

Report – zimný semester

Ročníkový projekt

Odhad repetitivnosti dát

Krajčiová Eliška

Contents

Úvod	1
Teoretický základ	1
Metodológia	2
Proces	2
Testy	3
Záver	3

Úvod

V súčasnej dobe rýchleho rastu objemu dát je kľúčová ich efektívna analýza. Môj projekt sa zameriava na meranie repetitívnosti dát pomocou delta miery, čo je dôležité pre rôzne aplikácie, od bioinformatiky po spracovanie jazyka. Pôvodným cieľom bolo vyvinúť metódu na odhad pseudodelty, alternatívy k delta miere, ktorá je výhodná v prípadoch, keď je dôležitejšia rýchlosť a efektívnosť než presnosť. Na scénu však prišiel tím ľudí, ktorí vyvinuli softvér, ktorý dokáže efektívne vypočítať pseudodeltu, to zmenilo smer môjho projektu na štúdium a testovanie tohto nástroja. Tento report dokumentuje tento proces, od teórie po praktické testovanie softvéru.

Teoretický základ

Môj prvý styk s delta miernou bol prostredníctvom článku "Indexing Highly Repetitive String Collections, Part I: Repetitiveness Measures", odporúčaného mojím garantom Adriánom Gogom. Tento článok, najmä jeho sekcia 3.10, mi poskytol základné pochopenie merania repetitívnosti. Aby som ešte hlbšie pochopila túto komplexnú tému, strávila som mnohé hodiny študovaním doplnkových online zdrojov. Medzi tieto zdroje patrí projekt Delta Measure ([GitLab projekt](#)) a texty z Pizzachili ([Pizzachili kolekcia](#)). Následne, po objavení sa

spomínaného tímu výskumníkov a ich prínosu v podobe nového softvéru na výpočet pseudodelty, som bola postavená pred ďalšiu výzvu - dôkladné naštudovanie ich práce ([arXiv](#)). Tento článok sa stal ďalším kľúčovým zdrojom informácií, ktorý som musela analyzovať a pochopiť, aby som mohla efektívne implementovať a testovať ich softvér v rámci môjho projektu.

Metodológia

V rámci projektu som použila softvér na výpočet a aproximáciu miery δ , ktorá sa zameriava na počet unikátnych faktorov určitej dĺžky k vstupného reťazca. Tento nástroj, dostupný na [GitHub](#), je navrhnutý pre efektívne spracovanie veľkých textových súborov a poskytuje rýchle a presné analýzy repetitívnosti.

Od garanta som dostala dataset s názvom chr19. Tento dataset obsahuje Chromozómové haplotypy 19 odvodené z projektu 1000 Genomes (1KG) Project, fázy 3, ktoré predstavujú významnú časť genetického výskumu zameraného na rozmanitosť ľudského genómu. Projekt 1000 Genomes, ktorého cieľom bolo sekvenovať genómy tisícok ľudí z rôznych etnických skupín, bol jedným z najambicióznejších projektov v oblasti genómovej diverzity. Chromozóm 19 je zvlášť zaujímavý kvôli svojej veľkosti a hustote génov. Obsahuje 58 miliónov bázových párov, čo je približne 1,9% z celkového ľudského genómu. Haplotypy sú súbory génov alebo DNA variácií, ktoré sú dedené spoločne. V prípade chromozómu 19 boli haplotypy odvodené z dát 1KG projektu fázy 3, ktoré poskytli hlbšie porozumenie genetických variácií v rámci tohto chromozómu.

Proces

Pred objavením spomínaného softvéru som experimentovala s Jellyfish, nástrojom na analýzu DNA sekvencií, ktorý bol súčasťou môjho pôvodného prístupu k problematike repetitívnosti dát. Tento skúsenostný základ mi pomohol lepšie pochopiť komplexnosť a výzvy súvisiace s analýzou veľkých dátových súborov. Následne som sa pokúsila spojiť nový softvér na operačnom systéme Windows. Pre jeho správne fungovanie bolo potrebné nainštalovať cmake a knižnicu SDSL, čo prinieslo množstvo komplikácií. Po čtených výzvach, neúspešných pokusoch a hľadaní riešení problémov na internete som dospela k záveru, že najlepším riešením bude opustiť tento prístup, vzdať sa hodín doterajšej práce a začať celý proces odznova s využitím WSL (Windows Subsystem for Linux).

Inštalácia a spojenie cmake a SDSL bolo pri práci s WSL (používam Ubuntu) oveľa rýchlejšie a priamočiarejšie v porovnaní s Windows. Pri sťahovaní datasetu nastala prvá menšia komplikácia – jeho veľkosť. Komprimovaný priečinok má veľkosť zhruba 1,52 GB, jeho sťahovanie trvalo môjmu notebooku približne 13 minút. Pri pokuse o extrahovanie došlo

k chybe a mojím jediným šťastím bolo, že projekt mám na externom disku. Dala som teda dataset extrahovať na stolný počítač a stalo sa to jeho jedinou úlohou. Tento proces zabral zhruba 2 hodiny a veľkosť súboru je 58,7 GB. Narazila som na ďalší problém – extrahovanie prefixov trvá dlhšie než som dúfala. To v spojení s ďalším problémom – počítanie delty a pseudodelty tiež trvá dlho – vyúsťuje v to, že do deadline-u tohto semestra zvládnem otestovať len veľmi malú vzorku. Na rýchlosť celého procesu vplýva aj málo voľnej pamäte na mojom notebooku.

Testy

Prvý test môjho projektu sa uskutočnil na súbore chr19.5.fa, ktorý obsahuje 5 sekvencií vybratých z pôvodného datasetu chr19.1000.fa. Hlavným cieľom bolo vypočítať pseudodeltu, a preto som spustila príslušný program. Počas procesu sa objavilo niekoľko komplikácií s kompiláciou, ale postupne som ich všetky vyriešila.

Keď bol program konečne pripravený a testovací súbor nastavený, spustila som test pomocou príkazu „./delta-stream -s -o "output_path" < file“. Očakávala som, že proces bude rýchly, ale na moje prekvapenie nebolo výpočet dokončený ani po 5 minútach. Nakoniec skončil neočakávane chybou – runtime_error, čo je zobrazené na obrázku z výsledku testu nižšie. V dôsledku tejto chyby sa moje testovanie skončilo skôr, než som predpokladala.

```
krajciov@DESKTOP-E1RM13A:/mnt/d/rocnikovy_projekt/substring-complexity$ ./delta-stream -s -o /mnt/d/rocnikovy_projekt/substring-complexity/outputs < /mnt/d/rocnikovy_projekt/chr19.5.fa
Selected prime number, q = 30754405573884523
Number of sampled lengths: 53
Stream length = 300557691
delta = 6.11643e+06
argmax_k = 18
terminate called after throwing an instance of 'std::runtime_error'
what(): Failed to dump
Aborted
```

Záver

V priebehu môjho projektu som zažila výrazný rozdiel medzi prácou na operačnom systéme Windows a využívaním Windows Subsystem for Linux (WSL). Bolo prekvapujúce, koľko problémov sa vynorilo pri spúšťaní softvéru – od nutnosti inštalovať rôzne doplnky po zvládanie chýb, opráv a úprav. V dôsledku týchto skúseností by som odporučila lepšiu dokumentáciu a presnejšie inštrukcie pre kompiláciu programov delta.cpp a delta-stream.cpp.

Rýchle ukončenie plánovaných testov na väčšej vzorke súborov a pomalosť extrahovania dát z pôvodného datasetu ma zaskočili. Je zrejmé, že tieto problémy súvisia s technickými obmedzeniami.

Počas letného semestra sa plánujem venovať týmto výzvam, s cieľom zlepšiť výkon mojej techniky alebo získať prístup k výkonnejšiemu zariadeniu pre hlbšie testovanie. Momentálne však nemám k dispozícii riešenie pre uvedené problémy, a preto sa moja práca za zimný semester končí.