Posudek oponenta diplomové práce

Autor práce: **Zuzana Mikolášová**

Název práce: **Analýza dat v sociálních sítích**

Obsah práce
Autorka se zabývá aplikací algoritmu PageRank na graf vytvořený z textových příspěvků na sociální síti Twitter za účelem zjišťování tzv. „informativnosti" zpráv vztahujících se k nějaké události. Navazuje tak na svůj oborový projekt s podobnou tématikou, do kterého zapracovává znalosti získané studiem současné literatury.

Kvalita řešení a dosažených výsledků
Kvalitu řešení a dosažených výsledků považuji za průměrnou. Velmi dobrou formální úroveň práce poněkud snižuje malý význam učiněných závěrů a poměrně skromný rozsah vytvořeného zdrojového kódu.

Formální úroveň
Formální úroveň diplomové práce (DP) je velmi dobrá. Chyb a překlepů ve čtivém textu je velmi málo a vhodné je i jeho logické členění. Mimořádně zdařilý je překlad či převyprávění publikace [1] na stranách 16 – 25. Naopak kapitola 8 (Experimenty) by měla mít přesnější název (experimenty s čím?) a nepůsobit jako trochu narychlo vložený dodatek. Na obr. 6.2 by bylo určitě vhodné doplnit kardinality vazeb a označit ho za ERA model. Kapitoly 6.3 a 6.4 by spíše patřily do uživatelského manuálu v příloze. Třídy v diagramech balíků (od s. 62) jsou nesprávně pojmenovány použitím sloves. Lze vytknout užití pojmu „power metoda" (např. na s. 20) místo obvyklého „metoda prostých iterací" a také vazby „pomocí" místo sedmého pádu, např. „pomocí regulárního výrazu" místo „regulárním výrazem" na s. 36.

Práce s literaturou
Práce s literaturou je poměrně slabá. Z 24 referencí jsou pouze 3 „newebové" a z textu práce je zřejmé, že se autorka obsáhle seznámila jen s publikací [1]. Při citaci z knih 7 a 10 by bylo vhodné uvádět i relevantní stránky.

Splnění zadání
Zadání DP bylo splněno, i když budoucí použití vytvořeného zdrojového kódu je sporné z důvodu jeho nedostačujícího okomentování.

Doplňující informace k práci
Největším nedostatkem DP je podle mého názoru nedostatečně ohraničený vlastní přínos od výstupů z předchozího oborového projektu a od převzatých výsledků. (Především na již výše zmiňovaných stranách 16 – 25 mi chyběly explicitní odkazy na práci [1] a musel jsem si je domýšlet.) Také použití externích knihoven a nástrojů není v DP zcela jasně popsáno. Vhodný by byl nějaký schematický diagram jejich začlenění do procesu řešení úkolů zadání DP. Nemohu se ztotožnit se závěrem, že „mezi množinami je silný a kladný vztah" (s. 49). Mezi proměnnými je zřejmě kladná korelace, ale jestli je i silná, to nelze z tak malého vzorku dat usuzovat. Je možné, že by se na větších datech tato vlastnost neprokázala, a proto je předčasné projektovat učiněný závěr i na obecný vztah mezi algoritmem *TwitterEventInfoRank* a logistickou regresí. Z drobných věcných chyb lze zmínit např. „rozdíl vektorů" (místo „rozdíl velikosti vektorů") na s. 39 a „dvě stejné hodnoty" (místo „dvě nebo více stejných hodnot") na s. 48.

Dotazy k práci
Kdo anotoval testovací data? (s. 41)
Kde se vzaly váhy vstupující do klasifikátoru a proč má počet slov negativní váhu? (s. 33)
Proč se formalita textu počítá právě podle uvedeného vzorce? (s. 34)

SOUHLASÍ S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

Z čeho usuzujete, že se „obecně na posledních příčkách objevují příspěvky spojené s konspiračními teoriemi“? (s. 44)

Navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni dne 31. května 2016

doc. Ing. Dalibor Fiala, Ph.D.

SOUHLASÍ S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky