Big Data Analytics: HOMEWORK 2
Professor: Mohamed Ndaoud
DUE ON: Monday, DECEMBER 5.

The homework can be done in groups of 4 students. Please type your solutions (LaTeX and code) on a Jupyter Notebook that you will upload on Moodle. Make sure to write the names of the 4 students of your group in the top of your notebook. It is enough that one students uploads the work of the whole group.
In case you struggle writing Latex on a Jupyter notebook, you can return a (single) pdf file that includes both the executed code and your solutions to the theoretical questions.

Problem 1, 30 points: (On the Flajolet-Martin Algorithm)

The goal of this algorithm is to compute an approximation of the number of distinct elements in a stream. Consider a stream S with N items. Let F be the number of distinct items in S, our goal is to estimate F. In order to do so, we have access to a perfect hash function h such that for all k in S, $h(k) \sim \text{Unif}[0, 2^w - 1]$ where $w = \lfloor \log(N) \rfloor$. In particular for $k_1 \neq k_2$ we have that $h(k_1)$ and $h(k_2)$ are independent.

We follow the following scheme:

- Let z_k be the number of 0 at the tail of the binary representation of $h(k)$.
- $Z = \max_{k \in S} z_k$.
- $\tilde{F} = 2^Z$.

For example, if $w = 5$ and $h(k) = 4 = (00100)_2$, $z_k = 2$. We want to prove that

$$\forall c \geq 3, \quad \mathbb{P}\left(1/c \leq \frac{\tilde{F}}{F} \leq c\right) \geq 1 - 3/c.$$

1. Show that for all $r \in [0, w]$, $\mathbb{P}(z_k \geq r) = 2^{-r}$.

   For $r \in [0, w]$ and $k \in S$, we define the random variable $X_k(r) = \mathbf{1}(z_k \geq r)$ and $X(r) = \sum_{\text{distinct } k \in S} X_k(r)$.

2. Compute the expectation and variance of $X_k(r)$ and $X(r)$.

   Let r_1 be the smallest integer r such that $2^r > cF$ and r_2 the smallest integer such that $2^r \geq F/c$.

3. Prove that the scheme is correct if $X(r_1) = 0$ and $X(r_2) \neq 0$.

4. Show that $\mathbb{P}(X(r_1) \geq 1) \leq 1/c$.

5. Show that $\mathbb{P}(X(r_2) = 0) \leq 2/c$.

6. Conclude.

Solution

1. Since $h(k)$ is uniformly distributed, the digits of its binary representation are independent Bernoulli variables. $z_k \geq r$ only if the last k digits are 0 or equivalently k independent bernoulli variables are 0. It comes out that for all $r \in [0, w]$, $\mathbb{P}(z_k \geq r) = 2^{-r}$.

2. We know that for Bernoulli(p) the expectation is p and variance $p(1-p)$. It comes out that $\mathbf{E}(X_k(r)) = 2^{-r}$ and $Var(X_k(r)) = 2^{-r}(1-2^{-r})$. Similarly we also get that $\mathbf{E}(X(r)) = F2^{-r}$ and $Var(X(r)) = F2^{-r}(1-2^{-r})$.

3. Let r_1 be the smallest integer r such that $2^r > cF$ and r_2 the smallest integer such that $2^r \geq F/c$. The scheme is correct if $1/c \leq \frac{\tilde{F}}{F} \leq c$. On one side, $1/c \leq \frac{\tilde{F}}{F}$ is equivalent to $2^Z \geq F/c$ or equivalently $Z \geq r_2$ and hence $X(r_2) \neq 0$. On the other side, $c \geq \frac{\tilde{F}}{F}$ is equivalent to $2^Z \leq cF$ or equivalently $Z < r_1$ and hence $X(r_1) = 0$.

   It comes out that the scheme is correct if $X(r_1) = 0$ and $X(r_2) \neq 0$.

4. $\mathbb{P}(X(r_1) \geq 1) \leq \frac{\mathbf{E}(X(r_1))}{1} \leq F2^{-r_1} \leq \frac{1}{c}$.

5. $\mathbb{P}(X(r_2) = 0) = \mathbb{P}(z_k < r_2)^F = (1 - 2^{-r_2})^F \leq \exp(-F2^{-(r_2-1)}/2) \leq \exp(-c/2) \leq 2/c$.

6. The scheme is wrong if $X(r_1) \geq 1$ or $X(r_2) = 0$. This event happens with probability at most $3/c$ based on the previous questions. Hence the scheme is correct with probability $1 - 3/c$.

Problem 2, 30 points: (Dead ends in PageRank computations)

Let the matrix of the Web M be an n by n matrix, where n is the number of Web pages. The entry m_{ij} in row i and column j is 0, unless there is an arc from node (page) j to node i. In that case, the value of m_{ij} is $1/k$, where k is the number of arcs (links) out of node j. Notice that if node j has $k > 0$ arcs out, then column j has k values of $1/k$ and the rest 0's. If node j is a dead end (i.e., it has zero arcs out), then column j is all 0's.

Let $r = [r_1, r_2, ..., r_n]^T$ be (an estimate of) the PageRank vector; that is, r_i is the estimate of the PageRank of node i. Define $w(r)$ to be the sum of the components of r; that is $w(r) = \sum_i r_i$.

In one iteration of the PageRank algorithm, we compute the next estimate r' of the PageRank as: $r' = Mr$. Specifically, for each i we compute $r'_i = \sum_j m_{ij} r_j$.

1. Suppose the Web has no dead ends. Prove that $w(r') = w(r)$.

2. Suppose there are still no dead ends, but we use a teleportation probability of $1 - \beta$, where $0 < \beta < 1$. The expression for the next estimate of r_i becomes $r'_i = \beta \sum_j m_{ij} r_j + (1 - \beta)/n$.

   Under what circumstances will $w(r') = w(r)$? Prove your conclusion.

3. Now, let us assume a teleportation probability of $1 - \beta$ in addition to the fact that there are one or more dead ends. Call a node "dead" if it is a dead end and "live" if not. Assume $w(r) = 1$. At each iteration, each live node j distributes $(1 - \beta) r_j / n$ PageRank to each of the other nodes, and each dead node j distributes r_j / n PageRank to each of the other nodes.

   Write the equation for r'_i in terms of β, M, r, n and D (where D is the set of dead nodes). Then, prove that $w(r')$ is also 1.

Solution

1. We need to show $\sum_i \sum_j m_{ij} r_j = \sum_i r_i$. Interchange order of summations so we have $\sum_j \left(\sum_i m_{ij}\right) r_j$ on the left. Since there are no dead ends, $\sum_i m_{ij} = 1$ for each j. Thus the equation holds.

2. If and only if $w(r) = 1$. To see why, sum over I to get $w(r') = \sum_i \beta \sum_j m_{ij} r_j + \sum_i (1-\beta)/n$. The second term on the right sums to $1-\beta$, and using the reasoning from part (1), the first term on the right sums to $\beta \sum_j r_j$. That is: $w(r') = \beta w(r) + (1-\beta)$. If $w(r) = 1$,

   then the equation tells us $w(r')$ is also 1. Conversely, if $w(r') = w(r) = x$, then x must satisfy

   the equation $x = \beta x + 1 - \beta$, from which it follows that $x = 1$.

3. The equation is $r'_i = \beta \sum_j m_{ij} r_j + (1-\beta)/n + \beta/n \sum_{\text{dead}_j} r_j$. If we sum over i and use the same trick of moving the summation on i to apply only to m_{ij} , we get $w(r') = \beta \sum_{\text{live}_j} r_j + (1-\beta) + \beta \sum_{\text{dead}_j} r_j$. The first and last terms on the right together give $\beta w(r)$, which is β, since $w(r)$ is assumed to be 1. Thus, the right side reduces to 1, as desired.