Universitat Politècnica de Catalunya
Facultat d'Informàtica de Barcelona

Degree: Grau en Enginyeria Informàtica **Academic year:** 2024–2025 (Final Exam)
Course: Randomized Algorithms (RA-MIRI) **Date:** January 17th, 2025
Time: 2 h 30 min

This exam has a total of 12 points in 4 questions. Each exam has 3 pages printed in 2 physical sheets of paper. Your score will be the sum of the points of the questions below, capped to a maximum of 10 points.

The exam poses several questions, all around the following scenario. We have a large collection of documents $\mathcal{D} = \{D_1, \ldots, D_T\}$. Each document D_i consists of a sequence of *words* (more generally, *shingles*): $D_i = z_1^{(i)}, z_2^{(i)}, \ldots, z_{N_i}^{(i)}$. The set of distinct words in D_i is the *vocabulary* $V_i = V(D_i)$ of the document. There exists a predefined list SW of frequently occurring words such as *a*, *and*, *that*, *the*, ... which will likely appear in most or all the documents, hence their presence in a given document conveys little or no information (these words are known as *stopwords*). The *normalized vocabulary* for document D_i is

$$V_i' = V_i \setminus SW = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}\}; \qquad n_i = |V_i'|$$

We denote $n_{\max} = \max_{1 \le i \le T}\{n_i\}$. Although $n_{\max}$ can be equal or close to $N_{\max} = \max_i\{N_i\}$, we will often have $n_{\max} \ll N_{\max}$. Likewise, the total size of the document collection $N = \sum_{1 \le i \le T} N_i$ will be often much less than $T \cdot N_{\max}$, since there are many more short documents than large ones.

In all questions, unless it is stated otherwise, you must justify your answers.

---

1. **(3 points)** For each document D_i, we will build a Bloom filter B_i as a sketch for the document. All Bloom filters are bitvectors of $M = \Theta(n_{\max})$ bits and use the same set of k hash functions.

```
procedure CreateSketch(D, SW)
    F := BloomFilter(M, k)
        ▷ Creates an empty Bloom filter
    for z ∈ D do
        if z ∉ SW then
            F.Insert(z)
        end if
    end for
    return F
end procedure
```

```
procedure AltCreateSketch(D, SW)
    F := BloomFilter(M, k)
    count := 0
    for z ∈ D do
        if z ∉ SW ∧ ¬F.Contains(z) then
            F.Insert(z)
            count := count + 1
        end if
    end for
    return F
end procedure
```

(a) What are the (expected) complexities of the functions CreateSketch(D, SW) and AltCreatSketch(D, SW) in terms of $|D|$, $|SW|$ and the size of the normalized vocabulary V' of D? Do they significantly differ? Which data structure would you use for SW to get these complexities?

(b) What is the difference, if any, between the Bloom filters produced by the calls $B_i :=$ CreateSketch(D_i, SW) and $B'_i :=$ AltCreateSketch(D_i, SW)? The parameters M and k (and the k hash functions) used for both B_i and B'_i are the same.

(c) What is the expected value of `count` after document D_i has been processed with algorithm AltCreateSketch? Give a lower bound using the probability of false positives. The probability of false positives evolves as you keep inserting elements in the Bloom filter, but you can use a pessimistic upper bound based on the maximum number of inserted elements.

Give also the answer for the particular case in which we set $M = 500000$, $k = 7$ and we can take for granted that $n_i \leq 50000$. Give the simplest possible form you can; but it's OK to leave your answer without carrying a few final floating point calculations, e.g., your expression might contain subexpressions such as $\sqrt{2 + e^{-2.4}}$.

---

2. **(4 points)**

(a) Devise an estimator for n_i using the number of 1s in the bitvector held in B_i. Use the approximation $\mathbb{E}[f(X)] \approx f(\mathbb{E}[X])$ to propose your estimator; you need not to prove that the estimator is (asymptotically) unbiased, nor to compute an explicit factor to compensate the bias.

(b) Given two documents D_i and D_j, $i \neq j$, their similarity is defined as

$$\sigma_{ij} = \text{Jaccard}(V'_i, V'_j) = \frac{|V'_i \cap V'_j|}{|V'_i \cup V'_j|}.$$

Provide a suitable scheme to estimate σ_{ij} given the "sketches" B_i and B_j of D_i and D_j, computed using CreateSketch (with identical bitvector size and hashes for both Bloom filters). That is, propose an estimator $\hat{\sigma}_{ij}$ such that $\mathbb{E}[\hat{\sigma}_{ij}] \approx \sigma_{ij}$. You have not kept vocabularies V'_i and V'_j, and you cannot process (again) the documents. Can you directly estimate $|V'_i \cap V'_j|$? What about $|V'_i \cup V'_j|$?

(c) What is the computational complexity of getting all possible $\binom{T}{2}$ estimates $\hat{\sigma}_{ij}$ for pairwise similarities, compared to getting all $\binom{T}{2}$ exact similarities σ_{ij}? Your computation of the costs of both alternatives must include the costs of processing every document to create a sketch or to extract its normalized vocabulary.

Assume that you have enough memory to store all normalized vocabularies V_i', $1 \leq i \leq T$, once you have extracted them. Express the costs in terms of N, T, and $n_{\max}$. Recall that the size of B_i is M bits, with $M = \Theta(n_{\max})$.

---

3. **(2 points)** The similarity between two documents can be the result of chance; we want to compute, even if only approximately, the expected similarity of two "random" documents. A random document with normalized vocabulary V_i' of size n_i is a random sample of n_i distinct words drawn without replacement from some *base vocabulary* $\mathcal{V}$ (N.B. the base vocabulary does not contain any stopword: $\mathcal{V} \cap SW = \emptyset$). The base vocabulary contains W distinct words. To get the sought answer, use

$$\mathbb{E}[\sigma_{ij}] \approx \frac{\mathbb{E}\big[|V_i' \cap V_j'|\big]}{\mathbb{E}\big[|V_i' \cup V_j'|\big]}.$$

Express your answer in terms of the relative sizes $\rho_i := n_i/W$ and $\rho_j := n_j/W$. A useful formula to express your final answer in a simple way is $x/(1-x) \sim x + \mathcal{O}(x^2)$, when $x \to 0$.

---

4. **(3 points)** Define a discrete finite Markov chain such that its states are the documents $\mathcal{D} = \{D_i\}_{1 \leq i \leq T}$ and let the transition probabilities $p_{i,j} = \frac{\sigma_{ij}}{\sigma_i}$, where $\sigma_i = \sum_{1 \leq k \leq T} \sigma_{ik}$.

   (a) Give sufficient conditions for the Markov chain to be regular. Express these conditions in terms of properties to be satisfied by the document set $\mathcal{D}$, avoid phrasing them in terms of the directed graph underlying the Markov chain. Aim for conditions which are not very restrictive. If all or some of the conditions are always met by any document set, tell so and justify why.

   (b) Assuming that the previous Markov chain is regular, show that $\pi = (\pi_1, \ldots, \pi_T)$ is the unique stationary distribution with

$$\pi_i = \frac{\sigma_i}{\sigma},$$

   and $\sigma = \sum_{1 \leq i \leq T} \sigma_i = \sum_{1 \leq i \leq T} \sum_{1 \leq j \leq T} \sigma_{ij}$.
