Quiz 6: Similar Items

Name: GRADER ID: ____________

Input: 10 million documents in English in the database and 1 new document
Output: in the database, all the documents that are similar to the new document with 80% Jaccard similarity and their Jaccard similarity to the new document

Describe your process to generate the output from the input using k-shingles, minhash, and LSH. You need to describe your input, process, and output using examples.

1. How to generate features describing each document using 10-shingles (1 pt)

For each document generate unique 10-shingles sets and then converts them into integers as features.

	D_1	D_2	. . . . .
S_1	0	1	0 . . .
S_2	1	0	0
S_3	0	0	1 . . . .
⋮			

2. How to efficiently generate 100 minhash signatures for each document from their 10-shingle features (2 pts)

[1 POINT] → Design 100 hash functions and find its 100 min hash signatures by scanning the table and indicate the row number where first '1' appears.

[1 point] →

	D_1	D_2	D_3	. . . .	D_{10000}	h_0	h_1	h_2 . .	h_{99}
S_1	0	1	0	. . . .		1	. . .	. .	
S_2	1	0	1	. . . .		0	. . .	.	
S_3	0	0	0	. . .		5	. . .	.	
⋮									

← minhash signature

3. How to use LSH to speed up the process of comparing signatures; what are the parameters in your LSH process? (3 pts); what are the false-positives and false negatives? (1 pt) how to set the parameters to control false-positives and false negatives? (3 pts)

[1.5 POINTS] To speed up the process, ~~we~~ sample the signatures with several bands and just compare the signatures in the same bands. if similar then put into candidate pairs

[1.5 POINTS] Parameters: r: rows of each band., b: number of bands.
s: Jaccard similarity, t: threshold.

[0.5] False positives: Items are dissimilar but hashed to same band/bucket.

[0.5] False negatives: Items are similar but hashed to different band/bucket.

[1.5 Points] r: ~~increase False~~ increase r it will increase false negative but reduces False +ve.

[1.5 POINTS] b: increase b it will increase false positive but reduces False negatives