Disclaimer

I wrote this to my best knowledge, however, no guarantees are given whatsoever.

Sources

If not noted differently, the source is the lecture slides and/or the accompanying book.

1 Approximate Retrieval

Nearest-Neighbor Find $x^* = \text{argmin}_{x\in X}\ d(x,y)$ given S, $y \in S$, $X \subseteq S$.

Near-Duplicate detection Find all $x,x' \in X$ with $d(x,x') \le \epsilon$.

1.1 k-Shingling

Represent documents (or videos) as set of k-shingles (a. k. a. k-grams). *k-shingle* is a consecutive appearance of k characters/words.

Let there be N documents and C k-shingles.
Binary *shingle matrix* $M \in \{0,1\}^{CxN}$ where $M_{i,j} = 1$ iff document j contains shingle i.

1.2 Distance functions

General $d : S \times S \to \mathbb{R}$ is a *distance function* iff $\forall x,x',x'' \in S$ it's positive definite except for $x = x'$ ($d(x,x') > 0 \iff x \neq x'$ and $d(x,x) = 0$), symmetric ($d(x,x') = d(x',x)$) and satisfies the Cauchy-Schwartz triangle inequality ($d(x,x'') \le d(x,x') + d(x',x'')$).

L_r-norm $d_r(x,y) = (\sum_i |x_i - y_i|^r)^{1/r}$. L_2-norm also called *Euclidean*.

Cosine similarity $\text{Sim}_c(A,B) = \dfrac{A \cdot B}{|A| \cdot |B|}$.

Cosine distance $d_c(A,B) = \dfrac{\arccos(\text{Sim}_c(A,B))}{\pi}$.

Jaccard similarity $\text{Sim}_J(A,B) = \dfrac{|A \cap B|}{|A \cup B|}$.

Jaccard distance $d_J(A,B) = 1 - \text{Sim}_J(A,B) = 1 - \dfrac{|A \cap B|}{|A \cup B|}$.

1.3 LSH – local sensitive hashing

Key Idea: Similiar documents have similiar hash.
Note: Trivial for exact duplicates (hash-collisions $\to$ candidate pair).

Min-hash $h_\pi(C)$ Hash is the *minimum (i. e. first) row index* with a one after permutation: $h_\pi(C) = \min_{i,C(i)=1} \pi(i)$, given binary vector C and (random) permutation π.
Note: $\Pr_\pi[h_\pi(C_1) = h_\pi(C_2)] = \text{Sim}_J(C_1,C_2)$ if $\pi \in_{\text{u.a.r.}} S_{|C|}$.

Min-hash signature matrix $M_S \in [N]^{n \times C}$ with $M_S(i,c) = h_i(C_c)$ given n hash-fns h_i drawn randomly from a universal hash family.

Pseudo permutation h_π with $\pi(i) = (a \cdot i + b) \mod p \mod N$, N number of shingles, $p \ge N$ prime and $a,b \in_{\text{u.a.r.}} [p]$ with $a \neq 0$.
Instead of real permutations (slow, inefficient, large storage) use pseudo permutations as hash family. Pseudo permutations only need to store a and b.

Compute Min-hash signature matix M_S For all columns $c \in [C]$ and rows $r \in [N]$ with $C_c(r) = 1$, set $M_S(i,c) = \min\{h_i(C_c), M_S(i,c)\}$ for all hash functions h_i.

Banding as boosting Reduce FP/FN by AND/OR-boosting, respectively.
This is done by grouping the signature matrix into b bands of r rows each. A candidate pair matches in at least one band completely. This corresponds to a b-way OR after a r-way AND boosting.

2 More stuff to come