**Posudek vedoucího diplomové práce**

## Jiří Martínek: Inteligentní vyhledávání dokumentů

Cílem práce je návrh a implementace systému pro vyhledávání dokumentů v podobě naskenovaných rastrových obrázků. Systém musí v první fázi převést dokumenty do textové podoby a dále je zaindexovat do fulltextové databáze pro jejich efektivní vyhledávání.

Práce autora začíná popisem relevantních pojmů v oblasti řešené problematiky. Zde je popsána úloha indexace a vyhledávání, které jsou pro vyřešení problému nezbytné. V další části autor přibližuje úlohu optického rozpoznávání znaků (OCR). Dále jsou popsány a srovnány dostupné systémy pro OCR a fulltextové vyhledávání. Na základě provedeného srovnání zvolil student pro OCR analýzu systém Tesseract a pro fulltextové vyhledávání Apache Solr. Následující kapitola se zabývá popisem možností detekce a opravy chyb vzniklých při převodu písma do textové podoby. Jako velmi perspektivní se jeví znakové jazykové modely.

Dále se diplomant zabývá vlastním řešením, kde je popsán návrh a implementace vyvíjeného systému. Další kapitola popisuje provedené experimenty. Zde autor ukazuje vliv rozlišení dokumentů na výsledky OCR, dále experimentálně stanovuje optimální poměr vah pro kombinaci výstupu systému Tesseract a znakových jazykových modelů. Výsledky OCR po opravě chyb dosahují přesnosti (word accuracy) kolem 80%, což považuji za velmi dobré. Vzhledem k provedené opravě chyb dokáže systém lépe najít relevantní dokumenty dle zadaného dotazu. Pro zlepšení přesnosti vyhledávání byla dále navržena a integrována automatická klasifikace dokumentů podle obsahu, která zúží množinu relevantních dokumentů.

Průvodní dokument (55 stran + přílohy) je vytvořen v systému LaTeX. Má přehlednou strukturu. Dokument je na dobré jazykové úrovni, neobsahuje pravopisné chyby ani překlepy. Práce obsahuje několik drobných nepřesností, které odpovídají znalostem studenta magisterského studia.

Přiložené DVD má logickou strukturu, obsahuje readme soubor s popisem celého DVD. Dle návodu se bez problémů podařilo systém nainstalovat a spustit. Při testování byl systém plně funkční, nebyly nalezeny žádné chyby.

Předložená diplomová práce splňuje zadání. Je třeba dále zdůraznit, že téma práce je rozsáhlé a složité a vyžadovalo nastudování řady informací z oblasti umělé inteligence. Autor zde prokázal dobré znalosti nejen z informatiky, ale i statistiky. Přesvědčivě ukázal, že dokáže samostatně analyzovat a řešit složité problémy. Práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

**„výborně“**

doc. Ing. Pavel Král, Ph.D.
vedoucí DP

V Plzni 5. června 2017

SOUHLASÍ
S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky
①