Posudek oponenta diplomové práce

Autor/autorka práce: **Jiří Martínek**

Název práce: **Inteligentní vyhledávání dokumentů**

Obsah práce

V rámci zpracování diplomové práce byl vytvořen poměrně komplexní nástroj pro zpracování naskenovaných dokumentů uložených v podobě rastrových obrázků. Postupně jsou řešeny problémy optického rozpoznávání znaků, opravy chyb vzniklých při rozpoznávání a indexace zpracovaných dokumentů vhodným nástrojem. Cílem uvedených operací je umožnit efektivní vyhledávání v množině dokumentů.

Po analýze byl vybrán OCR nástroj Tesseract. Oprava chyb je řešena kombinací Viterbiho algoritmu, jazykového modelu a volitelně Levenshteinovy vzdálenosti. Indexace dokumentů je provedena nástrojem Apache Solr. Dosažená hodnota word accuracy je 80 až 90 % v závislosti na kvalitě skenů. Navíc oproti zadání je přidána klasifikace dokumentů pomocí strojového učení.

Program je kvalitně zpracován. Obsah přiloženého DVD umožňuje bezproblémové spuštění a otestování funkčnosti systému.

Text práce je psán vhodnou formou s minimem překlepů a pravopisných chyb. Jednotlivé kapitoly by bylo dle mého názoru vhodnější uspořádat podle workflow programu. Tj. nejdříve OCR, pak oprava chyb a nakonec indexace dokumentů, která je i implementačně nejméně zajímavá. U analýzy OCR systémů by bylo vhodné přidat tabulku shrnující vlastnosti jednotlivých nástrojů a také neuvádět některé nástroje, které nevyhovují požadavkům zadání. V závěru je uvedena úspěšnost systému, ale není jasné, na jakých datech byla dosažena – počet dokumentů kvalita atp. Bylo by vhodné přidat ukázky dokumentů pro lepší představu (pokud to charakter dokumentů umožňuje).

Práce s literaturou je na vysoké úrovni a použité prameny jsou v práci dostatečně citovány.

Zadání bylo splněno bez výhrad.

Dotazy k práci

Provedl jste porovnání několika OCR nástrojů. Máte k dispozici nějaké číselné hodnoty vyjadřující jejich úspěšnost?

Podařilo se u některých OCR nástrojů zjistit, na jakých metodách jsou založeny?

Jak vypadá robustní dotaz? (použito v popisu nástroje Elasticsearch)

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

V Plzni 6.6.2017

Ing. Ladislav Lenc, Ph.D.