LABORATORIUM NR 5

HASHOWANIE I

ZADANIE AiSD.P.L5.1 (2 pkt.) Celem zadania jest sprawdzenie różnych wariantów funkcji haszującej — chcemy zbadać wynik haszowania ciągu kluczy, które są napisami, z użyciem łańcuchowej metody usuwania kolizji.

Ogólny sposób postępowania jest następujący. Należy zdefiniować listę `T` rozmiaru m, gdzie każde `T[i]` będzie listą tych kluczy (napisów) k, dla których $h(k) = i$. Zatem najpierw trzeba wypełnić tablicę `T` pustymi listami, a potem dla kolejnych kluczy k wyliczać $h(k)$ i dodawać k do odpowiedniej listy: `T[h(k)].append(k)`.

- Klucze-napisy do testowania są w pliku `3700.txt`.
- Haszowanie modularne: $h(k) = \mathsf{liczba}(k) \bmod m$, gdzie liczba() to zdefiniowana przez nas funkcja wykonująca pseudolosową konwersję napisu na liczbę.
- Przykładowe schematy konwersji napisu na liczbę:

  a) $abcdef\ldots \to ((256 \cdot a + b) \text{ XOR } (256 \cdot c + d)) \text{ XOR } (256 \cdot e + f)\ldots$;

  b) $abc\ldots x \to (\ldots((111 \cdot a + b) \cdot 111 + c) \cdot 111 + \ldots) \cdot 111 + x$.

  W drugim schemacie liczba 111 to przykładowa stała (niebędąca potęgą dwójki).

Do tablicy `T` należy wstawić około $2m$ kluczy, po czym wypisać, jaka jest:

– liczba pustych list w tablicy `T`;

– maksymalna długość listy w `T`;

– średnia długość niepustych list w `T`.

TESTY. Należy przeprowadzić następujące testy:

$$W.17, D.17, S.17, W.1031, D.1031, S.1031, W.1024, D.1024, S.1024,$$

gdzie w symbolu testu litera oznacza warianty funkcji haszującej:

(W): wbudowana w Pythonie funkcja hash;
(D): własna dobra funkcja haszująca h według jednego ze schematów powyżej;
(S): własna słaba funkcja haszująca, np. tylko według pierwszej litery lub według sumy kodów liter klucza;

a liczba oznacza wariant rozmiaru tablicy:

(17): $m = 17$, z wydrukiem całej tablicy
(1031): $m = 1031$ (liczba pierwsza)
(1024): $m = 1024$ (potęga 2); ewentualnie $m = 1026$, czyli $m = 19 \cdot 3 \cdot 3 \cdot 2$.

Jako rozwiązanie przesłać kod programu wykonującego pomiary z dołączonymi wynikami

wygenerowanymi przez ten program dla wszyskich dziewięciu wariantów powyżej. Wyniki te można dołączyć np. jako komentarz na końcu kodu programu. Proszę też w pliku wpisać odpowiedź na następujące pytania:

(P.1) Który z rozmiarów tablicy (1031 lub 1024) dawał lepsze wyniki?
(P.2) W jaki sposób wybór rodzaju funkcji haszującej (W, D, S) wpływał na jakość wyniku?

Lepszy wynik to taki, że maksymalna i średnia (liczona po niezerowych wartościach) długość listy są mniejsze.