THAO PHUNG (W07976257)

#Problem 1

Q1: How many distinct k-grams/shingles are there for each document with each type of k-gram/shingle

	D1	D2	D3	D4
SH1	241	268	261	298
SH2	321	400	400	511
SH3	181	249	238	318

Q2: Compute the Jaccard Similarity between all pairs of documents

	D1-D2	D1-D3	D1-D4	D2-D3	D2-D4	D3-D4
SH1	0.51	0.56875	0.47	0.516	0.4588	0.471
SH2	0.18	0.222	0.156	0.21	0.168	0.15
SH3	0.026	0.0997	0.006	0.0188	0.0107	0.0091

Problem 2

L0(D1, D2) RCH 10 = 0.1 (t = 0.006578)
L0(D1, D2) RCH 20 = 0.15 (t = 0.0096).
L0(D1, D2) RCH 60 = 0.15 (t = 0.0303)
L0(D1, D2) RCH 200 = 0.16 (t = 0.115)
L0(D1, D2) RCH 500 = 0.21 (t = 0.266)

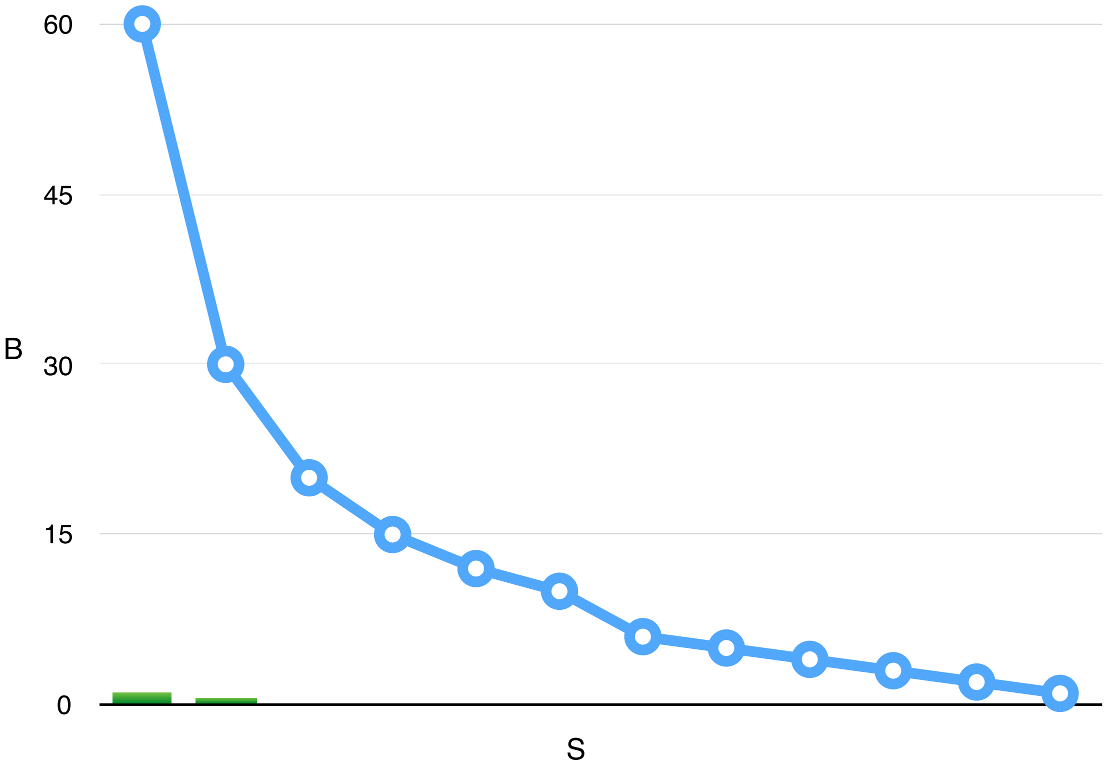
RCH-Good Value: rhc = 200

JUSTIFICATION FOR PRIOR ANSWER: With rhc = 200, the result is close to Jaccard Similarity when we use intersection and union. It also implemented fast.

#3 LSH

Table 1-1

R	S
1	0.999998467504459
2	0.706142356769295
3	0.148404332914851
4	0.0237330547657416
5	0.003833248803772
6	0.000639815711453329
10	0.000000614399842890911
12	0.0000000204799996961214
15	0.000000000131072042108826
20	0.00000000000000313082892944294
30	0
60	0


60
45
B
30
15
0
S