

# Sprawozdanie

Zbieranie i Analiza Danych

Labolatorium nr 5 – Pliki mapowane w pamięci



Oskar Kwidziński

S156013

Fizyka Techniczna

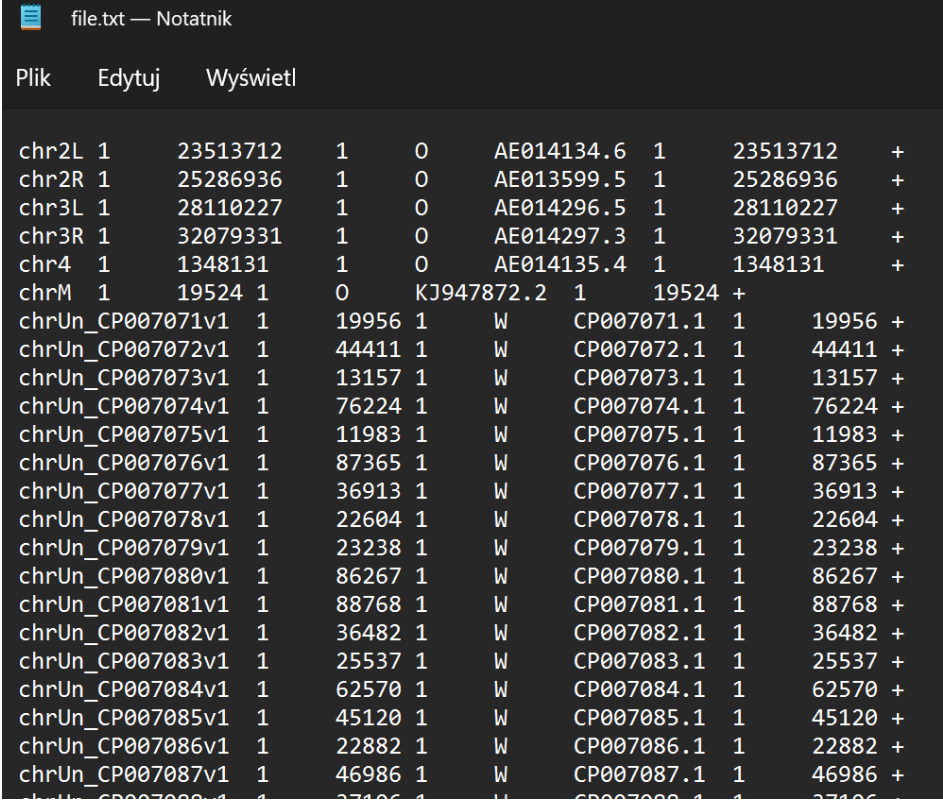
Semestr zimowy 2022/2023

## Wstęp

Wykonano zadanie mające na celu zapoznanie się z możliwością wykorzystywania optymalizacji operacji wyjścia-wejścia. W tym celu stworzono program wielowątkowy pozwalający na przeszukiwanie plików tekstowych w poszukiwaniu pożądanego wzorca. W celu przygotowania pliku tekstowego o rozmiarze 1GB, a także wykorzystano klasę `MemoryMappedFile` – w celu porównania czasu przeszukiwania pliku w porównaniu do tradycyjnego sposobu.

## Zadanie 1

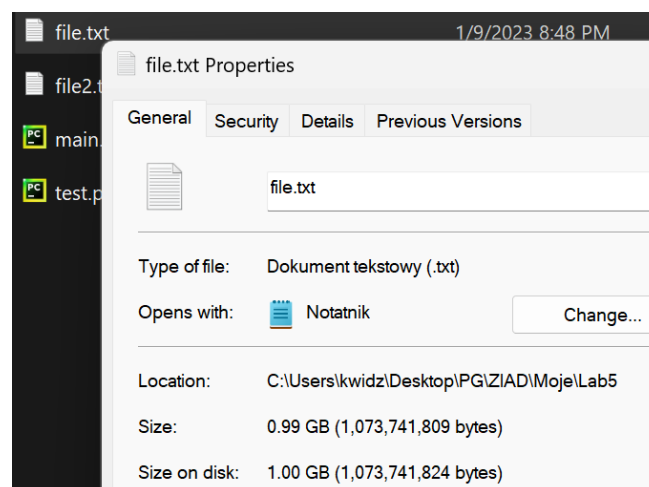
Stworzono duży plik tekstowy zawierający przykładowy zbiór danych:



```
file.txt — Notatnik

Plik    Edytuj    Wyświetl

chr2L 1    23513712    1    0    AE014134.6    1    23513712    +
chr2R 1    25286936    1    0    AE013599.5    1    25286936    +
chr3L 1    28110227    1    0    AE014296.5    1    28110227    +
chr3R 1    32079331    1    0    AE014297.3    1    32079331    +
chr4 1    1348131     1    0    AE014135.4    1    1348131     +
chrM 1    19524      1    0    KJ947872.2    1    19524      +
chrUn_CP007071v1 1    19956      1    W    CP007071.1    1    19956      +
chrUn_CP007072v1 1    44411      1    W    CP007072.1    1    44411      +
chrUn_CP007073v1 1    13157      1    W    CP007073.1    1    13157      +
chrUn_CP007074v1 1    76224      1    W    CP007074.1    1    76224      +
chrUn_CP007075v1 1    11983      1    W    CP007075.1    1    11983      +
chrUn_CP007076v1 1    87365      1    W    CP007076.1    1    87365      +
chrUn_CP007077v1 1    36913      1    W    CP007077.1    1    36913      +
chrUn_CP007078v1 1    22604      1    W    CP007078.1    1    22604      +
chrUn_CP007079v1 1    23238      1    W    CP007079.1    1    23238      +
chrUn_CP007080v1 1    86267      1    W    CP007080.1    1    86267      +
chrUn_CP007081v1 1    88768      1    W    CP007081.1    1    88768      +
chrUn_CP007082v1 1    36482      1    W    CP007082.1    1    36482      +
chrUn_CP007083v1 1    25537      1    W    CP007083.1    1    25537      +
chrUn_CP007084v1 1    62570      1    W    CP007084.1    1    62570      +
chrUn_CP007085v1 1    45120      1    W    CP007085.1    1    45120      +
chrUn_CP007086v1 1    22882      1    W    CP007086.1    1    22882      +
chrUn_CP007087v1 1    46986      1    W    CP007087.1    1    46986      +
chrUn_CP007088v1 1    27105      1    W    CP007088.1    1    27105      +
```



## Zadanie 2, 3

Utworzono skrypt mierzący okres potrzebny do przeszukania pliku w poszukiwaniu pożądanego wzorca (string '13906') w dwóch przypadkach:

- pliku mapowanego,
- w tradycyjny sposób:

```
def regular_io():
    with open(FILENAME, mode="r", encoding="utf8") as file_obj:
        text = file_obj.read()
        text.find("13906")

def mmap_io():
    with open(FILENAME, mode="r", encoding="utf8") as file_obj:
        with mmap.mmap(file_obj.fileno(), length=0, access=mmap.ACCESS_READ) as mmap_obj:
            mmap_obj.find(b"13906")

def main():
    print("OUTPUT FOR REGULAR SEARCH: ", timeit.repeat(
        "regular_io()",
        repeat=3,
        number=1,
        setup="from __main__ import regular_io"))

    print("OUTPUT FOR MEMORY MAPPED FILE: ", timeit.repeat(
        "mmap_io()",
        repeat=3,
        number=1,
        setup="from __main__ import mmap_io"))

if __name__ == "__main__":
    main()
```

Wyniki pomiarów prezentują się następująco:

```
OUTPUT FOR REGULAR SEARCH: [1.5121071999892592, 1.5879973999690264, 1.590463100001216]
OUTPUT FOR MEMORY MAPPED FILE: [0.000769799982663244, 0.00034410000080242753, 0.0002886999864131212]
```

Odnotowano spadek czasu potrzebnego do przeszukania przy pomocy plików mapowanych w pamięci w porównaniu do konwencjonalnej metody. Dla porównania przygotowano również plik o mniejszym rozmiarze (file2.txt – 2 megabajty) i ponownie przeprowadzono test:

```
OUTPUT FOR REGULAR SEARCH: [0.009022800019010901, 0.009191700024530292, 0.008215599984396249]
OUTPUT FOR MEMORY MAPPED FILE: [0.00032489997101947665, 0.00016110000433400273, 0.00012520002201199532]
```

## Podsumowanie

W obydwu przypadkach przeszukanie pliku przy pomocy plików mapowanych w pamięci przebiegło w krótszym czasie. Różnica w przypadku małego pliku nie była aż tak znacząca jak w przypadku pliku dużego.