Algorytmy Tekstowe

Laboratorium 2 – raport

Mateusz Kocot

1. Format pliku

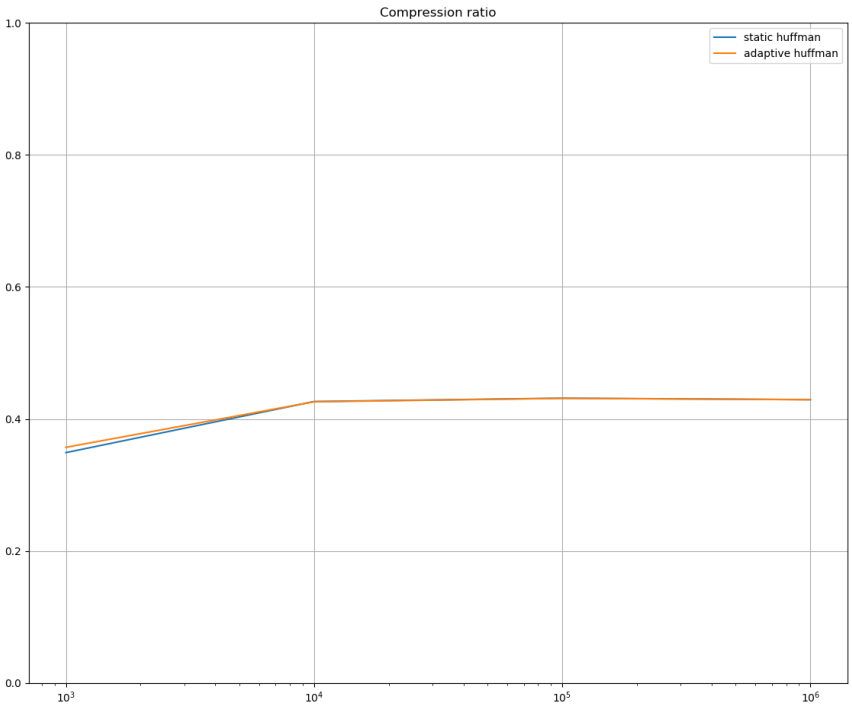
W przypadku kodowania dynamicznego, format pliku jest oczywisty (jest częścią algorytmu).

Dla kodowania statycznego, w pliku oprócz zakodowanego tekstu należy także zapisać strukturę drzewa. W moim rozwiązaniu wykorzystane zostało przeglądanie drzewa pre-order. Gdy napotykamy węzeł zewnętrzny, dodajemy 1 do wyniku i kontynuujemy przegląd. W przeciwnym przypadku, gdy napotykamy liść, dodajemy do wyniku 0 oraz kod ASCII odpowiedniej litery i wracamy do poprzedniego węzła zewnętrznego. Odpowiednią kolejność przeglądu zapewnia stos.

2. Testowane pliki

Testy przeprowadzono dla plików o rozmiarach $1kB$, $10kB$, $100kB$, $1MB$. Dla każdego rozmiaru wyróżniono tekst napisany językiem naturalnym (fragment „Pana Tadeusza”) oraz tekst wygenerowany losowo ze zbioru liter i cyfr. Wyniki dla drugiej grupy były bardzo podobne do wyników grupy pierwszej. Jedyną różnicą był fakt, iż wszystko działało nieco wolniej (z zachowaniem odpowiednich proporcji dla rozmiaru tekstu). Z tego powodu przedstawię tu tylko wyniki dla tekstów naturalnych.

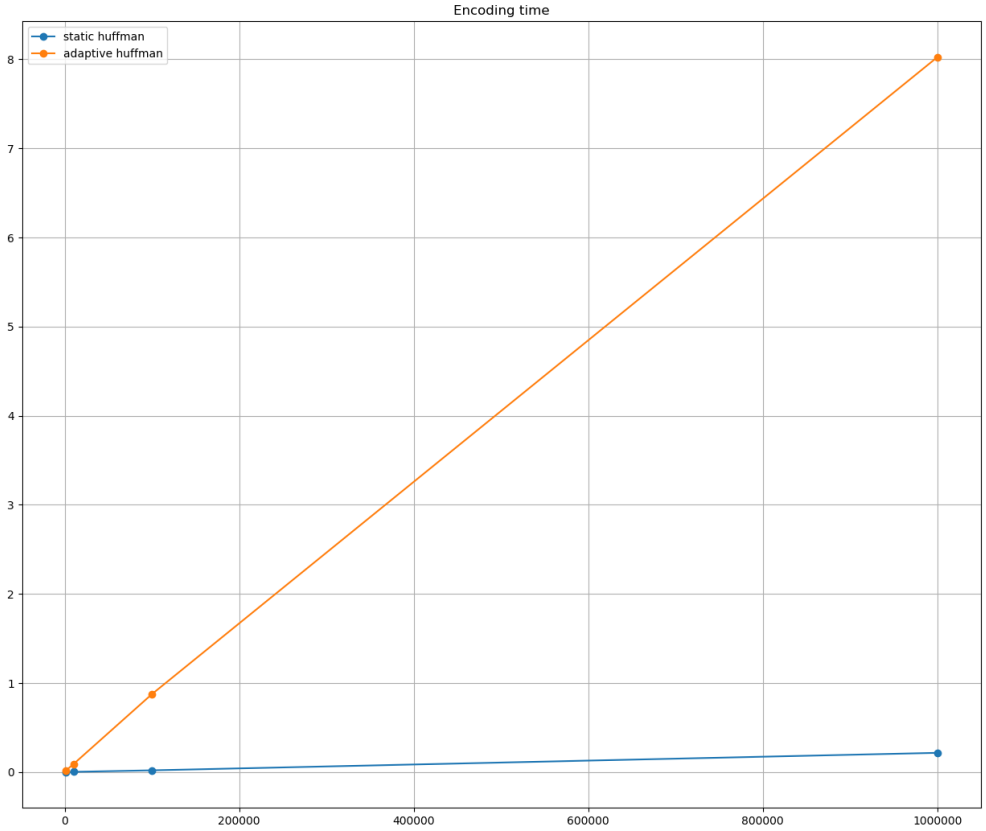
3. Współczynnik kompresji



Rys. 1. Wykres współczynnika kompresji od rozmiaru tekstu.

Wyniki dla obu sposobów kodowania są prawie identyczne. Gorszy współczynnik dla krótszego tekstu wynika w przypadku algorytmu statycznego z konieczności zakodowania struktury drzewa, a dla algorytmu dynamicznego – z konieczności kodowania znaków ASCII pierwszych wystąpień liter.

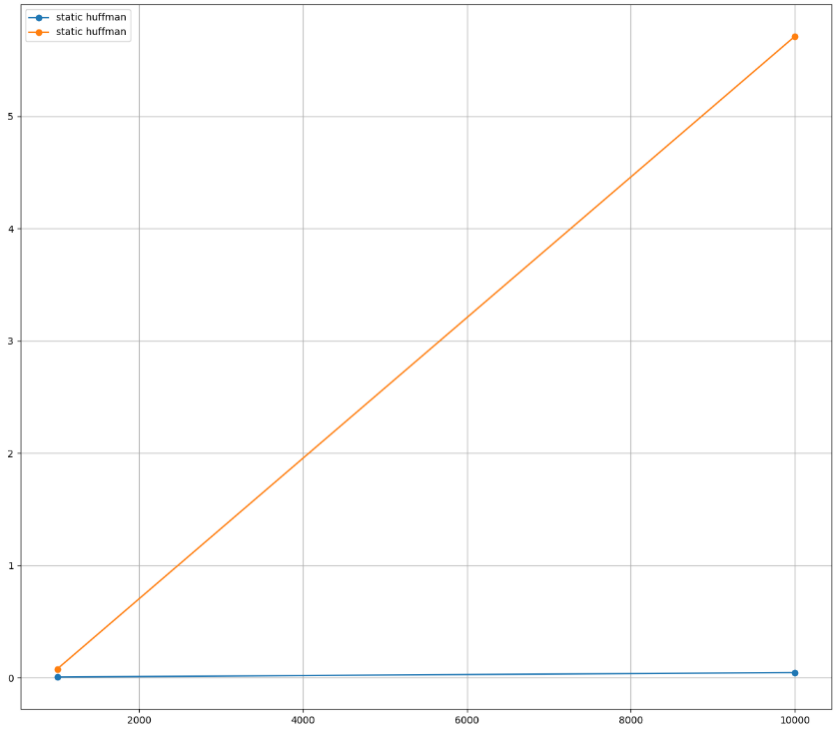
4. Czas kompresji



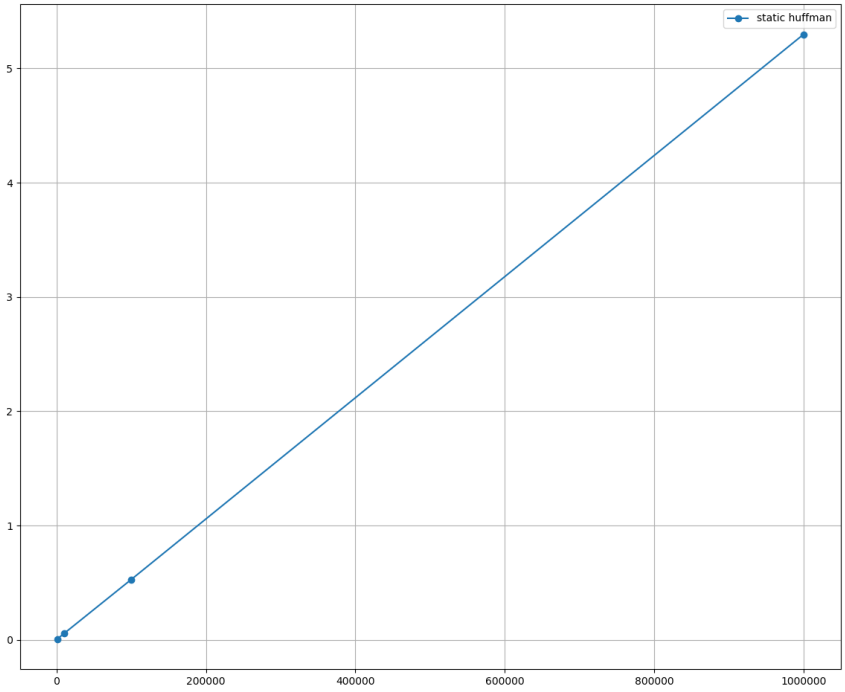
Rys. 2. Wykres czasu kompresji od rozmiaru tekstu.

Czas kompresji dla algorytmu dynamicznego jest znacznie większy od czasu dla algorytmu statycznego. Prawdopodobnie jest to spowodowane nieoptymalną implementacją. Jednakże, poprawne zaimplementowanie procedury $increment$ jest bardzo problematyczne i zajęło mi to bardzo dużo czasu. Zaletą mojej implementacji (algorytm Vittera) jest tworzenie mocno zrównoważonego drzewa, co pozytywnie wpływa na współczynnik kompresji (rys. 1).

5. Czas dekompresji



Rys. 3. Wykres czasu dekompresji od rozmiaru tekstu dla obu rodzajów kodowania.



Rys. 4. Wykres czasu dekodowania od rozmiaru tekstu dla kodowania statycznego

Stworzono dwa wykresy, gdyż czas dekodowania metodą dynamiczną jest bardzo długi. Wynika to najprawdopodobniej z nieoptymalnej implementacji. Ogólnie, czasy dekodowania dla obu algorytmów są dłuższe od czasów kodowania. Jest to spowodowane koniecznością przeglądania utworzonego drzewa w celu znalezienia liścia zawierającego odpowiednią literę.