Politechnika Wrocławska

Wydział Elektroniki, Fotoniki i Mikrosystemów

Projektowanie Algorytmów i Metody Sztucznej Inteligencji

Projekt 2

Zadanie na ocenę bdb

Prowadzący:
Dr inż. Łukasz Jeleń

Wykonała:
Zuzanna Mejer, 259382

Wrocław, 6 maja 2022r.

Spis treści

1. Wprowadzenie

Zadanie miało na celu zapoznanie się z algorytmami sortowania oraz przeprowadzenie analizy efektywności wybranych i zaimplementowanych sortowań. Z wymienionych algorytmów wybrałam sortowania: przez scalanie, szybkie oraz introspektywne.

2. Opis badanych algorytmów i ich złożoność obliczeniowa

2.1. Sortowanie przez scalanie

Jest to rekurencyjny algorytm sortowania danych, stosujący metodę „dziel i zwyciężaj". W algorytmie wyróżnia się trzy podstawowe kroki: podział danych wejściowych na 2 rozłączne podzbiory; rekurencyjnie zastosowanie sortowania dla każdego podzbioru, aż do uzyskania struktur jednoelementowych; scalenie posortowanych podzbiorów w jeden zbiór. Całkowita złożoność obliczeniowa dla sortowania przez scalanie wynosi $O(n \cdot logn)$, w związku z czym zastosowanie tego sortowania okaże się wydajniejsze dla bardzo dużych tablic.

2.2. Sortowanie szybkie

Również jest to algorytm sortowania danych stosujący metodę „dziel i zwyciężaj", nie wykorzystuje on jednak dodatkowych podtablic. Istnieje wiele implementacji sortowania szybkiego, jednak generalna idea jest taka, że wybierany jest jeden element w sortowanej strukturze, który nazywany jest piwotem. Może być to element środkowy, pierwszy, ostatni bądź losowy, przy czym należy pamiętać, że w przypadkach, kiedy piwot jest ciągle maksymalny lub minimalny, występuje najgorsza złożoność obliczeniowa $O(n^2)$. Przy optymalnych wyborach piwotu, złożoność wynosi $O(n \cdot logn)$.

2.3. Sortowanie introspektywne

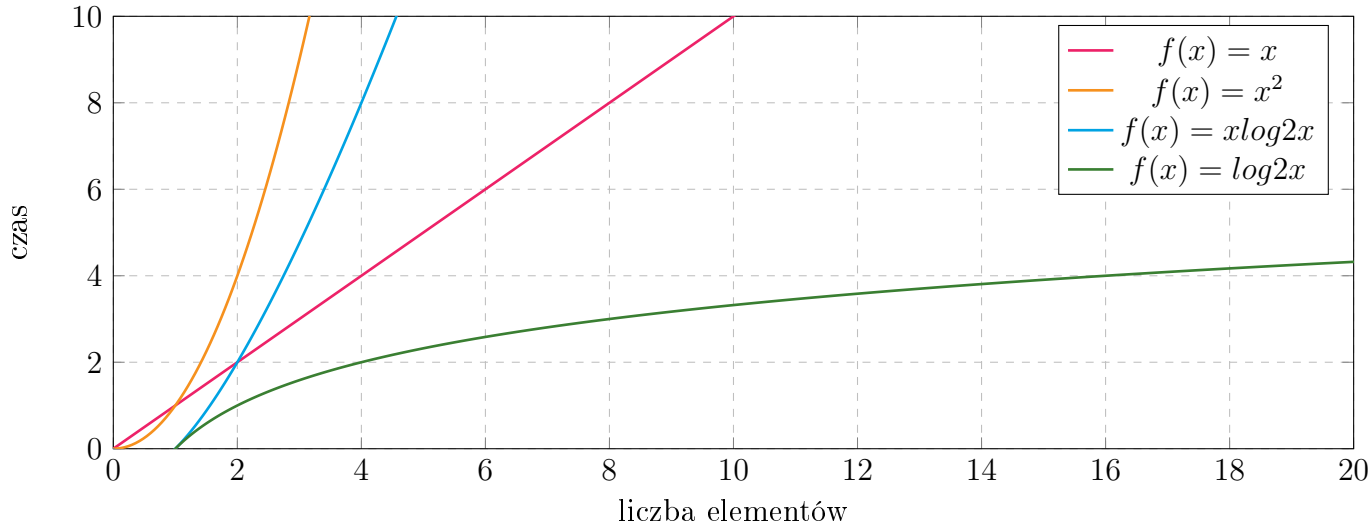
Jest to odmiana sortowania hybrydowego, które opiera się na spostrzeżeniu, że niewydajne jest wywoływanie ogromnej liczby rekurencji dla małych tablic w algorytmie sortowania szybkiego. Głównym założeniem algorytmu sortowania introspektywnego jest zatem wyeliminowanie problemu złożoności $O(n^2)$ występującej w najgorszym przypadku sortowania szybkiego. Sortowanie introspektywne jest połączeniem sortowania szybkiego i sortowania przez kopcowanie, które jest traktowane jako pomocnicze. Tym samym złożoność obliczeniowa wynosi $O(n \cdot logn)$.

2.4. Porównanie złożoności obliczeniowych wybranych algorytmów

Poniższa tabela zestawia oczekiwane i najgorsze przypadki złożoności wybranych algorytmów sortowania. Poniżej dodano także poglądowy wykres funkcji, na którym widać, że dla małej liczby danych sortowanie o złożoności kwadratowej będzie wydajniejsze niż dla logarytmicznej i przeciwnie dla dużej liczby elementów do posortowania.

Tab. 1: Porównanie oczekiwanych i najgorszych przypadków złożoności obliczeniowej dla wybranych algorytmów sortowania

	sortowanie		
	przez scalanie	szybkie	introspektywne
typowa złożoność	$O(nlong)$	$O(nlogn)$	$O(nlogn)$
najgorszy przypadek złożoności	$O(nlogn)$	$O(n^2)$	$O(nlogn)$



Rys. 1: Poglądowe wykresy funkcji możliwych złożoności obliczeniowych

3. Implementacja algorytmów sortowania

3.1. Sortowanie przez scalanie

3.2. Sortowanie szybkie

3.3. Sortowanie introspektywne

4. Zadanie 1 - przeszukanie i przefiltrowanie danych

4.1. Krótki opis

Plik udostępniony do sortowania był okrojoną bazą filmów „IMDb Largest Review Dataset" ze strony kaggle.com. Plik zawierał tytuły filmów oraz przypisane im oceny. Niektóre pola z ocenami były puste, zatem przed wykonaniem zadań związanych z sortowaniem, należało wykonać przeszukanie i usunięcie wpisów bez ocen. Do wykonania tego zadania, zastosowano gotową strukturę z biblioteki STL: *std::vector*. Mimo chęci wykonania sortowań na strukturze dwuelementowej: *std::vector< std::pair <std::string, float> >*, przechowującej i tytuł filmu, i ocenę, komputery, na których wykonywałam testy złożoności obliczeniowej, nie były w stanie wykonać sortowań dla maksymalnej liczby elementów z pliku. Podsumowując, wykonane zostało przeszukiwanie, wykorzystujące strukturę jednoelementową, a następnie sortowane były jedynie oceny filmów. Poniżej przedstawiono algorytm przeszukiwania struktury i usuwania pól z pustymi ocenami:

```
1     for (int i = 0; i< structure.size(); ++i)
2     {
3         if ( structure[i].second.empty() )
4         {
5             structure.erase(structure.begin() + i--);
6         }
7     }
```

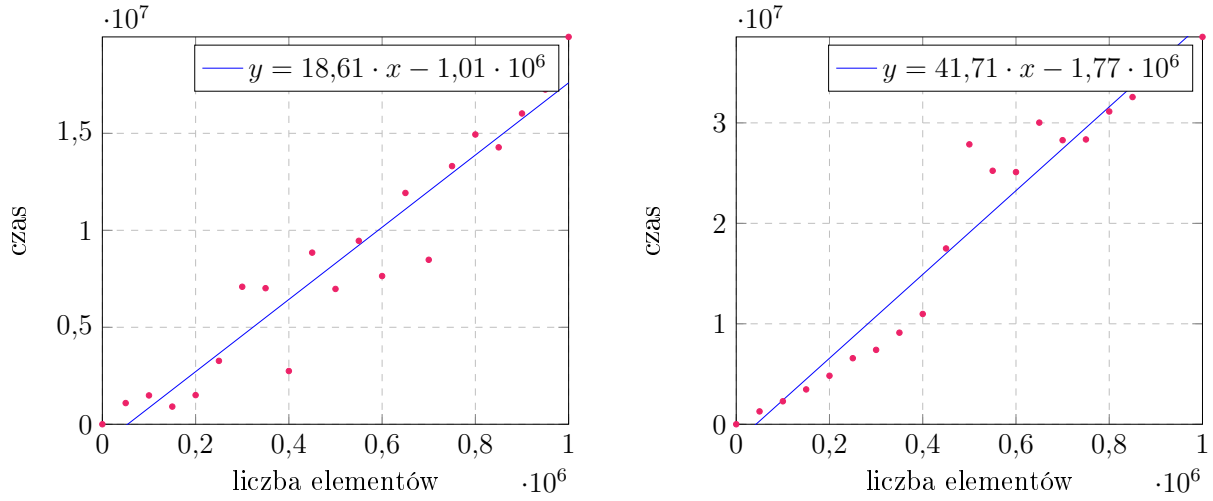
4.2. Analiza złożoności

Kluczową rolę w kodzie odgrywają 2 funkcje: *empty* oraz *erase*. Funkcja *empty* ma złożoność obliczeniową stałą dla jednego elementu, jednak zaimplementowana jak w powyższy sposób, wykona się dla n elementów, zatem jej złożoność w tym przypadku powinna być liniowa $O(n)$. Funkcja *erase* ma oczekiwaną złożoność obliczeniową także liniową $O(n)$. Przeprowadzone zostały testy dla różnych danych w pliku i zmierzone zostały czasy działania obydwu funkcji. Wyniki przedstawia poniższa tabela.

Tab. 2: Czas działania funkcji *empty* i *erase* dla różnej liczby elementów

liczba elementów	czas działania empty [ns]	czas działania erase [ns]
0	381	734
50000	1092756	1280059
100000	1486812	2285586
150000	910546	3468263
200000	1502636	4828635
250000	3267864	6575286
300000	7089495	7404888
350000	7015680	9116657
400000	2745055	10978655
450000	8849012	17499953
500000	6976789	27862306
550000	9457213	25233448
600000	7641937	25103366
650000	11925197	30027679
700000	8479812	28277875
750000	13310083	28351194
800000	14939825	31133886
850000	14277732	32570166
900000	16023509	34344044
950000	17258645	35836178
1000000	19971461	38559105

Na podstawie tabeli 2 wygenerowane zostały charakterystyki działania obydwu funkcji dla różnej liczby danych w pliku. Jak pokazują poniższe wykresy, obydwie funkcje przypominają oczekiwaną charakterystykę liniową. Zatem, funkcje *empty* oraz *erase* w przedstawionej implementacji, mają liniowe złożoności obliczeniowe $O(n)$.



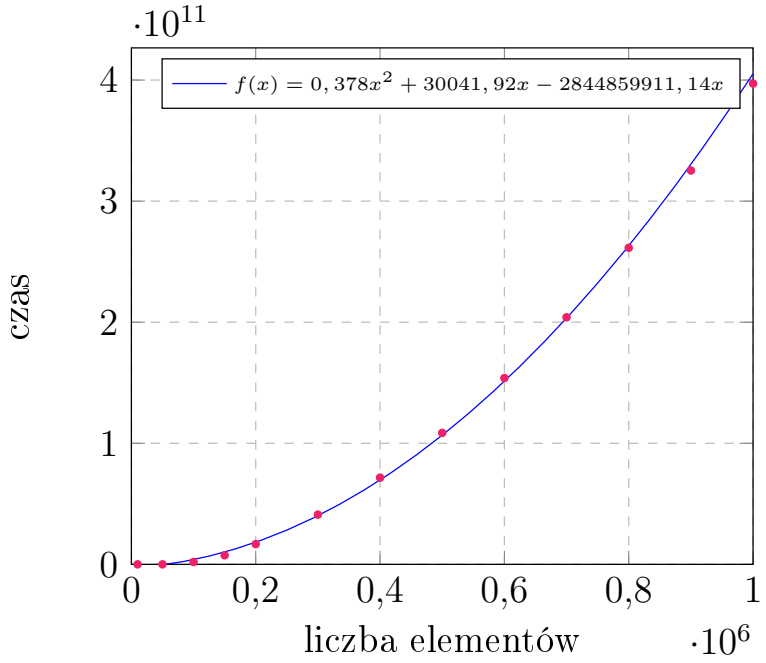
(a) empty (b) erase

Rys. 2: Złożoności obliczeniowe funkcji *empty* i *erase* w przedstawionej implementacji

Złożoności obydwu funkcji są liniowe, zatem całość - przeszukanie i przefiltrowanie danych powinno mieć złożoność kwadratową: $n \cdot n = n^2$. W poniższej tabeli przedstawiono zebrane pomiary działania całego algorytmu.

Tab. 3: Pomiary czasu działania całego algorytmu przeszukiwania i usuwania wybranych pól dla różnej liczby danych

liczba elementów	czas [ns]
10000	99010
50000	2233885
100000	1916298094
150000	7481184919
200000	16730176118
300000	41073112025
400000	71606117902
500000	108592553455
600000	153867195664
700000	204008003164
800000	261449177263
900000	325313418247
1000000	397092989768
1010294	426561982926



Rys. 3: Złożoność obliczeniowa całego algorytmu przeszukiwania i usuwania wybranych elementów

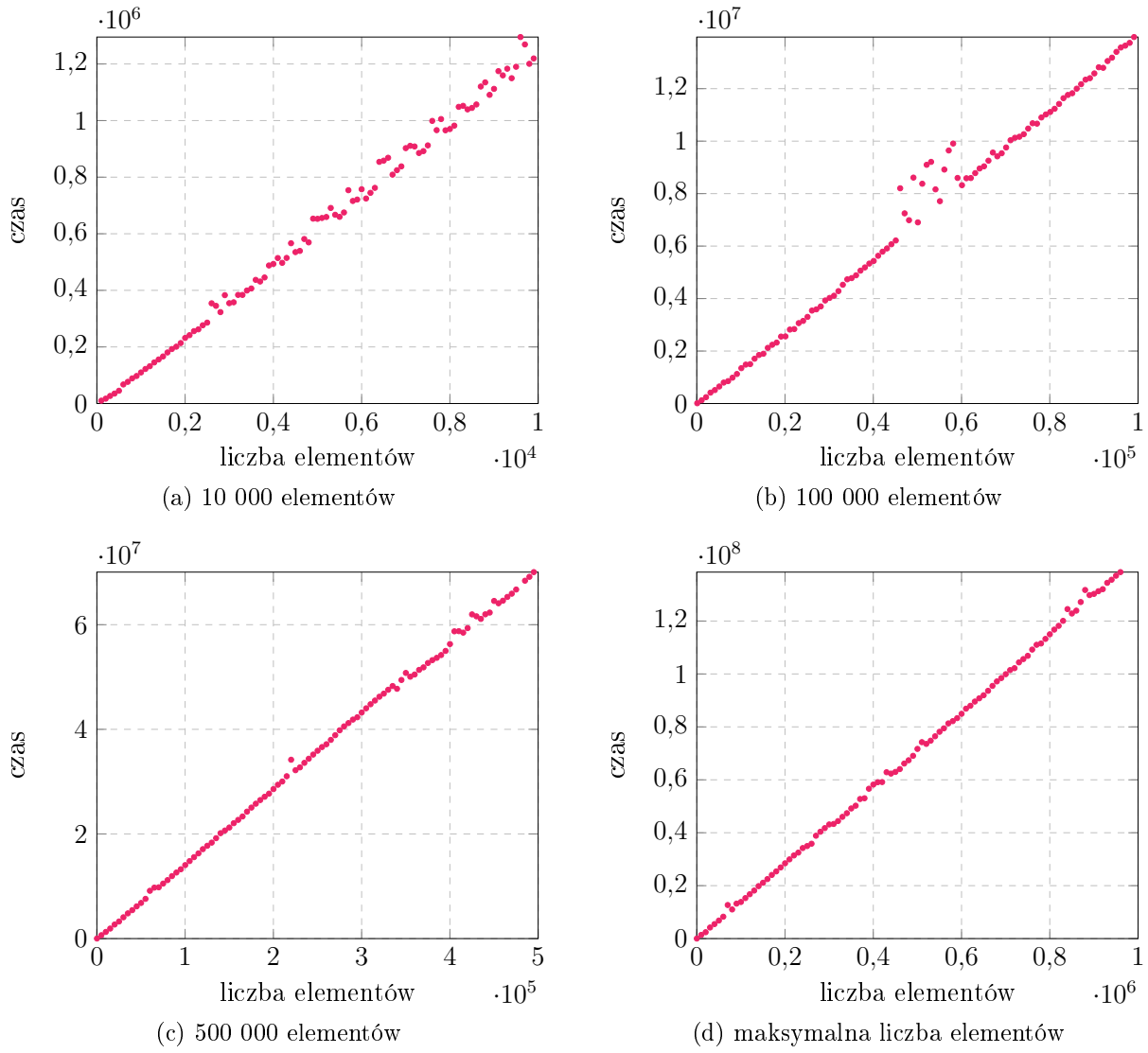
Jak widać na powyższym wykresie, przeszukanie i przefiltrowanie danych ma kwadratową złożoność obliczeniową $O(n^2)$. Dla ponad miliona danych, nie jest to optymalna złożoność. Łączny czas wykonywania przeszukiwania i usuwania wybranych pól zajęła: **426561982926 ns**, czyli około **7,10 min**.

5. Analiza złożoności algorytmów sortowań

5.1. Przebieg eksperymentów

Sortowane były jedynie oceny filmów. Sortowania odbyły się dla 10 000, 100 000, 500 000 oraz maksymalnej ilości danych z pliku po przefiltrowaniu. Dla każdego zestawu danych wykonano po 100 pomiarów.

5.2. Sortowanie przez scalanie



(a) 10 000 elementów

(b) 100 000 elementów

(c) 500 000 elementów

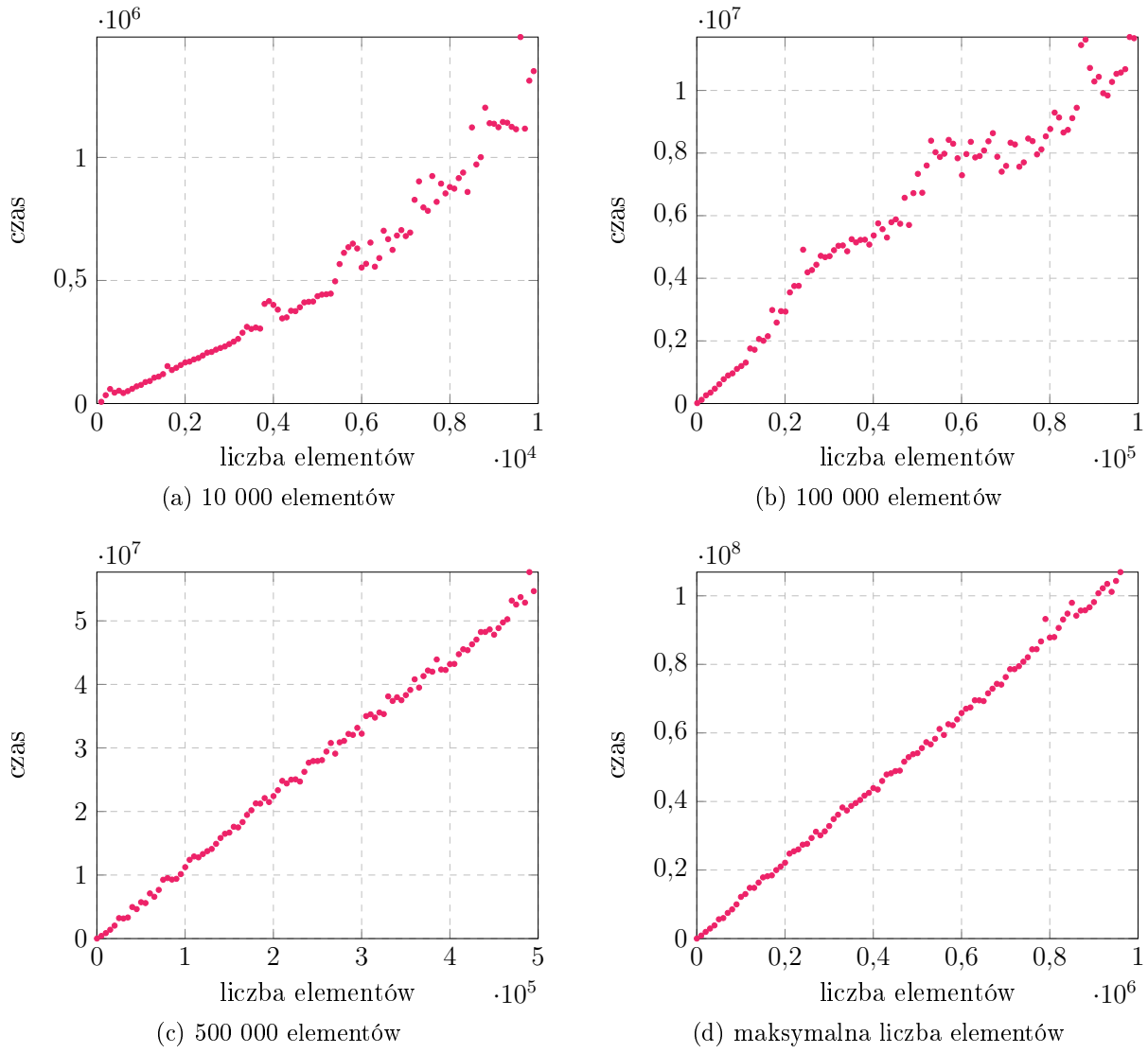
(d) maksymalna liczba elementów

Rys. 4: Sortowanie przez scalanie dla różnej liczby elementów

Tab. 4: Dokładny i przybliżony czas sortowania przez scalania

	liczba elementów			
	10 000	100 000	500 000	maksymalna
dokładny czas sortowań [ns]	1295207	13966928	70052081	138481693
przybliżony czas sortowań [ms]	1,29	13,97	70,05	138,48

5.3. Sortowanie szybkie



(a) 10 000 elementów

(b) 100 000 elementów

(c) 500 000 elementów

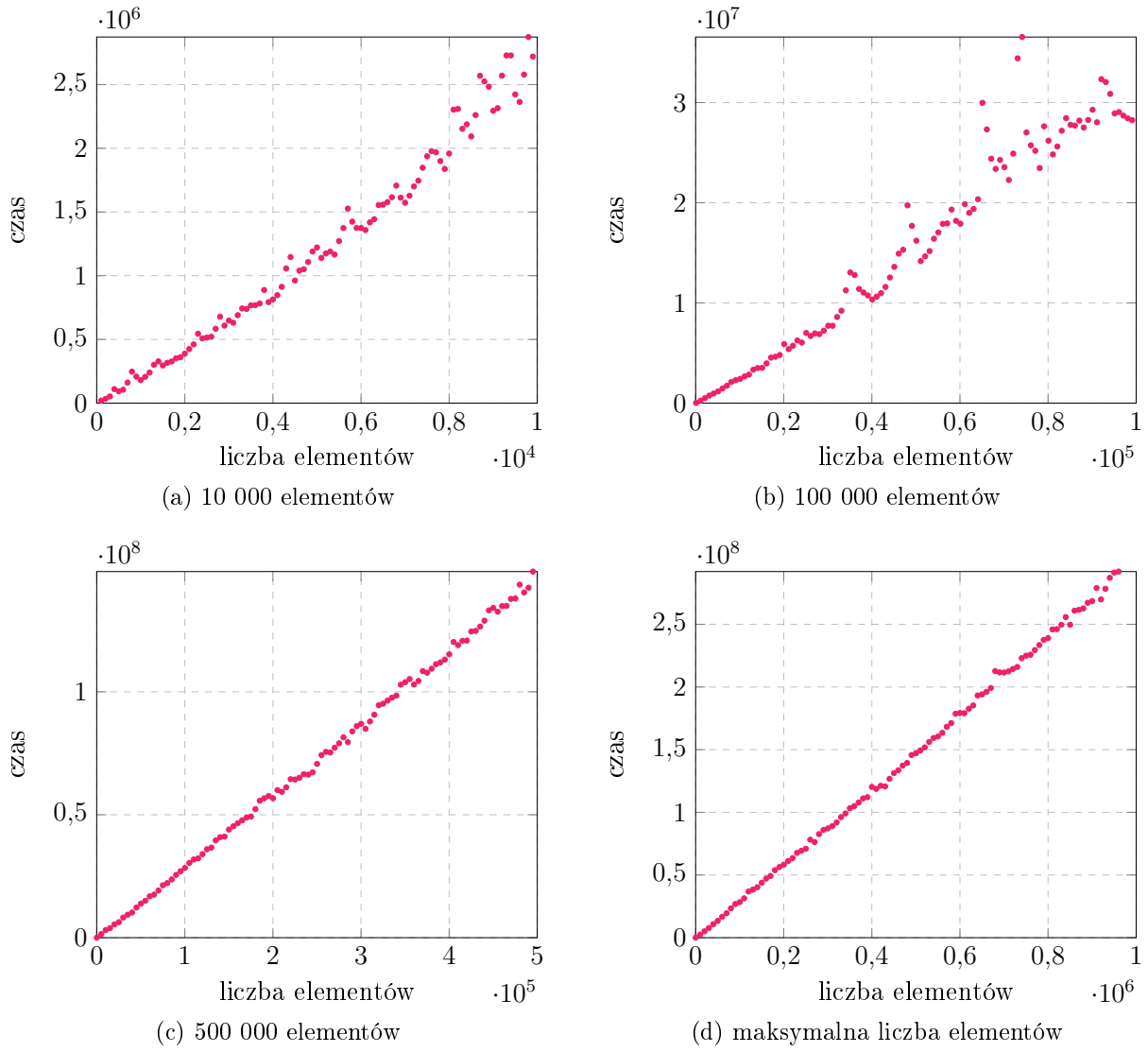
(d) maksymalna liczba elementów

Rys. 5: Sortowanie przez scalanie dla różnej liczby elementów

Tab. 5: Dokładny i przybliżony czas sortowania szybkiego

	liczba elementów			
	10 000	100 000	500 000	maksymalna
dokładny czas sortowań [ns]	1489031	11705054	57671388	106918054
przybliżony czas sortowań [ms]	1,49	11,71	57,67	106,92

5.4. Sortowanie introspektywne



(a) 10 000 elementów

(b) 100 000 elementów

(c) 500 000 elementów

(d) maksymalna liczba elementów

Rys. 6: Sortowanie introspektywne dla różnej liczby elementów

Tab. 6: Dokładny i przybliżony czas sortowania introspektywnego

	liczba elementów			
	10 000	100 000	500 000	maksymalna
dokładny czas sortowań [ns]	287312	36543660	148902464	292090659
przybliżony czas sortowań [ms]	0,29	36,54	148,90	292,09

6. Średnia wartość i mediana

Ponadto, dla każdego zestawu danych zostały wyznaczone średnie wartości oraz mediany rankingu, których wartości zostały przedstawione w poniższej tabeli:

Tab. 7: Średnia wartość oraz mediana wyznaczona dla każdego zestawu danych

	liczba elementów			
	10 000	100 000	500 000	maksymalna
średnia wartość	5,46	6,09	6,67	6,64
mediana	5	7	7	7

7. Podsumowanie i wnioski

pkt 3 - opis + zaznajomienie się na jutro ; wnioski ; ewentualnie komentarze

8. Bibliografia