Posudek diplomové práce pana Bc. Tomáše Prudkého nazvané “Data Science a vizualizace dat”

Oponent doc. Dr. Ing. Tomáš Brandejský V Pardubicích 6.9.2023

1. Jaké metody (příslušející navazujícímu magisterskému studiu) diplomant ve své práci uplatnil?

Student ve své práci vycházel především ze znalostí získaných v předmech věnovaných oblasti databází. Nad jejich rámec využil znalostí z oblasti BigData a datové analytiky.

2. Co diplomant při vypracování své DP vytvořil?

Dle zadání měl student popsat možnosti a způsoby využití dat science, což se mu podařilo. Dále měl podle tohoto zadání požít jazyk Python, Jupyter lab, nebo Jupyter notebook či Apache Zeppelin ve spojení s Apache Spark a EKL Stack. Použít metodiku …
V hodnocení využití jazyka Python a výše zmíněných nástrojů vydímnejvětší problém. Práce obsahuje v příloze jen výsledné CSV soubory, žádný programový kód. Také v textu práce není tento zdrojový kód ani obrázek dokládající např. Použití Jupyter nebo Zeppelin notebooků.

3. Jak diplomant prokázal správnost navrženého řešení problému?

Diplomant přiložil výše uvedené CSV soubory.

4. Podařilo se diplomantovi splnit cíle práce, které mu byly uloženy?

Diplomantpostupoval metodicky správně přiformulai množiny nástrojů pro řešení problému. Použití některých nástrojů a jazyka python neprokákaz, nepřiložil je k práci ani jako Pyhonovský kód, ani jako notebook.

5. Jaká je kvalita textu diplomové práce z hlediska jeho struktury, srozumitelnosti, jazykové a typografické úrovně?

Práce z formálního hlediska odpovídá doporučené šabloně a čítá 119 stran včetně všech požadovaných seznamů a dále práci tvoří samostatná příloha. K práci přiložený ZIP soubor obsahuje demonstrační *.CSV soubory. V práci je 61 obrázků.
Práce je logicky členěna, je srozumitelná a je v prvních kapitolácj na odpovídající jazykové a typografické úrovni. Kap. 3.3 dojem kazí – je plna překlepů a neuvážených hodnocení.
Kap. 4.2.3 přináší užitečný příklad získávání dat pomocí nástrojů Elastic Stack, především Elastic Search.
Při vlastní analýze postupoval diplomant metodicky správně. Definoval si otázky, které byly předmětem jeho zájmu. Získal potřebná data.

6. Jak byla vyhodnocena kontrola textu DP (případně zdrojových kódů softwaru) pomocí systému pro odhalování plagiátů mezi závěrečnými pracemi?

Systém detekce plagiátů uvádí minimální shodu ve výši 1%.

7. Které nejasnosti vyskytující se v DP by měl diplomant objasnit při obhajobě a jaké jsou další připomínky k DP?

V práci jsem nicméně nalezl následující formální nedostatky:
Nepovažuji za vhodné používání anglických termínů i v případech, kdy existují rozšířené české ekvivalenty – narážím především na názvy kapitoly 1 a podkapitol 2.1 až 2.4. Obzvláště nešťastné

je střídavé užívání anglického i českého termínu, viz kap. 1.1, její titulek na bezprosředně navazující věta.
Citovaný (převzatý) text bývá zvykem odlišit kurzívou, což v práci není dodrženo, viz kap. 1, 2. odstavec.
Str. 19, dvě věty tvořící poslední odstavec si protiřečí.
Str. 20, 2. odstavec – obávám se, že Hadoop není nerelační databáze...
Str. 21, nejsou “tyto data” ale “tato data”
Str. 21, předposlední odstavec – zábavný průmysl není totéž jako zábavní průmysl. Proč je v tomto odstavci copyright Microsoftu (a také o stranu dále)? Znak copyrightu se v práci vyskytuje i dále – podle mne do ní nepatří vzhledem k odkazům na původní text. Při jeho důsledném užívání bz musel být uveden i u názvy firmy Oracle na str. 37, IBM na str. 42 a pod.
Název kap. 3 je poněkud nesrozumitelný, s DataScience se nepracuje.
Kap. 3, 1. odstavec – především různé problémy vyžadují různé postupy.
Kap. 3.3 – obávám se, že skutečným důvodem rozšíření Pythonu je jeho snadnější zvladatelnost neprogramátory. Pro uvedené velké obejmy dat a efektivní výpočty je i Java nebo Julia vhodnějším jazykem – vždyť Python nepodporuje ani paralelní vykonávání kódu.
Kap. 3.3.4 – proč tedy popisujete Matplotlib, když (nejen) podle Vás je např. Ggplot modernější a existuje i pro Python?
Kap. 3.3.8 – autor nepochopil funkci knihoven Keras a TenzorFlow (nikoli TenzerFlow) – TenzorFlow je vlastní výkonná knihovna, Keras její nadstavba.
Kap. 3.5 – autor pominul, že v jazyce Scala je vytvořen samotný Spark a některé jeho funkcionality jsou dostupné jen z programů vytvořených v tomto jazyce. Spar nabízí interface pro čtyři jazyky (Python Scala, Java a R*) a množiny jejich funkcionalit jsou disjunktní…
Popis knihovny Epic není použitelný.
Kap. 3.5.3 – opět nepřesná formulace – Scala je zpravidla rychlejší než Python, protože je kompilována do Java Byte Code. Knihovna ScalaPy tedy nezvyšuje výkon, ale řekněme rozšiřuje aplikační doménu.
Kap. 3.7 – proč je diskutován jazyk Julia, když mu chybí to nejpodstatnější – interface pro Spark? Není v této souvislosti důležitější Java, která Sparkem odporována je?
Mezi kap. 4.1 a 4.2 chybí analýza struktury systému, ze které by vyplynula struktura a funkcionalita systému a tedy role jednotlivých nástrojů. Například není uveden žádný důvod pro použití kontejnerů.
Str. 71, 1. odstavec je poněkud nedopovězen. Končí větou “Při”, dokonce bez tečky na konci. Další odstavec rekapituluje kapitolu v2novanou jazyku R*.
Str. 79 – nesrozumitelná věta “Data byla sesbírána z maximálního možného roku, který byl k dispozici.”
Str. 83, VO2 - Důvodem pro volbu R* byla funkce summary, která dle popisu nabízí stejnou funkcionalitu jako metoda describe v Pandas?
V části věnované nikoli analýze historických dat, ale té druhé, věnované predikci mi chybí jakákoli zmínka o modelu použitém pro predikci. Byla to např. ARMA?
Dvě ze tří položek doporučené literatury jsem nenalezl v seznamu použité literatury (van Der Plas je v textu citován, v seznamu literatury chybí, Hill není ani v textu citován, ani uveden v seznamu literatury).

8. V závěru je nutno jednoznačně uvést, zda je práce doporučena či nedoporučena k obhajobě a jakým klasifikačním stupněm je hodnocena?

Vzhledem k výše uvedenému doporučuji práci k obhajobě a hodnotím ji známkou E – dostatečně.

Doc. Dr. Ing. Tomáš Brandejský