Posudek recenzenta diplomové práce

Téma diplomové práce: **Datamining výsledků experimentů**

Diplomant: Bc. Martin Bárta

Diplomant se v posuzované diplomové práci zabývá způsoby, jakými lze analyzovat výsledky experimentů, které generují velmi obsáhlé datové sady. Na to navazuje popis technik pro hledání užitečných informací s jejich následným zobrazením pomocí tabulek a grafů.

V teoretické části je vysvětlen proces a technologie dolovaní dat, společně s představením nástrojů použitelných pro tento úkol. Praktická část pak obsahuje postup analyzování dat s pomocí nástroje Apache Spark. Diplomant použil vzorovou datovou sadu, získanou při běhu algoritmu symbolické regrese, rovněž za pomoci nástroje Apache Spark. Rozdělení práce do kapitol je zcela logické a správné, jejich objem pokládám za přiměřený.

Předložená práce je rozdělena do 5 základních kapitol. Z toho v kapitole 2 jsou představeny použité technologie, zejména programovací jazyky. Přestože popis je věcný a správný, je zaměřen spíše prakticky; v porovnání mi chybí větší teoretický nadhled (např. jazyky staticky x dynamicky typované, funkcionální paradigmata apod.). Samostatná kapitola 3 je věnována nástroji SPARK. Nástroj v ní je podrobně představen popsán. To je správně, protože se v dalších kapitolách 4 a 5 používá pro praktické dolování dat. Spark byl vybrán pro zpracování praktické části především z hlediska rychlosti. V práci je vysvětleno, proč je Spark rychlejší i v případě dat použitých v diplomové práci. V tomto konkrétním případě se sice nevejdou do operační paměti počítače, na kterém byla zpracovávána, ale přesto je rychlost vyšší ve srovnání s jiným nástrojem Hadoop. Diplomant podrobně zdůvodnil, proč tomu tak je.

Kapitoly 4 a 5 tvoří těžiště vlastní práce diplomanta. Jsou to kapitoly „konstrukční", protož popisují a zdůvodňují rozhodování diplomanta při „konstrukci" experimentu, dále pak provedení experimentu a jeho výsledky. Všechny kroky v této části jsou logicky zdůvodněny, a proto je akceptuji jako správné. Poněkud nešťastné se mi jeví, že v práci jsou podrobné *screen-shoty* stránek pro stažení programů od jejich dodavatelů. Je to jednak zbytečné (stahování a instalace programů je pro IT technika běžná rutina), jednak tím objem diplomové práce zbytečně narůstá, ale především je jejich vypovídací hodnota malá, protože vzhled webových stránek se mění velmi rychle.

Pro praktický příklad byla zvolena data, která jsou výstupem algoritmu symbolické regrese. Jedná se o genetický algoritmus, vysvětlený na obrázku 14 a podrobně popsaný v následujícím textu. Po věcné stránce se jedná o zdařilou kapitolu, námitky ovšem mám k její formě. Nezdá se mi vhodné, v diplomové práci používat fráze jako *„výsledný JAR soubor poté bude naservírován spark-submit skriptu..."* a podobně. Osobně pokládám za nepatřičný i výklad formou *„naším cílem je do seznamu přidat proměnnou"*, *„nebudeme všechny TGZ soubory extrahovat na disk"* a podobně, protože DP má být výpovědí o vlastních výsledcích diplomanta a nikoliv nějakého abstraktního „my".

Navrhuji, aby se diplomant při obhajobě pokusil vysvětlit, proč existuje sedm experimentů, které nemají žádné výsledky (strana 78).

Celkově diplomovou práci hodnotím jako kvalitní. Diplomant prokázal, že se umí orientovat ve složitém prostředí práce s velkými daty, a to nejen teoreticky, ale především prakticky. Velmi oceňuji, že se mu podařilo vyřešit všechny problémy, ke kterým v praktickém použití došlo. Zejména ta část, ve které došlo k selhání při pokusu načítat a zpracovávat data z objemné databázové tabulky, svědčí o důkladném zvládnutí technologie diplomantem. Mimoto, vyřešení problému načítáním dat po částech pomocí integrovaných funkcí Apache Spark je dosti obtížné a je jasným důkazem, že diplomant úlohu zvládl na úrovni, která je vyšší, než je běžně od studentů požadováno.

Drobné námitky mám k jazykové stránce, protože na mnoha místech chybí interpunkce.

Protože diplomant prokázal svou schopnost samostatně, správně a s patřičným přehledem vyřešit zadané úlohy, **navrhuji přijmout jeho diplomovou práci k obhajobě a hodnotím ji známkou**

=== výborně ===.

V Praze dne 5. září 2020

doc. Ing. Josef Kokeš, CSc.

recenzent