**UNIWERSYTET GDAŃSKI**

**WYDZIAŁ MATEMATYKI, FIZYKI I INFORMATYKI**

**Wardziński Marcin**

*Kierunek: Informatyka*

*Numer albumu: 235423*

**Badanie skuteczności różnych odmian kodu Huffmana w języku polskim.**

Praca magisterska

napisana pod kierunkiem

prof. dr hab. Andrzeja Szepietowskiego

1. **Streszczenie**
2. **Spis Treści**

Streszczenie

Spis Treści

Wstęp

Rozdział 1: Kod Huffmana

Rozdział 2: Metody badawcze i opis narzędzi

Rozdział 3: Wyniki badań

Rozdział 4: Wnioski i podsumowanie

Bibliografia

2

3

4

5

1. **Wstęp**

**3.1 Wprowadzenie**

Niniejsza praca poświęcona jest zagadnieniom optymalizacji kompresji danych tekstowych przy użyciu kodów Huffmana. Kody te tworzone są z osobna dla każdego kompresowanego tekstu (a co za tym idzie, muszą być przesłane wraz ze skompresowaną wiadomością by można było ją zdekompresować) na podstawie algorytmu stworzonego w 1951, a opublikowanego rok później przez Davida Alberta Huffman’a, działającego w oparciu o prawdopodobnieństwo występowania poszczególnych znaków. Choć sam algorytm Huffmana obecnie praktycznie nie jest używany, ze względu na powstanie nowszych, optymalniejszych rozwiązań, stanowi on bazę dla wielu z nich, przez co one także bywają nazywane “kodami Hufmana”.

**3.2 Cel pracy**

Celem pracy jest zbadanie efektywności różnych odmian kodów Huffmana (kodowanie znaków pojedynczych, dwuznaków lub trójznaków, zastosowanie znacznika ELSE) i porównanie ich ze sobą w różnych sytuacjach (różne długości, tematyki i style tekstów). Eksperymenty obejmą również próby zastosowania kodu wygenerowanego na jednym tekście do kompresji innego, a co za tym idzie - sprawdzenia zasadności tworzenia ogólnego kodu dla poszczególnych grup tekstów, lub nawet dla całości języka. Dodatkowym celem jest także próba zaobserwowania różnic pomiędzy poszczególnymi kategoriami tekstów, możliwych do wykrycia w czasie procesu tworzenia kodu i kompresji.

Całość badań i rozważań dotyczyć będzie tekstów pisanych w języku polskim.

**3.3 Metody badawcze i narzędzia.**

Wyniki na bazie których oparta jest praca pozyskane zostały poprzez przeprowadzenie szeregu prób, polegających na tworzeniu kodu daną odmianą algorytmu Huffmana, skompresowaniu tekstu tym kodem i porównania rozmiaru tekstu przed i po kompresji. W części prób użyto kodu wygenerowanego w oparciu o kompresowany tekst, w innych tworzono kod w oparciu o jeden tekst by skompresować nim inny, w jeszcze innych kod tworzony był w oparciu o cały zbiór tekstów, uśredniając prawdopodobieństwo występowania danych znaków w nich wszystkich.

Wszystkie badania zostały przeprowadzone w oparciu o programy liczące prawdopodobieństwa występowania danych znaków i ich kombinacji, tworzące kody i kompresujące tekst stworzone przez autora pracy na potrzeby jej pisania.

**3.4 Podział pracy**

Praca składa się z czterech głównych części.

Pierwsza z nich zawiera szczegółowy opis algorytmu tworzenia kodów Huffmana i jego rozmaitych odmian użytych w czasie badań, przedstawia ogólną zasadę jego działania i najważniejsze właściwości, determinujące jego skuteczność i poprawność.

Druga część skupia się na wykonanych przez autora narzędziach. Opisuje ich konstrukcję, zależności między nimi i sposób ich wykorzystania.

Część trzecia zawiera opis poszczególnych eksperymentów, ich cele i przede wszystkim wyniki poszczególnych prób - zarówno suche liczbowe dane, jak i dostrzeżone wyjątki i zaobserwowane prawidłowości.

Ostatni rozdział to podsumowanie całości badań, wyciągnięte z analizy danych konkluzje i wnioski. Zamieszczone są w nim także propozycje ustandaryzowanych kodów Huffmana dla poszczególnych kategorii tekstów.