Posudek oponenta bakalářské práce

Student: Dvořáček Libor
Téma: Využití získávání znalostí pro data z PDF souborů (id 23895)
Oponent: Burgetová Ivana, Ing., Ph.D., UIFS FIT VUT

1. Náročnost zadání **průměrně obtížné zadání**

2. Splnění požadavků zadání **zadání splněno**

Zadání práce bylo splněno. Výsledkem práce je aplikace, která umožňuje extrakci tabulek z pdf dokumentů a aplikaci metod pro redukci dimenzionality na takto získaná data. Aplikace však umí zpracovat pouze určitý typ pdf dokumentů a ze zadání není jasné, zda to takto bylo myšleno.

3. Rozsah technické zprávy **je v obvyklém rozmezí**

4. Prezentační úroveň předložené práce **70 b. (C)**

Prezentační úroveň technické zprávy se v jednotlivých kapitolách značně liší. Kapitola 2 je poměrně nepřehledná a těžko pochopitelná. Kapitoly 4 a 6 mohly být podrobnější. Naopak kapitoly 3 a 5 jsou poměrně kvalitní. V práci také postrádám kapitolu, která by se věnoval návrhu aplikace a popisu požadavků na aplikaci. Problematické jsou také popisky obrázků, které jsou až příliš stručné. Až na uvedené nedostatky je logická struktura práce dobrá a jednotlivé kapitoly na sebe dobře navazují.

5. Formální úprava technické zprávy **66 b. (D)**

Z jazykového hlediska se jedná spíše o podprůměrnou práci, která obsahuje řadu gramatických chyb a překlepů (kromě kapitoly 3). Dále pak student používá celou řadu hovorových výrazů (updatována, footerem, callbacků, slideru, defaultně). Z typografického hlediska bych vytkla především odsazení prvního řádku po nečíslovaných podnadpisech.

6. Práce s literaturou **84 b. (B)**

Výběr studijních pramenů, které pokrývají řešenou problematiku z oblasti získávání znalostí je dobrý. Naopak v oblasti nástrojů pro extrakci dat z pdf dokumentů bych očekávala větší množství využitých zdrojů.

7. Realizační výstup **80 b. (B)**

Realizačním výstupem této práce je webová aplikace, která umoňuje zpracování dat pomocí metod pro redukci dimenzionality (PCA, t-SNE, UMAP) a metod shlukové analýzy (hierarchické shlukování a k-means). Implementace těchto metod byly převzaty z dostupných knihoven. Vstupní data mohou být extrahována z pdf dokumentů, které ovšem musí být specifického formátu, nebo z csv souborů.
Bohužel u všech zdrojových kódů chybí jméno autora a datum vytvoření.

8. Využitelnost výsledků

Myslím si, že se jedná o práci, která je dobře použitelná pro zpracování dat z dokumentů daného typu (výsledky laboratorních měření). Bohužel není možné zpracovat tabulková data z libovolných pdf dokumentů.

9. Otázky k obhajobě

- Proč jste pro extrakci dat z pdf dokumentů použil právě knihovnu pdfPlumber? Jaké jsou její výhody oproti jiným nástrojům?

10. Souhrnné hodnocení **80 b. velmi dobře (B)**

V rámci diplomové práce student vytvořil funkční aplikaci, která umožňuje aplikaci různých metod pro redukci dimenzionality a pro shlukovou analýzu na data extrahovaná z pdf dokumentů určitého formátu. Vzhledem k využití většího počtu metod pro redukci dimenzionality navrhuji mírně nadprůměrné hodnocení - 80 bodů (B).

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 3. června 2021

Burgetová Ivana, Ing., Ph.D.
oponent