



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 vedoucího diplomové práce

### Jakub Sido: Automatická extrakce příspěvků z diskusních fór

Diplomantovým úkolem bylo navrhnout a implementovat systém pro extrakci příspěvků z diskusních fór. Diplomát využil existujícího systému pro automatickou extrakci dat z webových stránek založenou na učení bez učitele. Výhodou tohoto přístupu je, že nepotřebuje pro svou práci žádná trénovací data. Nevýhodou je, že extrahovaná data nejsou sémanticky označena. Data jsou seskupena tak, že obsahují informaci stejného druhu, například jméno autora nebo datum příspěvku, ale není známo, o jakou konkrétní informaci jde, tedy zda jde o jméno autora nebo datum příspěvku. Diplomant navrhl metody založené na strojovém učení s učitelem a zpracování přirozeného jazyka tak, aby systém byl schopen korektně označit, o jakou informaci se jedná (tedy určit zda jde o jméno autora, datum příspěvku anebo nějakou další informaci, kterou systém umí rozpoznávat).

Diplomant soustředil po dohodě s vedoucím většinu prací na letní semestr. Pracoval intenzivně a pravidelně se dostavoval na dohodnuté konzultace. Vlastní přístup k řešení navrhl vesměs sám a konzultoval zejména využití metod zpracování přirozeného jazyka. V jisté fázi řešení se rozhodl využít existující systém pro extrakci dat z webových stránek, ač mu bylo vedoucím doporučeno vyvinout vlastní algoritmus. V závěru se ukázalo, že použitý systém *RoadRunner* má řadu nedostatků a diplomant strávil mnoho času napravováním chyb systému následnou korekcí výsledků. Možná by bylo lepší, kdyby diplomant vyvinul vlastní systém, ovšem taková rozhodnutí nelze vždy dobře predikovat a nejistota a riziko špatného rozhodnutí patří k vědecké práci.

Dosažené řešení je funkční a použitelné. Zároveň obsahuje velký potenciál pro vylepšení a tak má diplomant dobrý základ pro následné doktorské studium, na které se hlásí.

Zdrojové kódy jsou adekvátně komentovány. Dosažené výsledky se mi s malou odchylkou podařilo relativně snadno replikovat.

Práce je napsaná v sázecím programu  $\text{\LaTeX}$ . V práci se nachází jisté typografické prohřešky. Diplomant hojně používá rastrových obrázků, i když je mohl snadno nahradit obrázky vektorovými (například obrázky 2.2, 2.3, 2.4, 3.1, 4.1 a další). Několik obrázků obsahuje anglické texty (například obrázky 2.2, 3.2, 3.3, 6.5 a další) a to i přes to, že ve vlastním textu práce jsou používány české ekvivalenty. Diplomant mohl také lépe popsat použitá trénovací data. Není zřejmé, kolik webových stránek využil a počty jednotlivých tříd si musí čtenář dopočítat sám. Celkovou úspěšnost navrženého systému je nutno v práci dohledávat. Bylo by vhodnější toto číslo uvést do závěru a zhodnotit, zda je dostatečné pro praktické nasazení systému. Práce s literaturou je dobrá. Student dostatečně cituje relevantní prameny.

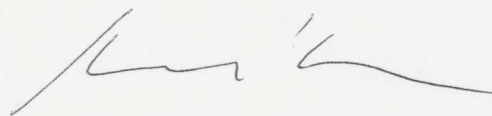
SOUHLASÍ 1  
S ORIGINÁLEM

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

Závěrem konstatuji, že diplomant splnil zadání ve všech jeho bodech. Velmi kladně hodnotím nápad využít pro extrakci dat algoritmus učení bez učitele a získaná data následně sémanticky označit metodami zpracování přirozeného jazyka. Myslím si, že práce má dobrý potenciál pro další vědecké zkoumání i pro praktické využití.

Dle mého názoru diplomant prokázal, že je schopen inženýrským způsobem řešit zadané problémy. Jeho práci doporučuji k obhajobě a vzhledem k dosaženým výsledkům práci hodnotím klasifikačním stupněm

“výborně”.



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

V Plzni 12. června 2017

**SOUHLASÍ  
S ORIGINÁLEM**



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

①