МФТИ, ФИВТ

Алгоритмы и структуры данных, осень 2022

Семинар №01. Базовые строковые алгоритмы

1. (Парадокс дней рождения) Пусть $|B| \geqslant |U|^2$, а функция $h : U \to B$ на элементах U принимает независимые значения, равномерно распределённые в B. Покажите, что $\mathbb{P}\left(\exists x \neq y : h(x) = h(y)\right) \leqslant 1/2$. Убедитесь, что чем больше $|B|$, тем меньше вероятность коллизии.

2. Китайская теорема об остатках гласит следующее: если $m_1, \ldots, m_k$ попарно взаимно просты, то для произвольного набора целых чисел $r_1, \ldots, r_k$ система сравнений $\{x \equiv r_i \pmod{m_i}\}_{i=1}^k$ имеет единственное решение среди чисел $\{0, 1, \ldots, p-1\}$, где $p = m_1 \cdot \ldots \cdot m_k$. Как можно расширить размер множества хешей до чисел порядка 10^{40}? *Указание:* Используйте несколько хеш-функций.

3. Почему префикс-функцию нельзя искать с помощью бинарного поиска и проверки равенства подстрок хешами? Можно ли так считать z-функцию?

4. Алгоритм Кнута—Морриса—Пратта ищет все вхождения шаблона s в текст t. Для этого рассматривается префикс-функция строки $s\#t$. Покажите, зачем нужен разделитель $\#$. Покажите, как оптимизировать алгоритм для потребления $O(|s|)$ памяти.

5. Строгим периодом строки s называется такая строка t минимальной длины, что существует целое $k \geqslant 1$, т. ч. $t^k = s$. Нестрогим периодом называется такая t минимальной длины, что существует целое $m \geqslant 1$, т. ч. $s \sqsubset t^m$ (то есть s является префиксом t^m). Найдите строгий и нестрогий периоды s за $O(|s|)$.

6. В строке s за $O(|s|^2)$ найдите количество её различных подстрок, то есть размер множества $\{s_l \ldots s_r \mid l \leqslant r\}$.

7. По z-функции строки постройте её префикс-функцию.

8. В данной строке s за $O(|s|)$ предподсчёта научитесь отвечать на запрос “равна ли подстрока $s_l \ldots s_r$ префиксу соответствующей длины?” за $O(1)$.

9. В данной строке найдите все префиксы, являющиеся палиндромами.

10. Дан шаблон p и текст t. Найдите все вхождения p в t не более чем с одной ошибкой (то есть нужно найти подстроки t, которые равны p с точностью до замены одного символа).

11*. По данной префикс-функции строки найдите количество строк над алфавитом мощности m с той же префикс-функцией.

1. Используйте union bound: вероятность существования не больше суммы вероятностей.
2. Возьмите 4 хеш-функции по 4 взаимно простым модулям порядка 10^{10}.
3. Для произвольного i множество таких j, что $s_j s_{j+1} \dots s_i = s_0 s_1 \dots s_{i-j}$, необязательно является отрезком.
4. Достаточно хранить префикс-функцию строки s, а также значение префикс-функции на последнем рассмотренном символе t.
5. Если $p[n-1]$ — последнее значение префикс-функции, рассмотрите $n - p[n-1]$.
6. Расширяйте строку s символами влево, насчитывайте префикс-функцию. Сколько новых строк добавляется после введения нового символа?
7. Если $z(i) \geqslant x$, то $\pi(i+x-1) \geqslant x$. Для каждого i поставьте пометку x в точке $i+x-1$, затем пройдите справа налево по строке.
8. Постройте дерево префикс-функции: если $\pi(k) = \ell$, проведите ребро $\ell \to k$. Тогда нужно будет научиться проверять, что одна вершина является предком другой.
9. Запишите подряд (или с разделителем) s и s^R, тогда нужно будет найти, какие суффиксы равны соответствующим префиксам.
10. В позиции i возьмите максимальную подстроку t, начинающуюся в i и равную префиксу p. Пропустите один символ. Проверьте, что остаток равен суффиксу p. Отдельно обработайте точные вхождения p.
11*. Я не знаю, как решать :(