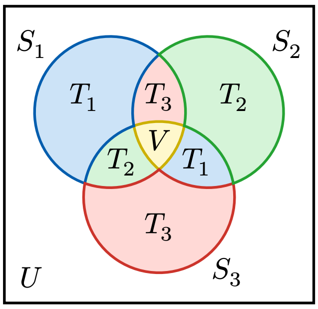
Distance between Sets

Define the *Jaccard distance* (measure of dissimilarity) between two sets S_i and S_j as:

$$d_J(S_i, S_j) = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|} = \frac{|S_i \triangle S_j|}{|S_i \cup S_j|}$$



Let $U = \bigcup S_i$ and $V = \bigcap S_i$. Define sets T_i:

$$T_1 = (S_1 \setminus (S_2 \cup S_3)) \cup ((S_2 \cap S_3) \setminus S_1)$$
$$T_2 = (S_2 \setminus (S_1 \cup S_3)) \cup ((S_1 \cap S_3) \setminus S_2)$$
$$T_3 = (S_3 \setminus (S_1 \cup S_2)) \cup ((S_1 \cap S_2) \setminus S_3)$$

Or more generally:

$$T_i = (S_i \setminus (S_j \cup S_k)) \cup ((S_j \cap S_k) \setminus S_i)$$

Let's prove that the Jaccard distance satisfies the *triangle inequality*:

$$d_J(S_i, S_j) + d_J(S_j, S_k) \geq d_J(S_i, S_k)$$

Step 1: The sum of T_i is exactly U without the triple intersection V:

$$|T_1| + |T_2| + |T_3| = |U| - |V|$$

Which can be rearranged to:

$$\frac{|T_1| + |T_2| + |T_3|}{|U|} = 1 - \frac{|V|}{|U|}$$

Step 2: Take any pair S_i, S_j and let k be the remaining index. Compute the symmetric difference:

$$\begin{aligned} S_i \triangle S_j &= (S_i \setminus S_j) \cup (S_j \setminus S_i) \\ &= ((S_i \setminus (S_j \cup S_k)) \cup ((S_i \cap S_k) \setminus S_j)) \cup \\ &\quad \cup ((S_j \setminus (S_i \cup S_k)) \cup ((S_j \cap S_k) \setminus S_i)) \\ &= T_i \cup T_j \end{aligned}$$

Since T_i are pairwise disjoint, we have $|S_i \triangle S_j| = |T_i| + |T_j|$. Therefore, the Jaccard distance is:

$$d_J(S_i, S_j) = \frac{|S_i \triangle S_j|}{|S_i \cup S_j|} = \frac{|T_i| + |T_j|}{|S_i \cup S_j|}$$

Step 3: Use two monotonicity facts about set sizes:

1. For the union, $|S_i \cup S_j| \leq |U|$. Dividing by a *smaller* number makes the fraction *larger*, so:

$$d_J(S_i, S_j) = \frac{|T_i| + |T_j|}{|S_i \cup S_j|} \geq \frac{|T_i| + |T_j|}{|U|}$$

2. For the intersection: $|S_i \cap S_j| \geq |V|$. Dividing by a *larger* number makes the fraction *smaller*, so:

$$\frac{|S_i \cap S_j|}{|S_i \cup S_j|} \geq \frac{|V|}{|U|}$$

Now recall (see Step 1) that $1 - |V|/|U| = (|T_1| + |T_2| + |T_3|)/|U|$, so this is the upper bound:

$$d_J(S_i, S_j) \leq \frac{|T_1| + |T_2| + |T_3|}{|U|}$$

Thus for every pair i, j we have the sandwich:

$$\frac{|T_i| + |T_j|}{|U|} \leq d_J(S_i, S_j) \leq \frac{|T_1| + |T_2| + |T_3|}{|U|} = 1 - \frac{|V|}{|U|}$$

Step 4: Combine the inequalities for the distances $d_J(S_1, S_2)$ and $d_J(S_2, S_3)$.

Using the lower bounds, we have:

$$d_J(S_1, S_2) + d_J(S_2, S_3) \geq \frac{|T_1| + |T_2|}{|U|} + \frac{|T_2| + |T_3|}{|U|} = \frac{|T_1| + 2 \cdot |T_2| + |T_3|}{|U|}$$

Since $2 \cdot |T_2| \geq |T_2|$, we have:

$$\frac{|T_1| + 2 \cdot |T_2| + |T_3|}{|U|} \geq \frac{|T_1| + |T_2| + |T_3|}{|U|} = 1 - \frac{|V|}{|U|}$$

But from the upper bound, we have $d_J(S_1, S_3) \leq 1 - |V|/|U|$. Therefore:

$$d_J(S_1, S_2) + d_J(S_2, S_3) \geq 1 - \frac{|V|}{|U|} \geq d_J(S_1, S_3)$$

Hence $d_J(S_1, S_2) + d_J(S_2, S_3) \geq d_J(S_1, S_3)$. The same argument works for any other permutation of indices, so the triangle inequality for the Jaccard distance is proven. $\square$