Home Work 3
Com S 435/535

Due Oct 19: 9:30AM

There are 5 problems and each problem is worth 50 points.

1. Consider the following documents $D_1 = \{1, 4, 6, 7, 8\}$ and $D_2 = \{2, 3, 9, 4, 7\}$
   (a) What are the *binary* term-frequency vectors of D_1 and D_2?
   (b) What is the Jaccard Similary of D_1 and D_2 (with respect to binary term-frequency vectors)
   (c) What is the cosine similarity of D_1 and D_2 (with respect to binary term-frequency vectors)

2. Let D_1 and D_2 be two documents. Let C be the cosine similarity of the documents with respect to binary term-frequency vectors and J be the jacquard similarity with respect to binary term-frequency vectors. Show that
   (a) $C^2 \leq J$
   (b) $J \leq \frac{C}{2-C}$

3. Consider the following term-document matrix.

	D_1	D_2	D_3	D_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1

   Suppose we picked the following permutations $(2x+1)\%5$, $(3x+4)\%5$, and $(x+3)\%5$. Compute the MinHash matrix.

4. Suppose that D_1 and D_2 are two documents with $D_1 \cup D_2 = \{1, 2, \cdots, n\}$. We showed that if we randomly pick a permutation $\Pi : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$, then

$$\Pr[\min[\Pi(D_1)] = \min[\Pi(D_2)]] = Jac(D_1, D_2).$$

   Suppose that we randomly pick a one-one function h from $\{1, 2, \cdots, n\}$ to $\{1, 2, \cdots, n+1\}$. Show that

$$\Pr[\min[h(D_1)] = \min[h(D_2)]] = Jac(D_1, D_2).$$

5. Let s be a similarity measure on M-dimensional vectors. We say that a family of hash functions $\mathcal{H}$ is *similarity preserving*, if for every pair of vector U and V

$$\Pr_{h \in \mathcal{H}}[h(U) = h(V)] = s(U, V).$$

Consider the following similarity measure: Let $U = \langle u_1, u_2, \cdots, u_M \rangle$, and $V = \langle v_1, v_2, \cdots, v_M \rangle$ be two M-dimensional vectors.

$$s(U, V) = \frac{|\{i \mid u_i = v_i\}|}{M}$$

Describe a similarity preserving hash family for the above similarity measure.

Guidelines:

- You are allowed to discuss with your classmates, while doing your homework. However, I strongly suggest that you think about the problems on your own before discussing.
- Definition of *classmates*: Students who are taking CS 435/535 in Fall 17.
- However, You should write the final solutions alone, without consulting your classmates. Your writing should demonstrate that you understand the proofs completely. If I suspect that you wrote the proofs without understanding, I may ask you to explain the proofs to me in person. In such scenarios, failure to explain proofs will be taken as evidence of *academic dishonesty*.
- For each problem, you should acknowledge the students with whom you discussed. This will not affect your grade. Failure to acknowledge is considered *academic dishonesty*, and it will affect your grade.
- Any student found guilty of academic dishonesty will receive "F" in the. course.
- When proofs are required, make them both clear and rigorous. Do not hand wave. Even when proofs are not required, you should justify your answers and explain your work.
- Late homeworks are not accepted.