Posudek vedoucího diplomové práce

Student:	Bc. Lukáš Rejmont
Číslo studenta:	E160000
Název diplomové práce:	Extrakce informace z textových dokumentů pro potřeby České obchodní inspekce
Cíl práce:	Cílem práce je charakterizovat současné přístupy k extrakci informace z textu (zejména ručně definované šablony a strojové učení), charakterizovat zvolenou metodu extrakce, na příkladu textů pro potřeby České obchodní inspekce provést předzpracování dokumentů, navrhnout jmenné entity pro extrakci, provést asociaci jmenných entit do relací a zhodnotit přesnost modelu na reálných datech.
Vedoucí práce:	doc. Ing. Petr Hájek, Ph.D.
Studijní program:	N6209 Systémové inženýrství a informatika
Akademický rok:	2017/2018

Náročnost tématu

	výborně	velmi dobře	vyhovující	nevyhovující	nelze hodnotit
Teoretické znalosti	☒	☐	☐	☐	☐
Vstupní údaje a jejich zpracování	☒	☐	☐	☐	☐
Použité metody	☒	☐	☐	☐	☐

Kritéria hodnocení práce

	výborně	velmi dobře	vyhovující	nevyhovující	nelze hodnotit
Stupeň splnění cíle práce	☒	☐	☐	☐	☐
Původnost zapracování tématu	☒	☐	☐	☐	☐
Adekvátnost použitých metod	☒	☐	☐	☐	☐
Hloubka provedené analýzy (ve vztahu k tématu)	☒	☐	☐	☐	☐
Logická stavba práce a rozsah	☒	☐	☐	☐	☐
Práce s českou a zahraniční literaturou včetně citací	☒	☐	☐	☐	☐
Formální úprava práce (text, grafy, tabulky)	☒	☐	☐	☐	☐
Jazyková úroveň (styl, gramatika, terminologie)	☒	☐	☐	☐	☐

Využitelnost výsledků práce

	vysoká	střední	nízká	nelze hodnotit
Pro teorii	☐	☒	☐	☐

Pro praxi	☒	☐	☐	☐

Ostatní připomínky k práci

Automatická extrakce informace je s rostoucím množstvím textových dokumentů vysoce aktuálním tématem. Zvláště u veřejných institucí může vést ke značné úspoře času a nákladů. Jedním z důležitých problémů ČOI je kontrola internetových obchodů z hlediska dodržování občanského zákoníku. Diplomant v této práci navrhuje systém pro podporu rozhodování v rámci této kontroly. Tento systém dokáže automaticky extrahovat informace z obchodních podmínek internetových obchodů. Problém se ukázala nedostupnost české mutace nástrojů pro tuto extrakci. Proto se diplomant zaměřil na obchody, které mají své webové stránky v anglickém jazyce. To je možné označit za hlavní omezení z hlediska praktické využitelnosti výsledků práce. Identifikace problémových oblastí kontroly internetových obchodů je ale provedena v českém jazyce. Ke splnění cílů práce diplomant navrhl několik subystémů, včetně automatického vyhodnocování problémových oblastí kontroly internetových obchodů, automatické extrakce názvů nejčastěji hodnocených obchodů, automatické stažení jejich obchodních podmínek, extrakce klíčových jmenných entit z textu a jejich asociace s databází Wikidata. V těchto oblastech provedl také předzpracování textu, což v případě webových stránek není triviální. Na závěr diplomant provedl zhodnocení přesnosti extrakce a výsledky vhodně diskutoval. Práce je technicky zpracovaná velice pečlivě, kapitoly na sebe logicky navazují a výsledky jsou snadno reprodukovatelné díky přiloženému software na CD.

Vyjádření k výstupům ze systému Theses

Kontrola plagiátorství: nejvyšší míra podobnosti 0 %, práce není plagiát.

Otázky a náměty k obhajobě

V rámci obhajoby diskutujte omezení navrženého systému a jeho další možná rozšíření.

Závěrečné hodnocení

Práci **doporučuji** k obhajobě.
Tuto diplomovou práci navrhuji hodnotit známkou: **A**

V Pardubicích 9.1.2019

Podpis .....