

Report - letný semester

Ročníkový projekt

Odhad repetitivnosti dát

Krajčiová Eliška

Contents

Abstrakt	1
Úvod	1
Metodológia	2
Analýza dát	2
Pohľad na grafy.....	3
Výsledky.....	8
Diskusia.....	9
Záver	9
Zdroje.....	10
Prílohy.....	10

Abstrakt

Tento projekt sa zameriava na analýzu a porovnanie výkonu dvoch metód spracovania dát, delta a delta-stream, pomocou rôznych súborov typu fasta. Hlavnými cieľmi projektu bolo zhodnotiť efektívnosť a presnosť oboch metód a identifikovať ich výhody a nevýhody. Pomocou rôznych metrík, ako sú hodnoty delta, čas používateľa, systémový čas a využitie CPU, sme analyzovali správanie sa týchto metód na rôznych súboroch dát.

Úvod

Tento projekt sa zaoberá testovaním softvéru pre odhad repetitívnosti dát, konkrétne nástrojmi delta a delta-stream, ktoré sú súčasťou projektu na GitHub: [Substring Complexity](#). Tento softvér počíta/približuje mieru δ , kde δ je počet rôznych faktorov dĺžky ℓ vstupného reťazca. Nástroj delta presne počíta túto mieru a poskytuje aj súvisiace štatistiky, ako $\delta \leq K$, n/δ , a $\text{argmax}_k(d_k/k)$ pre malé hodnoty k . Algoritmus je popísaný v práci:

- Tomasz Kociumaka, Gonzalo Navarro, Nicola Prezza, "Toward a Definitive Compressibility Measure for Repetitive Sequences". IEEE Trans. Inf. Theory 69(4): 2074-2092 (2023).

Nástroj delta-stream počíta aproximáciu δ pomocou sublineárneho pracovného priestoru.

Práca: vo vývoji.

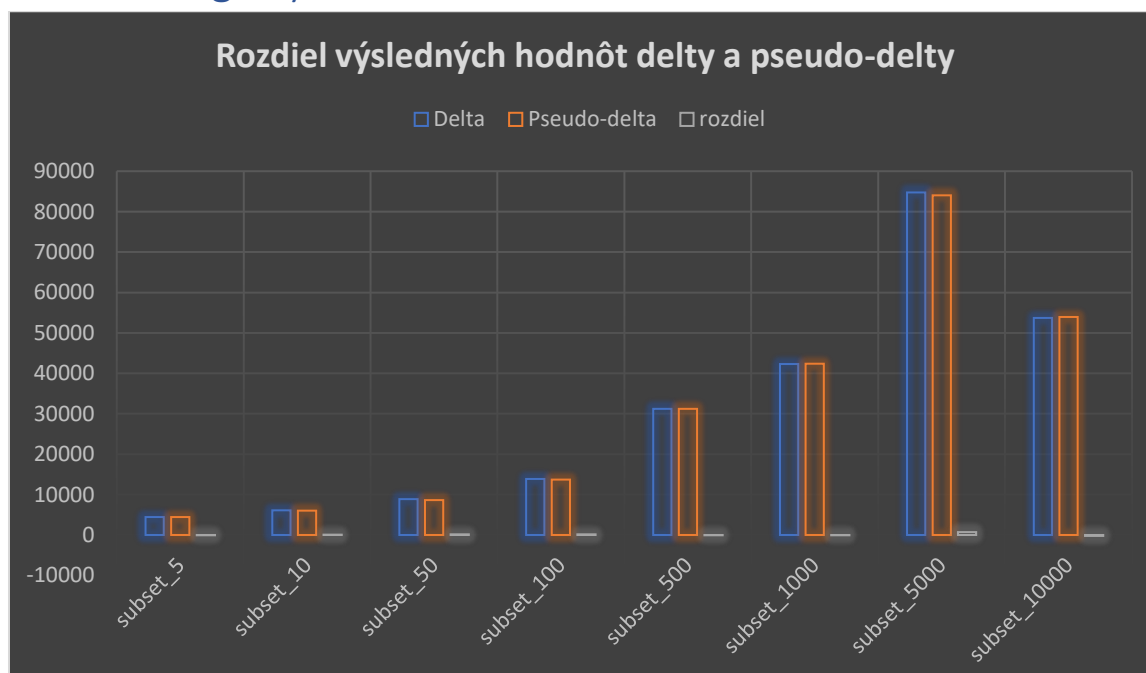
Metodológia

Pre analýzu sme použili osem rôznych súborov dát: subset_5.fasta, subset_10.fasta, subset_50.fasta, subset_100.fasta, subset_500.fasta, subset_1000.fasta, subset_5000.fasta a open_sc2_10k.fasta. Na každý súbor sme aplikovali metódy delta a delta-stream a zaznamenali sme kľúčové metriky ako hodnoty delta, čas používateľa, systémový čas a využitie CPU. Na vizualizáciu dát sme použili Excel, kde sme vytvorili scatter ploty a stĺpcové grafy.

Analýza dát

Analýza sa sústredila na porovnanie hodnôt delta, času používateľa, systémového času a využitia CPU medzi metódami delta a delta-stream. Vizualizovali sme dáta, aby sme jasne videli rozdiely a identifikovali trendy. Scatter ploty poukázali na vzťah medzi využitím CPU a uplynulým časom, zatiaľ čo stĺpcové grafy poskytli jasné porovnanie hodnôt delta a spotreby pamäte.

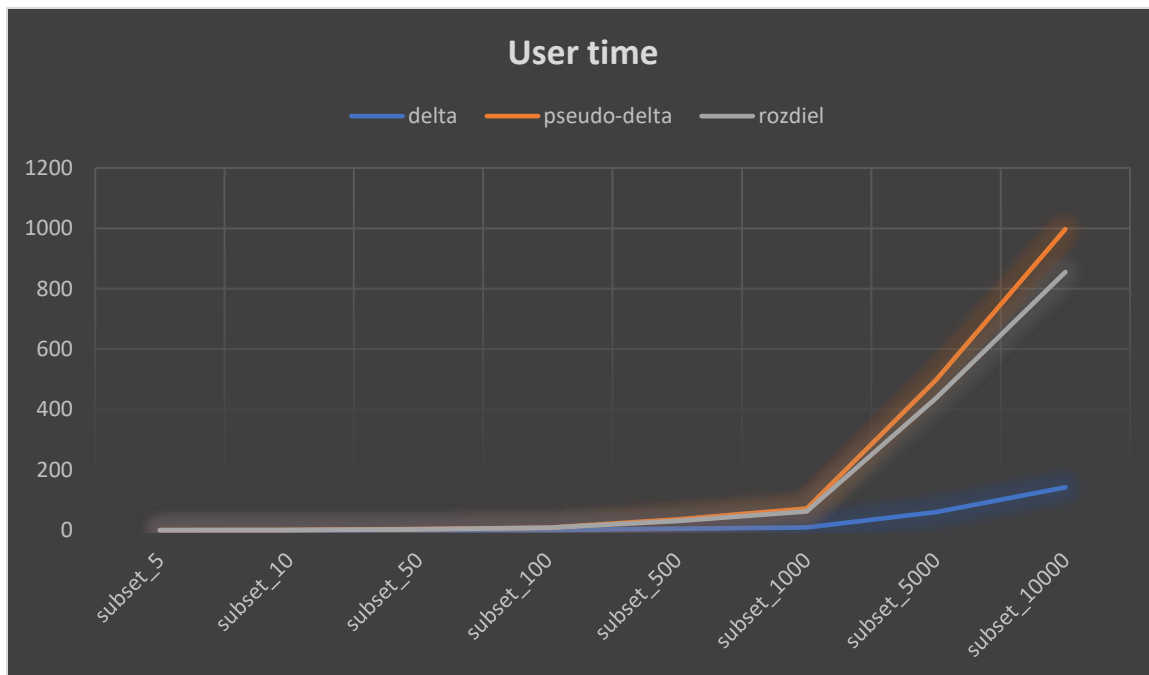
Pohľad na grafy



Tento graf porovnáva hodnoty delty a pseudo-delty pre rôzne súbory. Modré stĺpce reprezentujú hodnoty získané metódou delta, zatiaľ čo oranžové stĺpce predstavujú hodnoty získané metódou delta-stream. Posledné, na prvý pohľad menej viditeľné sivé stĺpce ukazujú rozdiel v hodnotách prvých dvoch.

Analýza:

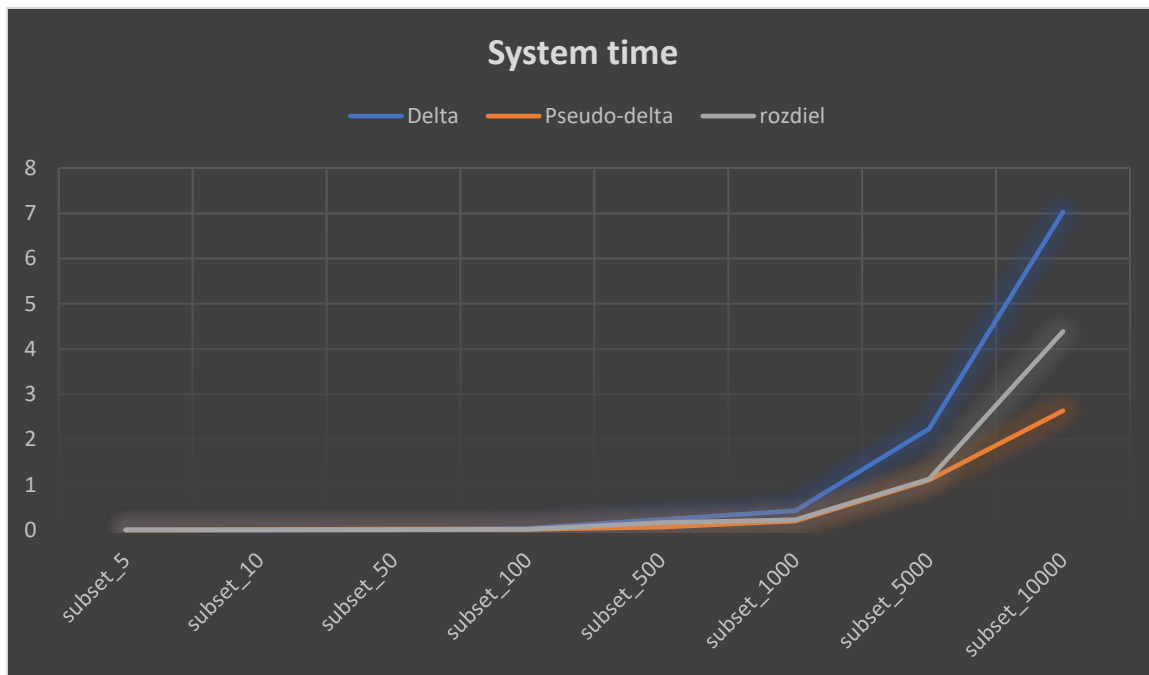
- Pre malé súbory (subset_5 a subset_10) sú hodnoty delty a pseudo-delty veľmi podobné, čo naznačuje, že obe metódy sú pri týchto veľkostiach veľmi presné.
- Pri väčších súboroch (subset_1000 a vyššie) vidíme, že metóda delta-stream produkuje hodnoty pseudo-delty, ktoré sú mierne odlišné od hodnôt delty. To je dôsledok približnej povahy metódy delta-stream, ktorá využíva sublineárny pracovný priestor.
- Rozdiely medzi hodnotami delta a pseudo-delta sa zvyšujú so zvyšujúcou sa veľkosťou súborov, čo poukazuje na výzvy spojené s presnosťou metódy delta-stream pri spracovaní veľkých dát.



Tento graf zobrazuje čas používateľa (user time) potrebný na spracovanie každého súboru metódami delta a delta-stream. Modrá čiara predstavuje čas používateľa pre metódu delta, zatiaľ čo oranžová čiara predstavuje čas používateľa pre metódu delta-stream. Sivá čiara predstavuje rozdiel hodnôt delta a pseudo-delta

Analýza:

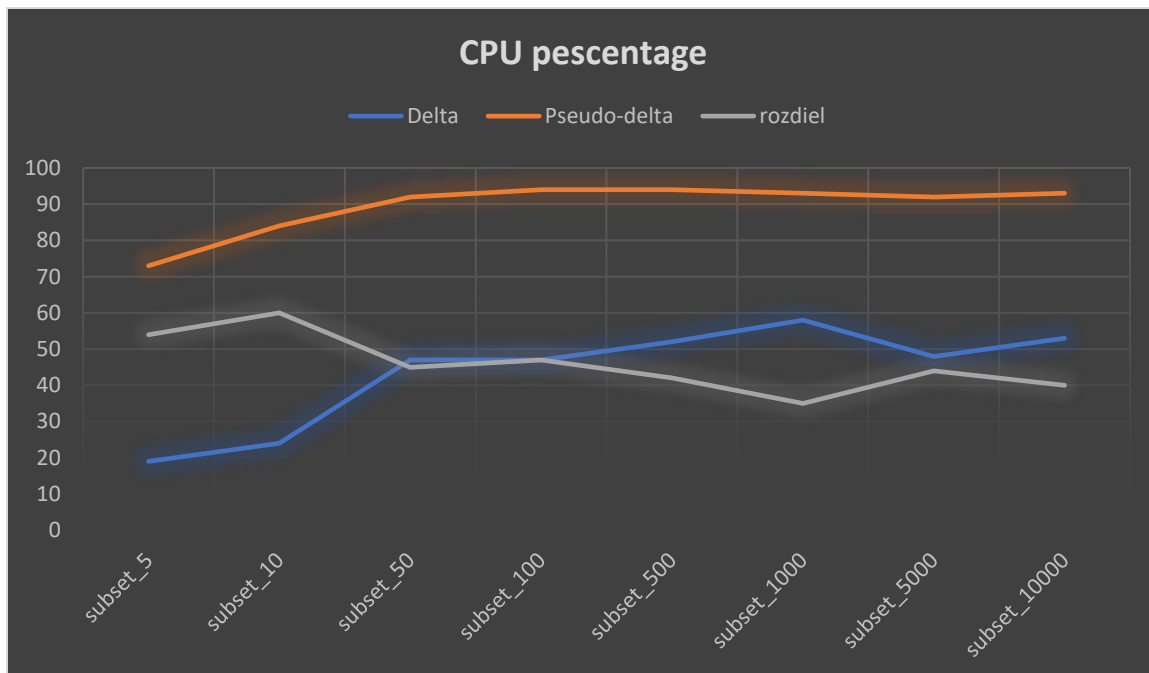
- Čas používateľa pre metódu delta je výrazne nižší v porovnaní s metódou delta-stream, najmä pri väčších súboroch. To naznačuje, že metóda delta je efektívnejšia z hľadiska CPU-bound procesov.
- Pri menších súboroch je rozdiel v čase používateľa medzi oboma metódami minimálny, čo znamená, že obe metódy sú pri malých dátach podobne efektívne.
- Pri najväčšom súbore (open_sc2_10k) je čas používateľa metódy delta-stream výrazne vyšší, čo poukazuje na náročnosť tejto metódy pri spracovaní veľkých dát.



Tento graf zobrazuje systémový čas (system time) potrebný na spracovanie každého súboru metódami delta a delta-stream. Modrá čiara predstavuje systémový čas pre metódu delta, zatiaľ čo oranžová čiara predstavuje systémový čas pre metódu delta-stream. Sivá čiara predstavuje rozdiel hodnôt delta a pseudo-delta.

Analýza:

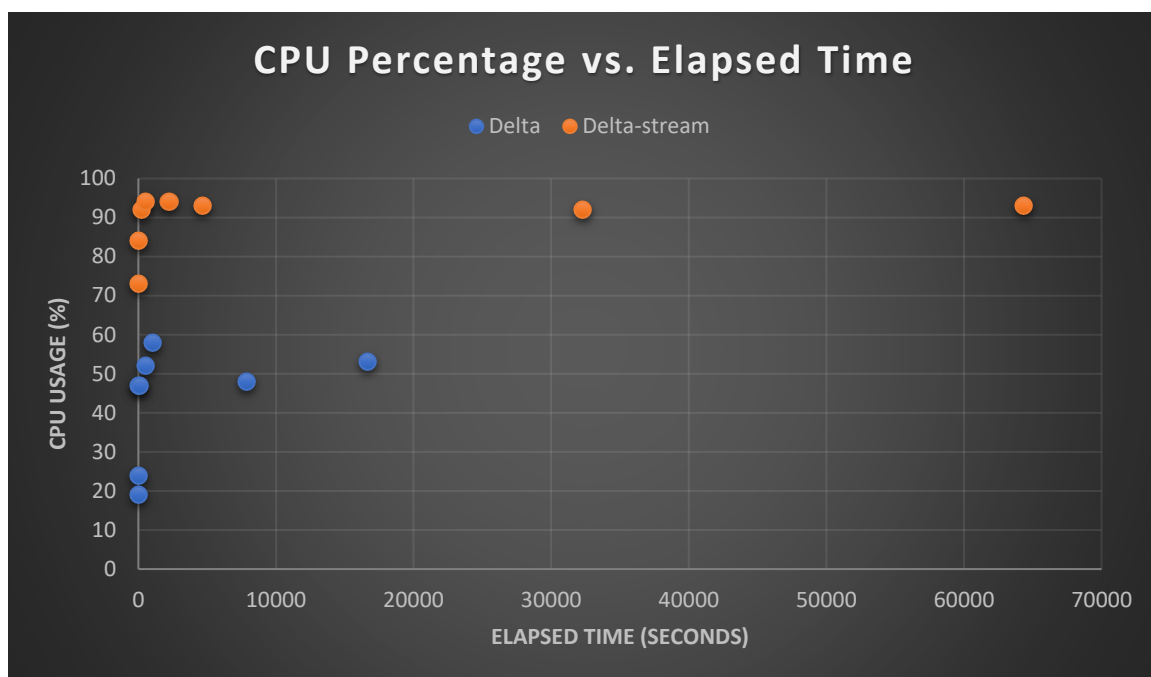
- **Subset 5 až Subset 100:** V týchto menších súboroch je systémový čas pre obe metódy veľmi nízky a takmer rovnaký. To naznačuje, že pri malých množstvách dát sú obe metódy rovnako efektívne z hľadiska systémového času.
- **Subset 500:** V tejto kategórii začíname vidieť mierne rozdiely, pričom metóda delta vykazuje o niečo vyšší systémový čas ako delta-stream. To môže byť dôsledok efektívnejšieho využitia I/O operácií metódou delta-stream.
- **Subset 1000 a vyššie:** Pri veľkých súboroch je systémový čas pre metódu delta výrazne vyšší v porovnaní s delta-stream. To poukazuje na to, že metóda delta-stream je pri veľkých množstvách dát systémovo efektívnejšia, je to vďaka svojmu sublineárnemu prístupu.



Tento graf zobrazuje percento využitia CPU pri spracovaní každého súboru metódami delta a delta-stream. Modrá čiara predstavuje využitie CPU pre metódu delta, oranžová čiara predstavuje využitie CPU pre metódu delta-stream a sivá čiara predstavuje rozdiel medzi nimi.

Analýza:

- **Subset 5:** Metóda delta vykazuje približne 20% využitie CPU, zatiaľ čo delta-stream dosahuje približne 70% využitie CPU. Rozdiel je významný, čo naznačuje, že delta-stream je náročnejšia na CPU zdroje už pri malých dátach.
- **Subset 10 až Subset 50:** V týchto kategóriách vidíme nárast využitia CPU pre metódu delta, zatiaľ čo delta-stream zostáva konzistentne vysoká, okolo 85-90%. To opäť poukazuje na vyššiu náročnosť metódy delta-stream na CPU zdroje.
- **Subset 100:** V tejto kategórii sa CPU využitie pre metódu delta a delta-stream priblížilo, ale delta-stream stále dominuje s približne 90% využitím CPU.
- **Subset 500 až Subset 5000:** Metóda delta ukazuje variabilné využitie CPU, ktoré kolíše medzi 40-60%, zatiaľ čo delta-stream zostáva stabilná nad 90%. Stabilné vysoké využitie CPU pre delta-stream naznačuje jej intenzívne požiadavky na CPU počas spracovania veľkých súborov.
- **Subset 10000:** Aj pri najväčšom súbore je využitie CPU pre metódu delta výrazne nižšie (približne 50%) v porovnaní s delta-stream, ktorá stále využíva približne 90% CPU.



Tento scatter plot zobrazuje vzťah medzi percentom využitia CPU a uplynulým časom (v sekundách) pre metódy delta a delta-stream. Modré body reprezentujú metódu delta, zatiaľ čo oranžové body predstavujú metódu delta-stream.

Analýza:

- **Využitie CPU:** Metóda delta-stream (oranžové body) vykazuje konzistentne vysoké percento využitia CPU (okolo 90%) naprieč všetkými uplynulými časmi. To naznačuje, že delta-stream je veľmi náročná na CPU zdroje počas celého procesu.
- **Uplynulý čas (Elapsed Time):** Metóda delta (modré body) ukazuje rôzne percentá využitia CPU, ktoré sú vo všeobecnosti nižšie ako pri delta-stream, a to aj pri dlhších uplynulých časoch.
 - Pri nižších uplynulých časoch (pod 10 000 sekúnd) vidíme väčšinu bodov pre delta a delta-stream.
 - Pri vyšších uplynulých časoch (nad 10 000 sekúnd) sú viditeľné iba body pre delta-stream, čo naznačuje, že táto metóda trvá dlhšie na spracovanie väčších súborov.

Výsledky

Počas testovania a analýzy metód delta a delta-stream na rôznych súboroch typu fasta sme dosiahli niekoľko dôležitých zistení. V nasledujúcich sekciách sú zhrnuté výsledky pre jednotlivé grafy, spolu s celkovými závermi.

Rozdiel výsledných hodnôt delty a pseudo-delty

- **Zistenia:** Pri malých súboroch sú hodnoty delta a pseudo-delta veľmi podobné, čo naznačuje vysokú presnosť oboch metód. Pri väčších súboroch však metóda delta-stream vykazuje väčšie odchýlky od hodnôt delta, čo môže byť dôsledok jej približnej povahy.
- **Záver:** Metóda delta-stream poskytuje približné hodnoty, ktoré sú pri malých dátach veľmi presné, ale pri väčších dátach sa môžu objaviť odchýlky.

User time

- **Zistenia:** Čas používateľa (user time) pre metódu delta je výrazne nižší v porovnaní s metódou delta-stream, najmä pri väčších súboroch. Metóda delta je efektívnejšia z hľadiska CPU-bound procesov.
- **Záver:** Pre úlohy vyžadujúce rýchle spracovanie s nízkou náročnosťou na CPU zdroje je metóda delta vhodnejšia.

System time

- **Zistenia:** Systémový čas (system time) pre metódu delta je pri malých a stredných súboroch takmer rovnaký ako pre delta-stream. Pri väčších súboroch delta-stream vykazuje výrazne nižší systémový čas.
- **Záver:** Metóda delta-stream je efektívnejšia pri spracovaní veľkých dát, čo sa prejavuje nižším systémovým časom.

CPU percentage

- **Zistenia:** Metóda delta-stream vykazuje konzistentne vysoké využitie CPU (okolo 90%) naprieč všetkými súbormi, čo znamená vysokú náročnosť na CPU zdroje. Metóda delta má variabilné a nižšie využitie CPU.

- **Záver:** Metóda delta je vhodnejšia pre úlohy s obmedzenými CPU zdrojmi, zatiaľ čo delta-stream je náročnejšia, ale môže poskytnúť vyššiu presnosť.

CPU Percentage vs. Elapsed Time

- **Zistenia:** Metóda delta-stream má vysoké využitie CPU a dlhší uplynulý čas v porovnaní s metódou delta. Delta vykazuje nižšie využitie CPU a kratší uplynulý čas.
- **Záver:** Pre úlohy, kde je čas kritickým faktorom, je metóda delta vhodnejšia. Delta-stream, aj keď náročná na zdroje, môže byť vhodná pre úlohy vyžadujúce vyššiu presnosť napriek vyšším nákladom na čas.

Diskusia

Na základe analýzy a testovania môžeme vyvodiť niekoľko kľúčových záverov:

- Metóda delta je efektívnejšia z hľadiska využitia CPU a systémového času pri menších súboroch, čo ju robí vhodnou pre menej náročné úlohy a obmedzené zdroje.
- Metóda delta-stream, aj keď náročnejšia na CPU zdroje a uplynulý čas, poskytuje presnejšie výsledky pri veľkých dátach, čo je výhodné pre náročnejšie úlohy.
- Výber medzi metódami delta a delta-stream by mal byť založený na konkrétnych požiadavkách projektu, pričom je potrebné zvážiť kompromisy medzi presnosťou a efektívnosťou.

Záver

Na záver, tento projekt poskytol detailné porovnanie metód delta a delta-stream. Každá metóda má svoje výhody a nevýhody, a výber medzi nimi by mal byť založený na konkrétnych požiadavkách úlohy.

Zdroje

- Tomasz Kociumaka, Gonzalo Navarro, Nicola Prezza, "Toward a Definitive Compressibility Measure for Repetitive Sequences". IEEE Trans. Inf. Theory 69(4): 2074-2092 (2023).
- [Substring Complexity on GitHub](#)

Prílohy

Poskytnem prístup k súboru, na ktorom mám uložené výstupy z konzoly a pomoc pri vytváraní subsetov.