Ročníkový projekt – report – letný semester

Minimalizácia počtu runov v BWT

Na začiatku semestra som napísal ešte niekoľko väčších a detailnejších testov na overenie správnosti programu. Po doladení menších chýb sme prešli k spracovaniu reálnych dát. Pôvodný cieľ bol spracovať dataset - 10k covid genomes (about 300MB). Dáta avšak obsahovali až okolo 100 000 unikátnych tagov. Veľkosť abecedy labels týchto dát teda bola príliš veľká vzhľadom na našu interpretáciu. Spracovali sme teda len dáta bez labels.

Namerané dáta:

Percentage of data	Computation time
1%	45 min
10%	24h 24min / 1464 min

Časy som sa ďalej počas semestra snažil zlepšiť. Nedosiahol som však dostatočného zlepšenia a projekt zostáva týmto smerom otvorený. Veľká časť neefiktivnosti je aj v používaných knižniciach. Napríklad, len vytvorenie sufixového stromu pre celé dáta trvá viac ako 2h.