FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Ing. Miloslav Konopík, Ph.D.
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd
Technická 8, CZ-30614 Plzeň
Tel.: (+420) 377 63 2418

Posudek oponenta diplomové práce

Ladislav Hlom: Automatická klasifikace vícejazyčných dokumentů

Diplomantovým úkolem bylo provést poměrně náročný vědecký experiment s klasifikací cizojazyčných dokumentů do tříd daných českou kolekcí. Student se ve své práci nejprve věnuje teoretické rovině strojového překladu, metodě LDA (nad rámec zadání) a klasifikaci dokumentů. Pochopení teorie strojového překladu a metody LDA považuji za slabou stránku práce. V textu týkajícím se této problematiky se nachází řada nepřesností (viz otázky v závěru práce).

Popis řešení je srozumitelný, i když některé pasáže nejsou příliš přesné. Opět se zejména jedná o použití metody LDA. Na str. 33 se píše, že musíme mít k dispozici stejně anotovaná data a že se klasifikuje na základě kosinové podobnosti. Zcela chybí definice toho, co to jsou stejně anotovaná data a jak se kosinová podobnost použije pro klasifikaci. Ze zdrojových kódů jsem zjistil, že je použit tzv. paralelní LDA model, který používá paralelní data (dvoujazyčná), která však nejsou nijak anotována. Navíc, tento paralelní model není vůbec v teorii popsán. I model klasifikace jsem si musel odvodit ze zdrojových kódů. Klasifikace probíhá tak, že se spočte kosinová podobnost daného dokumentu se všemi dokumenty v kolekci a vektor těchto podobností se vloží na vstup klasifikátoru. Toto ovšem není standardní řešení a není zřejmé, jak na něj student přišel. Myslím si, že diplomant v této oblasti podcenil studium literatury. Nejčastěji se používá buďto Supervised LDA [1] a nebo se na vstup klasifikátoru dává přímo vektor distribuce pravděpodobností témat.

Také považuji za celkem nešťastné, že nebyla použita optimalizace parametrů překladových submodelů (v práci nepřesně ladění systému). Překlad s implicitními hodnotami vah může dosahovat nevyzpytatelných výsledků. Diplomant si mohl situaci zjednodušit tím, že by použil hotové modely, které lze stáhnout ze stránek Mosesu[1]. Pokud chtěl diplomant záměrně zhoršit výsledky překladu, mohl snadno nastavit parametry překladu, které překlad zrychlí a zhorší. Výsledek by byl lépe predikovatelný.

Konfigurace experimentů a jejich výsledky jsou prezentovány přehledně a srozumitelně. Práce obsahuje velice stručnou kapitolu zhodnocení výsledků, v té diplomant výsledky jen popisuje, avšak vůbec se nezabývá jejich analýzou.

Formální úroveň práce je na dobré úrovni. Byl jsem schopen nalézt jen zanedbatelné prohřešky (např. rastrový obrázek v anglickém jazyce na str. 2, přetečení řádky na str. 8 nebo formulace ... *podařilo najít nalézt lepší...* na str. 36).

Student pracuje s literaturou vesměs dobře, avšak někdy uvádí on-line zdroje i v případech, kdy lze nalézt adekvátní tištěnou publikaci (např. [4] Moses – existují doporučené citace, [20] Brainy – existuje doporučená citace na stránkách systému).

Zdrojové kódy práce jsou přehledné a dobře komentované. Výjimku tvoří soubory *ParallelLDA.java*, *ParallelSLDA.java* a *ParallelStemmerPipe.java*, které jsou zcela prosté komentářů a nejspíše převzaté. Převzaté části zdrojových kódů by měly být jasně označeny.

Dle mého názoru splnil student bez pochyb všechny body zadání práce. V práci nad rámec zadání pracuje s LDA modelem. Tato část se mu však příliš nepovedla a student by udělal lépe, pokud by jí vůbec do práce nezahrnul. Vzhledem k tomu, že se jedná o rozšíření nad rámec zadání, jsou výtky týkající se LDA do hodnocení promítnuty jen okrajově.

[1] `http://www.statmt.org/moses/RELEASE-3.0/models/`

SOUHLASÍ
S ORIGINÁLEM

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

Vzhledem k vědecké podstatě zadání a jeho velké náročnosti, navrhuji hodnocení známkou **výborně** a práci doporučuji k obhajobě.

Otázky

- BLEU – Definujete $Count_{clip}$ jako maximální počet shodných n-gramů. Z čeho se bere maximum?
- LDA – Popište vztah mezi multinomiálním a Dirichletovo rozdělením. Z jakého rozdělení se samplují náhodné veličiny témat a z jaké se samplují náhodné veličiny slov?
- LDA – Jak se liší použitý vícejazyčný LDA model od standardního LDA?
- Truecasing (str. 31) – Je velikost písmen konvertována pouze u prvního slova ve větě nebo u všech slov? Proč se Truecasing dělá?

Ing. Miloslav Konopík, Ph.D.
(oponent DP)

V Plzni 2. června 2016

Reference

[1] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc., 2008.

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