Assignment 3
COMP 3801- Fall 2021

1 Instructions

Each question is worth 12.5 marks, but the maximum you can get is 100.
The assignment is due in Brighspace by November 19th before 11:59PM. Please write clearly and
answer questions precisely. As a thumb rule, the answer should be limited to ≤ 2 written pages,
with ample spacing between lines and in margins, for each question. Always start a new question
on a new page, starting with Question 1, followed by Question 2, ..., Question n. Please cite all the
references which you have used/consulted as the source of information for each of the questions.
If a question asks you to Prove or Show, please clearly spell out the proof - the technique used -
and show each step of the proof. Don’t expect (partial) marks if the main idea isn’t clear to us.

2 Problems

1. Prove that the Jaccard distance satisfies the metric properties. Recall that the Jaccard
distance between two sets $S$ and $T$ is defined to be $JD(S, T) = 1 - \frac{|S \cap T|}{|S \cup T|}$. Note that the metric
properties are (a) identity: $JD(S, T) = 0 \Leftrightarrow S = T$, (b) Symmetry: $JD(S, T) = JD(T, S)$,
and (c) Triangle Inequality: For three sets $S, T, U$, $JD(S, T) \leq JD(S, U) + JD(U, T)$.
(Hint: We related Jaccard similarity to the probability that the MinHash signatures match
for two sets.)

2. This problem is related to the similarity of documents using the locality-sensitive hashing
technique. Assume that we have the signature matrix for documents consisting of $n$ rows.
We partitioned the rows in $b$ bands for the standard banding technique, where each band
consists of $r$ rows. Assume that $n = br$. If the two sets $S$ and $T$ have the same signatures in
all the rows in any band, we identify $S$ and $T$ to be similar. This scheme follows the AND-OR
construction (AND corresponds to what happens within a band, and OR corresponds to what
happens among the bands). We showed that if the Jaccard similarity between $S$ and $T$ is
$s = \frac{|S \cap T|}{|S \cup T|}$, then the probability that $S$ and $T$ will be identified similar is $1 - (1 - s)^b$. Suppose
that we modify the banding technique by reversing the roles of AND and OR. We say that
two sets $S$ and $T$ are similar with respect to a band if their signatures agree in at least one of
the rows of that band. We report that the two sets are similar if their signatures agree in all
the bands. Estimate the probability of reporting that $S$ and $T$ are similar with this modified
scheme. Let us call the modified scheme as OR-AND scheme.

3. Suppose we modify the standard Banding technique into a 2-level scheme as follows. The
1st-level partitions the $n$ rows of the signature matrix into $b$ bands, and each band consists of
$r$ rows. In the 2nd level, rows within each 1st level band are further partitioned into $b'$ bands,
where each of the sub-bands consists of $r'$ rows. Assume that $r = b'r'$. Within each 1st level band, we follow the OR-AND scheme of the previous problem, with parameters $b'$ and $r'$.

Eventually, two sets $S$ and $T$ are reported to be similar if all of the 1st level bands identify them to be similar. Estimate the probability of identifying that the two sets with Jaccard similarity $s$ are similar using the 2-level scheme. Let us call this scheme the AND-OR-AND scheme.

4. Take at least 5 different sets of values for $n, r, b, b'$ and tabulate the probability of reporting that the two sets $S$ and $T$ are similar when their Jaccard similarity is $s = 0, 1, 2, 4, 5, 6, 8, 9, 1$ for the three schemes outlined above: (a) Standard AND-OR scheme, (b) Modified OR-AND scheme, and (c) 2-level AND-OR-AND scheme. Looking at the table, can you recommend which of these schemes is best suited for finding similar documents using the MinHash signatures. Please choose some meaningful values for these parameters. Justify your answer.

5. Suppose we have a case of fake currency bills of $100 circulating in the market. The association of ⟨FraudBusters⟩ employs a cumbersome method that is computationally inefficient. It is based on scanning and digitizing the bill. Let us assume that the digital copy has a resolution of $2048 \times 1536$ pixels (approximately 3 million pixels). To check whether the bill is actual, this organization compares each pixel of the digital copy with the corresponding pixel of a digitized copy of a true bill. It is given that if the bill is original, then the probability that any pixel will match with the corresponding pixel of the true bill is 97%. If the bill isn’t original, then the likelihood of the match decreases to 69%. Given this information, can you devise an LSH based scheme to determine most of the fake bills?

6. This question is about Stream Statistics Over Sliding Windows. You need to determine the value of $r$, where we use up to $r \geq 2$ buckets of type $B_i$ for $i \geq 0$ so that the count of the number of 1s reported by the algorithm is within 5% of the actual count of 1s among the last $N$ bits seen in a data stream. (E.g. if there are 500 1s in the last $N$ bits, your answer must be in the range $500 \pm 25$.) Justify your choice of $r$. Furthermore, analyze the total space used by the data structure that maintains all the required buckets to achieve the desired accuracy?

7. In the Basic Counting algorithm for counting the number of 1s among the last $N$ bits received in a binary data stream, why do we need to make multiple copies of the buckets of type $B_i$ for $i \geq 0$? What will happen in the analysis if we only take one copy of each of the bucket types?

8. Assume that we have a stream consisting of numbers from the set $\{-1, 0, +1\}$ and we are interested in maintaining the sum of the last $N$ bits of the stream. In this exercise, we will show that we need $\Omega(N)$ bits to maintain an approximate sum that is within a constant factor of the exact sum. Suppose we have an algorithm $\mathcal{A}$ that maintains the approximate sum. Assume that we have a bit string consisting of $\frac{N}{2}$-bits composed of 0s and 1s. We replace each 0-bit by a pair of bits $(1, -1)$ and each 1-bit by the pair $(-1, 1)$. Now this sequence of $N$-bits is presented to our algorithm $\mathcal{A}$ that maintains the approximate sum within a constant factor. Note that the exact sum of these $N$-bits is 0. In addition to these $N$ bits, the next set of $N$ bits received in the stream are only 0-bits. Answer the following:

(a) Show that if the next bit (i.e. the $(N + 1)$-st bit) in the stream is 0, the output to the sum query on receiving this bit will be $+1$ (respectively $-1$) if and only if the 1st bit in
10. (Bonus Problem) This problem is about classifying vectors that are oriented in a similar direction. Assume that we have a collection $V$ of $d$-dimensional vectors from the origin and assume that $d$ is large. If we take any two distinct vectors $p, q \in V$, there is a unique plane $\Pi_{pq}$ passing through the origin that contains $p$ and $q$. (You may want to visualize this first for the vectors in three dimensions.) Suppose $p = (p_1, p_2, \ldots, p_d)$ and $q = (q_1, q_2, \ldots, q_d)$. The angle $\theta$ between $p$ and $q$ is related to their dot-product by the expression $\cos(\theta) = \frac{p \cdot q}{|p||q|}$, where $|p|$ and $|q|$ is the length of vectors $p$ and $q$. For example, if $p = (1, 3)$ and $q = (-3, 4)$, $|p| = \sqrt{1^2 + 3^2} = \sqrt{10}$ and $|q| = \sqrt{3^2 + 4^2} = 5$ and $\cos(\theta) = \frac{p \cdot q}{|p||q|} = \frac{1 \cdot -3 + 3 \cdot 4}{5\sqrt{10}} = \frac{9}{5\sqrt{10}}$, or $\theta \approx 55^\circ$.

Choose a random vector $\eta_h$ at origin and consider the (hyper)plane $h$ (i.e., plane in 3-dimensions, or a line in 2-dimensions) normal to $\eta_h$ passing through origin. Any hyperplane $h$ partitions the space into two half spaces, usually referred to as $h^+$ and $h^-$. Assume that none of the vectors $p$ and $q$ are incident on $h$. The vectors $p$ and $q$ may lie in the same half-space with respect to $h$ or one resides in $h^+$ and the other in $h^-$. It should also be clear that if the angle between $p$ and $q$ is small, then there is more chance that $p$ and $q$ will reside in the same half-space of $h$, as compared to when the angle is large. Observe that if the dot products $p \cdot \eta_h$ and $q \cdot \eta_h$ have the same sign, then $p$ and $q$ are within the same half-space of $h$. Otherwise, they are in different half-spaces.

Suppose we want to identify vectors that are close together, i.e., they point approximately in the same direction. (For example, if we take a cone of small-angle centred in the direction of vector $u$ with the apex at origin, we want to find all the vectors that are within this cone. All these vectors are pointing in (approximately) the same direction as $u$. See the Figure.) For a random vector $\eta_h$ and its corresponding normal hyperplane $h$, we assign a signature $h(u) \in \{+, -\}$ to each vector $u \in V$ based on the sign of $u \cdot \eta_h$. For any two vectors $u, v \in V$, we say $h(u) = h(v)$ if sign of $u \cdot \eta_h$ is the same as the sign of $v \cdot \eta_h$. Show that $Pr(h(u) = h(v)) = 1 - \frac{\theta}{\pi}$, where $\cos(\theta) = \frac{u \cdot v}{|u||v|}$. Observe that if the angle between $u$ and $v$ is small, there is high chance that $h(u) = h(v)$. We can construct a family of hash functions by choosing several random hyperplanes $h_1, \ldots, h_n$, and use the banding technique to amplify the stream was a 1 (respectively, $-1$).

(b) For a positive integer $i < \frac{N}{2}$, show that after receiving the $(N + 2i - 1)$-th 0 bit, the output to the sum query will be $+1$ (respectively $-1$) if and only if the $i$-th bit in the stream was a 1 (respectively, $-1$).

(c) Show that after receiving the $2N$-th 0 bit, we would have completely recovered the first $N$-bits of the stream.

(d) Conclude that to estimate the approximate sum within a constant factor in a sliding window of size $N$ in a stream of (positive and negative) numbers, we need to store $\Theta(n)$ bits.

9. (Bonus Problem:) We have considered several problems with data streams where we make one pass over all the stream elements to answer some relevant questions (e.g. frequency statistics) using only small storage space. Assume that you are allowed to make two passes over the elements of the data stream. Can you think of some problems that can now be addressed while maintaining the small storage space that wasn’t possible with a single pass?
the probability of classifying the similar vectors. (You may want to look at Section 3.7.2 in the MMDS book for further details.)