FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student:	Bc. Jakub Novák
Oponent práce:	Ing. Jan Trávníček, Ph.D.
Název práce:	Komprese souborů FASTQ
Obor:	Teoretická informatika

Datum vytvoření: 25. 8. 2020

Hodnotící kritérium:	*Způsob hodnocení – následující škálou 1 až 4:*
1. Splnění zadání	*1=zadání splněno,* ***2=zadání splněno s menšími výhradami,*** *3=zadání splněno s většími výhradami,* *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Komentář:
Cílem zadání bylo nastudovat formát FASTQ pro ukládání dat vzniklých při sekvenování DNA, implementovat algoritmus komprese tohoto formátu a porovnat vlastnosti implementovaného algoritmu s dalšími existujícími algoritmy. Zadání hodnotím jako splněné s menšími výhradami, protože porovnání s výsledky z dvou zadáním odkazovaných článků bylo provedeno v práci jen okrajově a celá kapitola Výsledky se na toto téma zaměřuje jen jednou stranou.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
2. Písemná část práce	*70 (C)*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3. Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Komentář:
Kapitola základních pojmů je na diplomovou práci velmi stručná. Postrádám formální definice datových struktur, se kterými se v práci pracuje. Za zmínku stojí Suffix Trie, která je v implementaci používaná, ale chybí i definice řetězce a entropie a další, které jsou neméně důležité.

Stejně tak by kapitola základních pojmů měla popisovat kompresní algoritmy, které jsou ve výsledku použity, a to ještě detailněji než jak jsou v textu popsány v pozdějších kapitolách, kde jsou, pocitově, jakoby jen mimochodem.

Text práce by uvítal jazykovou korekturu viz "Tato u?vaha se uka?zala by?t spra?vnou podle vy?sledc??ch vy?s?e zm??ne?ne? studie, ..." a další příklady.

Na straně 11, 19 a 30 přetéká text pravou hranu zarovnání.

Sekce 5.5 Klient je spíše uživatelskou příručkou, která by měla být v příloze.

V kapitole Měření je zavedena slabší a silnější verze algoritmu, které se liší nastavením a počtem modelů predikce. Hodnoty nastavení vycházejí z předpokladu struktury DNA, nebylo však provedeno ověření zda tyto hodnoty poskytují optimální výsledky.

Nevidím důvod neuvádět referenční hodnoty komprese testovacích souborů pomocí GZip do tabulky společně s výsledky získanými pomocí implementovaných algoritmů. Podobně také nevidím důvod oddělení výsledků komprese souboru SRR007215_1 od hlavních (Z. marina, exander2, fey2), i když jsou tyto výsledky převzaty.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
3. Nepísemná část, přílohy	*62 (D)*

Popis kritéria:
Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce
– opakovatelnost experimentů

Komentář:

Nepísemná část práce je ve formě zdrojového kódu/java balíčku zpracovávajícího FASTQ soubory a zdrojového kódu/java balíčku aritmetického kodéru upraveného tak, aby pracoval s v práci popsaným modelem predikce bitů. Aritmetický kodér byl v rozšiřované knihovně kompresních algoritmů již implementovaný, a byl tedy zřejmě studentem jen v kopii upraven pro kompresi a dekompresi podle již zmíněných modelů. Celkově je rozsahem implementace na diplomovou práci spíše kratší a jednodušší.

Jediný způsob testování studentova příspěvku do knihovny, který jsem nalezl ve zdrojových kódech knihovny je zprostředkován třídou FASTQTest. Vzhledem k použité standardní struktuře implementace jako maven balíčku je tento test nekorektně mezi zdrojovými soubory nikoli mezi testy. Implementace není jinak otestována ani unit testy a, vzhledem k charakteru nástroje do kterého bylo přispíváno, ani integračními testy.

Javadoc se v kódu vyskytuje u přinejlepším 50% veřejných metod. Jediné rozhraní ve studentově kódu není dokumentované vůbec.

Ve třídě FASTQClient bych návratové hodnoty deklaroval jako Enum a v metodě FASTQClient::main bych spíše využil příkazu switch než dvou vnořených příkazů if.

Třída FASTQProvider využívá přinejmenším zvláštně principu zapouzdření. Proč jsou některé instanční proměnné třídy FASTQProvider veřejné? Navíc, proměnné, které jsou privátní, mají jak gettery tak settery poskytující přímý přístup k nim a nejsou tedy technicky vzato sémanticky rozdílné od veřejných. Ve třídě ContextMixingParams je dodaný setter k jedné z veřejných instančních proměnných.

Metoda FASTQData::StringList2ByteArray může být statická.

Jaký je význam optimalizace "přesunutí připravení modelu" (metoda setupModel) ve třídě ContextMixing z konstruktoru, kde by dle mého názoru dával větší smysl, do metod compress a decompress? I kdyby měl zůstat kde je, není nutné duplikovat informaci o neexistujícím modelu ve formě instanční proměnné first, když v takovém případě bude instanční proměnná models nastavena na null.

Instanční proměnné encoder, decoder a models jsou ve třídě ContextMixingCoder deklarované jako package private, i když se této vlastnosti nikde v kódu nevyužívá.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Hodnocení výsledků, jejich využitelnost	*75 (C)*

Popis kritéria:
Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Komentář:

Implementaci jsem otestoval i na souboru který nebyl formátu FASTQ a, i když komprese i dekomprese proběhla bez nahlášené chyby, soubor na výstupu nebyl stejný jako ten na vstupu. Čekal bych že program nahlásí chybu.

Na druhou stranu, implementace je pro FASTQ soubory funkční a dokáže je zkomprimovat vyšším kompresním poměrem než porovnávané nástroje.

Myslím, že implementace je spíše prototypem, který by především z uživatelského hlediska potřeboval ještě vylepšit.

Hodnotící kritérium:	*Způsob hodnocení – nehodnotí se*
5. Otázky k obhajobě	

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:
Proměřil jste vašimi algoritmy i datový soubor SRR007215_1 ze zadáním odkazovaného článku [3]?
Prosím okomentujte porovnání vašeho algoritmu s výsledky ze zadáním odkazovaného článku [4], které jsem v práci nenašel.

[3] Bonfield JK, Mahoney MV (2013) Compression of FASTQ and SAM Format Sequencing Data. PLoS ONE 8(3): e59190. https://doi.org/10.1371/journal.pone.0059190
[4] El Allali, A., Arshad, M. MZPAQ: a FASTQ data compression tool. Source Code Biol Med 14, 3 (2019). https://doi.org/10.1186/s13029-019-0073-5

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Celkové hodnocení	*68 (D)*

Popis kritéria:
Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.

Text hodnocení:
Pro odůvodnění bodového hodnocení zopakuji nejdůležitější výtky k odevzdané závěrečné práci:
- V textu jsem bohužel nenašel formalismy a konstrukce, které bych očekával v závěrečné práci oboru Teoretické informatika.
- Obsah některých sekcí textu je ve vyloženě nevhodných kapitolách.
- V práci je jen na jedné straně diskutováno porovnání výsledků s výsledky v zadáním odkazovaných článcích, a přitom bych toto čekal jako jedno z hlavních obsahových témat.
- Implementace je dle mého názoru rozsahem spíše jednodušší, automaticky netestovaná a na diplomovou práci dle mého názoru kratší a jednodušší.

Především z těchto důvodů hodnotím práci 68 body tedy stupněm D (uspokojivě).

Podpis oponenta práce: