Test 1: Similarity, Clustering, Frequent Items - PRACTICE

NAME: FINAL SCORE:

This test allows one 8×11.5 inch notes sheet, front and back. Nothing electronic that can transmit/receive information is not (e.g., computers, phones, calculators, ipads). Unlimited blank scratch paper is allowed.

Absolutely no talking allowed, unless a TA is present and you are asking a question. Talking students will have their tests confiscated.

1 k-Grams and Jaccard (35 points)

Consider the following two phrases:

$$P_1: \texttt{mississippiriver}$$
$$P_2: \texttt{mississippistate}$$

A: (15 points) List the sets of distinct 3-character-grams for phrases P_1 and P_2. This should produce two sets S_1 and S_2.

B: (10 points) What is the Jaccard Similarity $\mathsf{JS}(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$ between S_1 and S_2?

C: (10 points) What is the Andberg Similarity $\mathsf{Andb}(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cap S_2| + 2|S_1 \triangle S_2|}$ between S_1 and S_2?

2 MinHashing (20 points)

Consider three sets T_1, T_2, and T_3, and the 10-dimensional min-hash vectors for each set.

$$T_1: \; v_1 = (\texttt{17, 3, 4, 11, 07, 03, 02, 10, 72, 18})$$
$$T_2: \; v_2 = (\texttt{11, 3, 4, 01, 17, 30, 12, } \mathbf{11}\texttt{, 41, 24})$$

*(Note: **each entry** in a vector can be thought of as the result of creating a table, permuting the rows, and then returning the index of the highest element in the set.)*

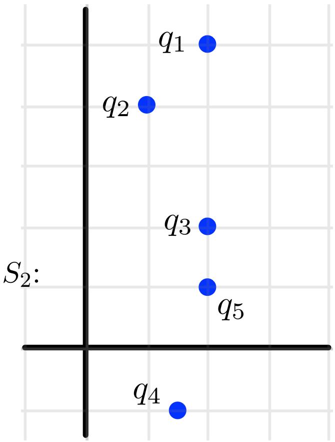
A: (10 points) Use v_1 and v_2 to estimate the Jaccard Similarity between T_1 and T_2?

B: (5 points) Will changing the 8th entry in v_2 (shown in bold) from an 11 to a 10 decrease the estimated Jaccard Similarity between T_1 and T_2? Explain your answer.

C: (5 points) Say we get a new set T_3, and we still remember the hash functions used to create v_1, and v_2. So we create a min-hash vector v_3 to represent T_3. What part of v_2 if any, do we need to recalculate to estimate the Jaccard similarity between T_3 and T_2? Explain your answer.

3 Clustering (45 points)

Consider 5 points, and mainly running a form of hierarchical agglomerative clustering (HAC) using the Euclidean distance:



$q_1 = (2, 5)$
$q_2 = (1, 4)$
$q_3 = (2, 2)$
$q_4 = (1.5, -1)$
$q_5 = (2, 1)$

We will consider two ways to measure the distance between clusters S_1 and S_2:

- Single-Link uses $d(S_1, S_2) = \min_{(s_1,s_2)\in S_1\times S_2} \|s_1 - s_2\|_2$ it measures the closest pair of points.
- Complete-Link uses $d(S_1, S_2) = \max_{(s_1,s_2)\in S_1\times S_2} \|s_1 - s_2\|_2$ it measures the furthest pair of points.

A: (10 points) Run HAC using Single-Link until there are 3 clusters. Report the resulting 3 clusters as sets.

B: (5 points) Run HAC using Single-Link until there are 2 clusters. Report the resulting 2 clusters as sets.

C: (10 points) Run HAC using Complete-Link until there are 3 clusters. Report the resulting 3 clusters as sets.

D: (5 points) Run HAC using Complete-Link until there are 2 clusters. Report the resulting 2 clusters as sets.

E: (10 points) Run Gonzalez Algorithm (the greedy algorithm for k-center clustering, assignment-based clustering) starting with the first center $c_1 = q_1$ until there are 2 clusters. Report the resulting 2 clusters as sets.

F: (5 points) Run Gonzalez Algorithm (the greedy algorithm for k-center clustering, assignment-based clustering) starting with the first center $c_1 = q_1$ until there are 3 clusters. Report the resulting 3 clusters as sets.