txt : 0.0


***** hindi4.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 0.0
hindi3.txt : 0.0
hindi4.txt : 1.0


===========================================
```

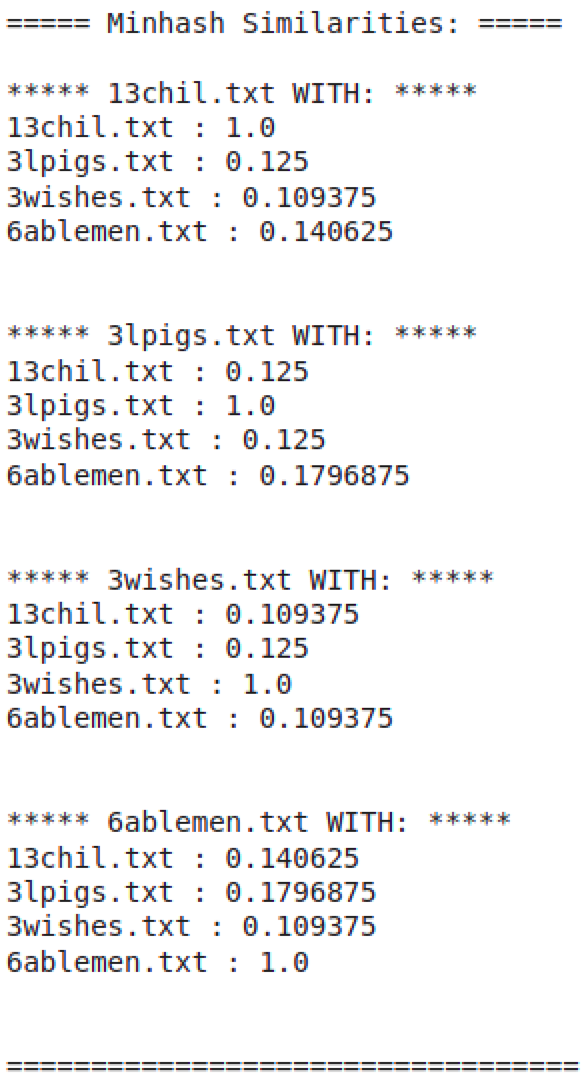


```
======= Normalized Levenshtein Distances:

***** hindi1.txt WITH: *****
hindi1.txt : 0.0
hindi2.txt : 1.0
hindi3.txt : 1.0
hindi4.txt : 1.0


***** hindi2.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 0.0
hindi3.txt : 1.0
hindi4.txt : 1.0


***** hindi3.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 1.0
hindi3.txt : 0.0
hindi4.txt : 1.0


***** hindi4.txt WITH: *****
hindi1.txt : 1.0
hindi2.txt : 1.0
hindi3.txt : 1.0
hindi4.txt : 0.0


=========================================
```

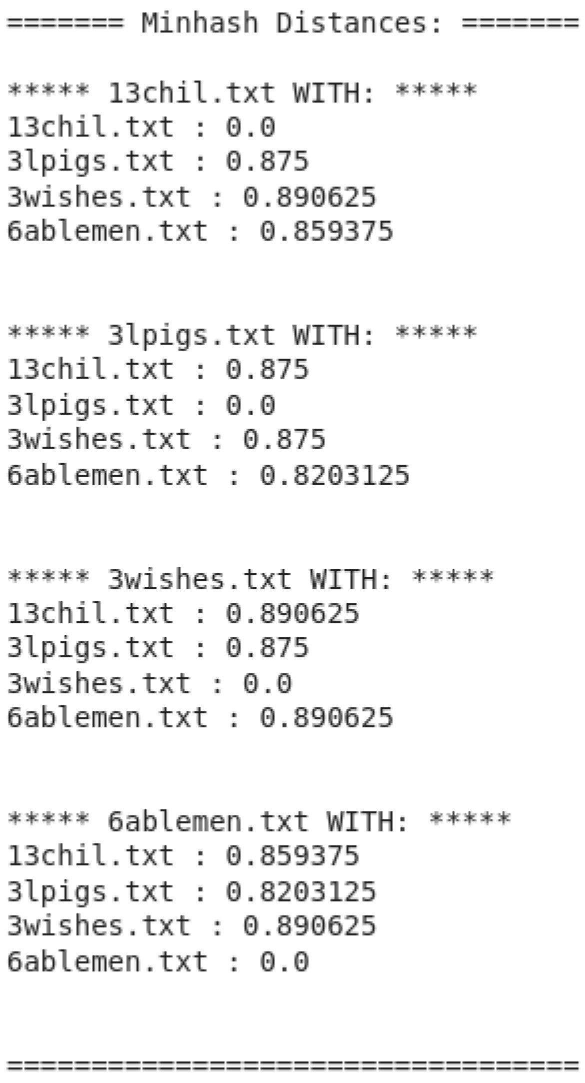
Q5. Build a min hash signature for the above experiment and provide your conclusions for the entire experiment.

Min hashing has been performed from scratch. The code can be found in the colab notebook. Only the English texts have been used to compute min hashing and the signature matrices obtained are used to compute similarity between each type of shingle (5-char, 8-char and 4-word).

(a) 5-character shingles

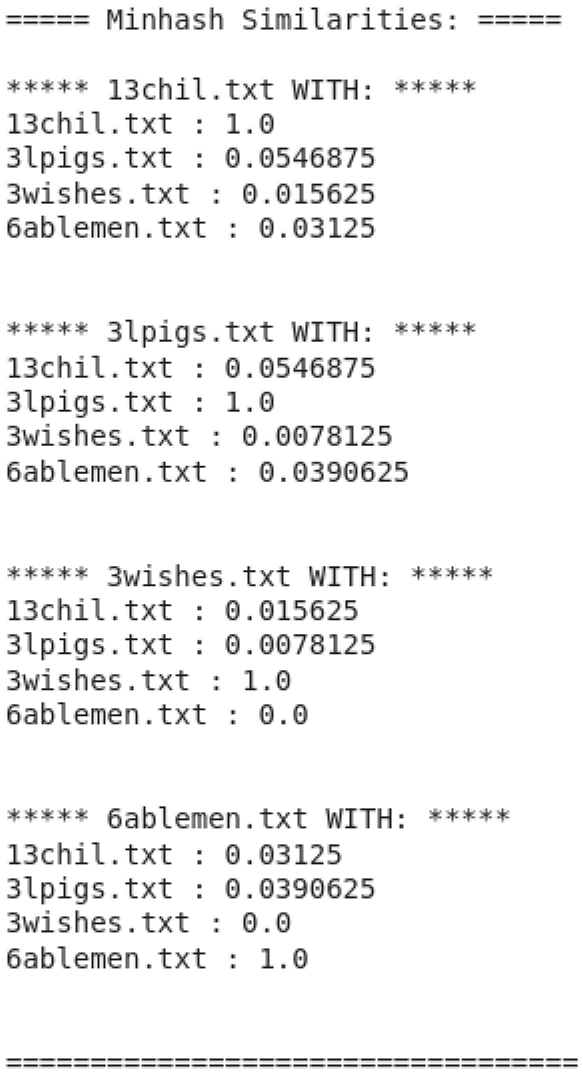


```
===== Minhash Similarities: =====

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.125
3wishes.txt : 0.109375
6ablemen.txt : 0.140625


***** 3lpigs.txt WITH: *****
13chil.txt : 0.125
3lpigs.txt : 1.0
3wishes.txt : 0.125
6ablemen.txt : 0.1796875


***** 3wishes.txt WITH: *****
13chil.txt : 0.109375
3lpigs.txt : 0.125
3wishes.txt : 1.0
6ablemen.txt : 0.109375


***** 6ablemen.txt WITH: *****
13chil.txt : 0.140625
3lpigs.txt : 0.1796875
3wishes.txt : 0.109375
6ablemen.txt : 1.0


==================================
```

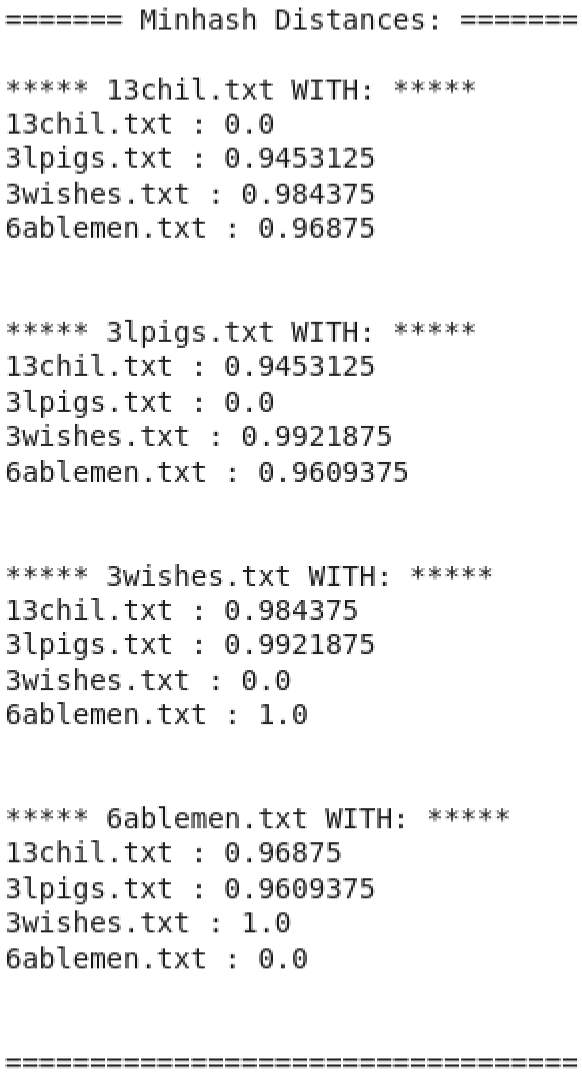


```
======= Minhash Distances: =======

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.875
3wishes.txt : 0.890625
6ablemen.txt : 0.859375


***** 3lpigs.txt WITH: *****
13chil.txt : 0.875
3lpigs.txt : 0.0
3wishes.txt : 0.875
6ablemen.txt : 0.8203125


***** 3wishes.txt WITH: *****
13chil.txt : 0.890625
3lpigs.txt : 0.875
3wishes.txt : 0.0
6ablemen.txt : 0.890625


***** 6ablemen.txt WITH: *****
13chil.txt : 0.859375
3lpigs.txt : 0.8203125
3wishes.txt : 0.890625
6ablemen.txt : 0.0


==================================
```

(b) 8-character shingles

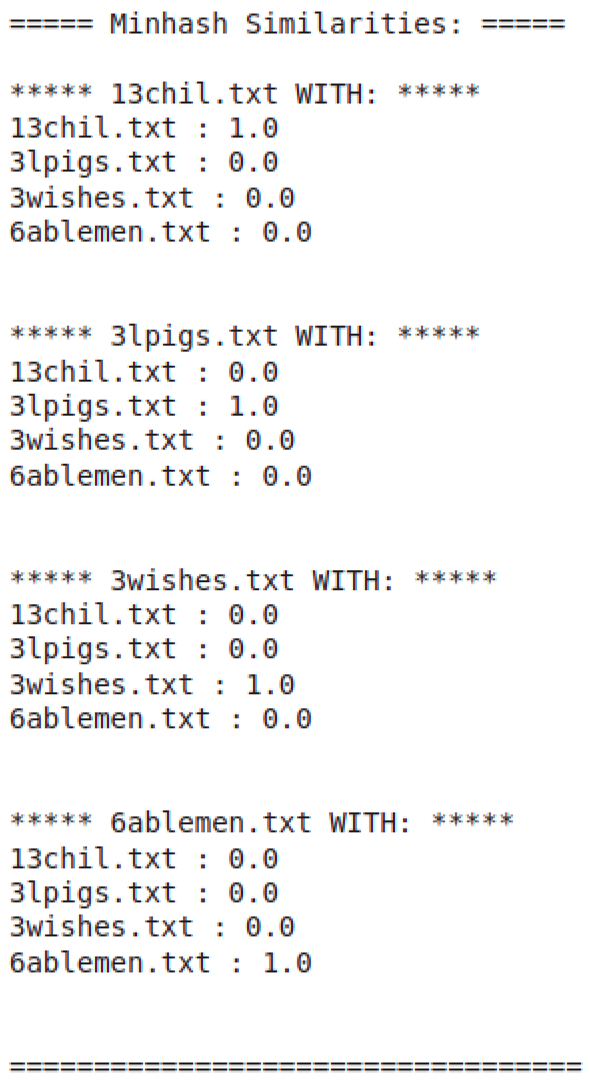


```
===== Minhash Similarities: =====

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.0546875
3wishes.txt : 0.015625
6ablemen.txt : 0.03125


***** 3lpigs.txt WITH: *****
13chil.txt : 0.0546875
3lpigs.txt : 1.0
3wishes.txt : 0.0078125
6ablemen.txt : 0.0390625


***** 3wishes.txt WITH: *****
13chil.txt : 0.015625
3lpigs.txt : 0.0078125
3wishes.txt : 1.0
6ablemen.txt : 0.0


***** 6ablemen.txt WITH: *****
13chil.txt : 0.03125
3lpigs.txt : 0.0390625
3wishes.txt : 0.0
6ablemen.txt : 1.0

==================================
```

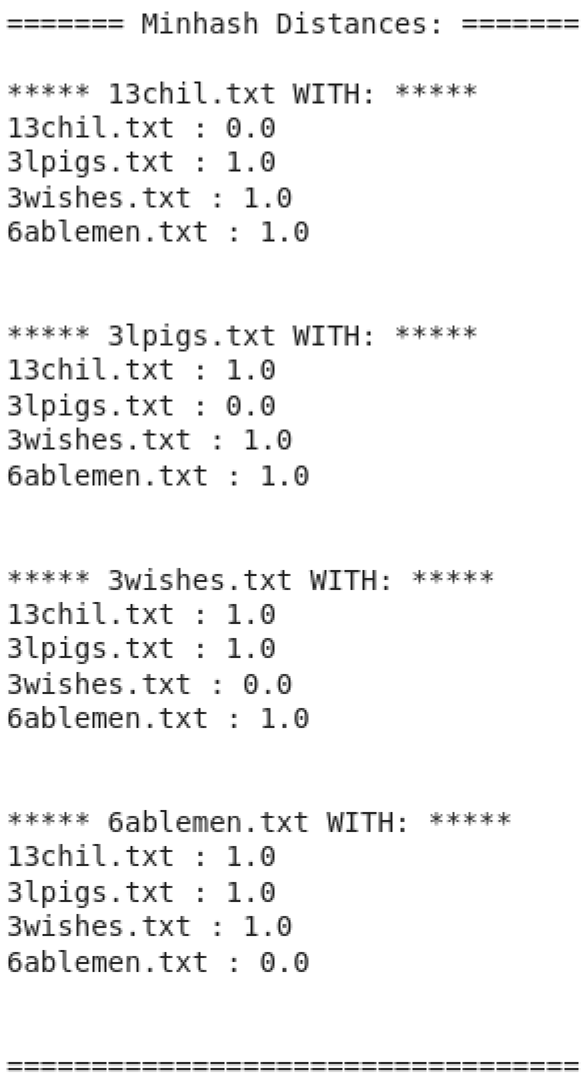


```
======= Minhash Distances: =======

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.9453125
3wishes.txt : 0.984375
6ablemen.txt : 0.96875


***** 3lpigs.txt WITH: *****
13chil.txt : 0.9453125
3lpigs.txt : 0.0
3wishes.txt : 0.9921875
6ablemen.txt : 0.9609375


***** 3wishes.txt WITH: *****
13chil.txt : 0.984375
3lpigs.txt : 0.9921875
3wishes.txt : 0.0
6ablemen.txt : 1.0


***** 6ablemen.txt WITH: *****
13chil.txt : 0.96875
3lpigs.txt : 0.9609375
3wishes.txt : 1.0
6ablemen.txt : 0.0

==================================
```

(c) 4-word shingles



```
===== Minhash Similarities: =====

***** 13chil.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.0
3wishes.txt : 0.0
6ablemen.txt : 0.0


***** 3lpigs.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 1.0
3wishes.txt : 0.0
6ablemen.txt : 0.0


***** 3wishes.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.0
3wishes.txt : 1.0
6ablemen.txt : 0.0


***** 6ablemen.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 0.0
3wishes.txt : 0.0
6ablemen.txt : 1.0


===================================
```



```
======= Minhash Distances: =======

***** 13chil.txt WITH: *****
13chil.txt : 0.0
3lpigs.txt : 1.0
3wishes.txt : 1.0
6ablemen.txt : 1.0


***** 3lpigs.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 0.0
3wishes.txt : 1.0
6ablemen.txt : 1.0


***** 3wishes.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 1.0
3wishes.txt : 0.0
6ablemen.txt : 1.0


***** 6ablemen.txt WITH: *****
13chil.txt : 1.0
3lpigs.txt : 1.0
3wishes.txt : 1.0
6ablemen.txt : 0.0


===================================
```

Conclusion:

We can see that the similarity changes as we use different similarity functions/ Particularly for the English texts, the similarity decreases when we use min hashing instead of the traditional shingling. This suggests that min hashing is a much efficient way to get finer document similarities. On the other hand, when we use Normalized Levenshtein similarity, the similarities change but the changes are not very significant. In some cases like 4-word shingles the change in distances is significant but it decreases as we decrease the length of a shingle. The conclusions drawn for normalized levenshtein distance may change if some other similarity function would’ve been used.

THANK YOU