Oponentský posudek diplomové práce

Autor a název předložené práce

Lucia Bečvarová: *Content classification in legal documents*

Téma práce

Předložená práce se zabývá automatickým zpracováním českých a anglických textových, částečně strukturovaných dat a jejich automatickou klasifikací na základě strojového učení. Zadání konkrétní úlohy vzniklo z podnětu komerčního subjektu *Datlowe, s.r.o.* a z jejich potřeby automaticky zpracovávat dokumenty psané v přirozeném jazyce. Jde o formální právní smlouvy a obchodní dokumenty jednak české, které jsou veřejně dostupné, jednak anglicky psané, které jsou vlastnictvím klienta firmy *Datlowe*. Významnou okolností je fakt, že trénovací a testovací příklady pro strojové učení mohla autorka převzít již hotové.

Cílem práce bylo navrhnout, implementovat a vyhodnotit klasifikační modul, který kategorizuje jednotlivé odstavce dokumentů. Tato klasifikace má být předstupněm pro automatickou extrakci klíčových informací. Zadání práce předpokládá využití existujících nástrojů pro automatickou analýzu jak anglických, tak českých jazykových dat.

Struktura práce: rešerše – data – experimenty – evaluace

Téma práce je vysoce praktické a očekává se nasazení autorčiných výsledků do provozního systému firmy *Datlowe*. Zaměření práce je implementační a experimentální. Autorka však nejprve musela pečlivě analyzovat problém samotný, povahu vstupních dat a metody automatického zpracování, které se ve společnosti *Datlowe* již užívají. Tomu je věnováno prvních cca 40 stran práce (kap. 1 až 4). Na dalších cca 25 stranách (kap. 5 až 7) jsou popsány metody, které autorka sama implementovala (s využitím dostupných knihoven pro podporu strojového učení), a experimentální výsledky včetně diskuse a závěrečného shrnutí. Posledních cca 25 stran předložené práce tvoří odkazy na literaturu a přílohy, zejména tabulky výsledků.

Uspořádání kapitol předložené práce je vcelku standardní. Krátká úvodní kapitolka představuje stručný úvod do problematiky, vysvětlení motivace, zadání projektu a rozdílu mezi českými a anglickými vstupními dokumenty. Anglické dokumenty jsou ve srovnání s českými spíše delší a obsahují podrobnější informace, české dokumenty jsou stručnější a tudíž pravděpodobně také se snadněji rozpoznatelnou a poněkud jednodušší strukturou. Rešerše v kapitole 2 se zaměřuje především na v práci využité tradiční metody strojového učení. V kapitole 3 jsou podrobně popsána dostupná data včetně anotací, pomocí nichž autorka získala vzorové příklady, tzv. „zlatou sadu“. V kapitole 4 autorka vysvětluje, jak jsou navrženy experimenty a jaké užívá evaluační metriky a hotové fungující metody pro předzpracování vstupních textů.

Těžiště autorčiny vlastní práce je v kapitolách 5 a 6. Pro předzpracování dat užívá systém GATE a pro strojové učení a automatickou klasifikaci užívá knihovny Pythonu. Kapitola 5 popisuje nejprve konstrukci vektoru příznaků a pak ladění hyperparametrů učících algoritmů. Obé je realizováno pomocí knihoven dostupných v Pythonu. V kapitole 6 autorka analyzuje výsledky jednotlivých metod individuálně pro jednotlivé výstupní třídy a přidává experimenty s kombinacemi klasifikátorů a analýzy významnosti příznaků. Odevzdaná práce obsahuje elektronickou datovou přílohu se zdrojovými kódy a s několika ukázkovými příklady zpracovávaných dokumentů.

Připomínky a otázky k obhajobě

- Kapitola 4.2.3 se zdá být problematická. Nerozumím, proč mezianotátorská shoda, jež je důležitým indikátorem pro posouzení smysluplnosti a úspěšnosti automatické klasifikace, není měřena vzhledem ke kategorizaci celých odstavců, což je v této práci cílem strojového učení. Vypadá to, jakoby měření mezianotátorské shody bylo navrženo pro jinou úlohu, než jakou ve skutečnosti řešíte. – Další otázka s tím souvisí: Co vyvozujete ze zjištění sepsaných v tabulce 4.8?

- Nedočetl jsem se, jaká je velikost „slovníku“ textových příznaků popsaných v 5.1.1? Je to pouze kolem tří tisíc unigramů a bigramů, jak by naznačovala poznámka v 5.1.2? Byly textové příznaky filtrovány nějak jinak než pomocí parametru max_df? Není to popsáno dostatečně.

- V 5.1.2 se vyjadřujete o „ztrojení“ prostoru příznaků slovy „it might help“ – bylo experimentálně ověřeno, že to skutečně pomáhá a v jaké míře?

- Vysvětlete, co konkrétně v daném kontextu znamená výraz „vice versa“ v 5.1.2? Není uvedeno, jak pracujete se záporně zkorelovanými příznaky?

- Byly nějak exaktně evaluovány heuristiky popsané v 5.4?

- Uvažovala jste o možnosti použít nejprve binární klasifikátor pro predikci, zda je odstavec pro extrakci hledaných informací relevantní, a teprve v druhém kroku trénovat klasifikaci? Jaký na to máte názor?

- Úvaha uvedená v 6.4 o tom, že všechny tři typy příznaků jsou relevantní pro klasifikaci, je sice platná, ale nedokazuje, že všechny tři typy příznaků jsou potřebné. Víte proč?

- V práci nenacházím vůbec žádnou analýzu chyb. To by myslím bylo velmi žádoucí a mohlo by to být možná i významné pro další zlepšování klasifikačního modulu.

- V práci nenacházím srovnání s podobnými projekty – ani co do výsledků, ani co do metod. Je jasné, že přímé srovnávání s podobnými projekty by stěží mohlo být vodítkem pro posouzení kvality výsledků samotné autorky. Numerické výsledky by byly víceméně nesrovnatelné. Přesto se domnívám, že přehled podobných projektů a jejich porovnání by bylo užitečné a inspirující.

Celkové hodnocení

Předložená práce zřejmě splňuje požadované zadání. Její celkovou úroveň hodnotím jako standardní, s výše uvedenými výhradami. Za největší nedostatek jednoznačně považuji fakt, že nebyla provedena žádná detailnější analýza chyb. Provedené experimenty jsou popsány dosti dobře a srozumitelně. Rozsah textu je adekvátní. Formální požadavky kladené na diplomovou práci jsou splněny. Autorka dostatečně cituje použitou literaturu, jejíž seznam na konci práce obsahuje 16 zpravidla cizojazyčných položek. Práce je napsána slušnou a snadno čitelnou angličtinou prakticky bez závažnějších chyb. Proto předloženou práci doporučuji k obhajobě.

V Praze, 1. června 2017

RNDr. Martin Holub, Ph.D.