Q3 — Locality Sensitive Hashing (17 points)

Setup

- Essay P: |S_P| = 1500 unique shingles
- Publication Q: |S_Q| = 4500 unique shingles
- Shared shingles: |S_P ∩ S_Q| = 1000
- Union: |S_P ∪ S_Q| = 1500 + 4500 − 1000 = **5000**

Part 1 (6 points) — Jaccard Distance & MinHash

Jaccard similarity:

$$J(S_P, S_Q) = \frac{|S_P \cap S_Q|}{|S_P \cup S_Q|} = \frac{1000}{5000} = \boxed{\frac{1}{5}}$$

Jaccard distance:

$$d_J = 1 - J = 1 - \frac{1}{5} = \boxed{\frac{4}{5}}$$

MinHash probability (by the fundamental MinHash property):

$$\Pr[\text{MinHash}(S_P) = \text{MinHash}(S_Q)] = J(S_P, S_Q) = \boxed{\frac{1}{5}}$$

Part 2 (8 points) — Cosine Similarity & SimHash

Represent S_P, S_Q as binary vectors x_P, x_Q of length m (all possible shingles). Position i is 1 iff the document contains shingle i.

Dot product:

$$x_P \cdot x_Q = |S_P \cap S_Q| = 1000$$

(Both are 1 only at positions corresponding to shared shingles.)

Norms:

$$\|x_P\| = \sqrt{|S_P|} = \sqrt{1500}, \quad \|x_Q\| = \sqrt{|S_Q|} = \sqrt{4500}$$

Cosine similarity:

$$\cos\theta = \frac{x_P \cdot x_Q}{\|x_P\| \cdot \|x_Q\|} = \frac{1000}{\sqrt{1500} \cdot \sqrt{4500}} = \frac{1000}{\sqrt{6{,}750{,}000}} = \frac{1000}{1500\sqrt{3}} = \boxed{\frac{2\sqrt{3}}{9} \approx 0.385}$$

Simplification: 1500 × 4500 = 1500² × 3, so √(1500² × 3) = 1500√3. Then 1000/(1500√3) = 2/(3√3) = 2√3/9.

SimHash probability (using the given hint):

$$\theta = \arccos\left(\frac{2\sqrt{3}}{9}\right) \approx 1.175 \text{ radians}$$

$$\Pr[\text{SimHash}(x_P) = \text{SimHash}(x_Q)] = 1 - \frac{\theta}{\pi} = \boxed{1 - \frac{\arccos(2\sqrt{3}/9)}{\pi} \approx 0.626}$$

Part 3 (3 points) — Which Metric to Pick?

Metric	Similarity	Pr(hash match)
Jaccard + MinHash	1/5 = 0.20	0.20
Cosine + SimHash	2√3/9 ≈ 0.385	≈ 0.626

Answer: Cosine similarity + SimHash

Reasoning: In plagiarism detection, the essay P (1500 shingles) is much smaller than the publication Q (4500 shingles). Jaccard similarity is penalized by the large union — even though 1000/1500 = 2/3 of P's content comes from Q, Jaccard only reports 0.2. Cosine similarity normalizes by each document's individual size (via the norms), yielding 0.385 — nearly double the Jaccard value. The SimHash match probability of 0.626 vs. 0.20 makes plagiarism significantly easier to detect.

Key insight: When document sizes are asymmetric (small essay vs. large publication), cosine similarity is far more suitable because it doesn't dilute the signal with the larger document's unrelated content.