Posudek oponenta bakalářské práce

Autor/autorka práce: **Josef Baloun**

Název práce: **Prohledávání dokumentů podle automaticky extrahovaných vzorů**

Obsah práce

Práce se zabývá vyhledáváním slov v ručně psaných dokumentech. Úloha vyhledávání je prováděna na již segmentovaných datech, kde dokument je rozdělen na obrázky jednotlivých slov. Slova jsou pak porovnávána na základě vizuální podobnosti. Pro řešení jsou použity neuronové sítě.

Kvalita řešení a dosažených výsledků

Práce je naprogramována v jazyce Python s využitím knihovny Keras pro implementaci neuronových sítí. Software je dodán jako množství spustitelných souborů, kdy každý soubor odpovídá jedné konfiguraci. Z pohledu uživatele je spouštění jednoduché, nicméně je zde velké množství opakujícího se kódu. Za zvážení by stálo spouštění z jednoho skriptu s nastavením parametrů.

Prezentované výsledky dosahují hodnot MAP nad 90 procent. Srovnání s literaturou je ale problematické, protože jsou uvedeny výsledky pouze z jedné publikace a není stoprocentně jisté, že postup vyhodnocení je stejný.

Bez ohledu na srovnání s literaturou je zde velký potenciál pro využití prezentovaných metod v rámci projektu řešeného na KIV.

Formální úroveň

Na začátku práce by bylo vhodné přesně specifikovat úlohu. Není zde zcela patrné, že se jedná o obdobu „information retrieval", kde pro hledané slovo (obrázek) je nalezena množina relevantních dokumentů. Není také jasné, jestli systém hledá pouze jednotlivá slova, řádky textu, nebo celé stránky, které slova obsahují. S tím souvisí i uvedení evaluačních metrik až v kapitole 6. Jejich uvedení na začátku práce by přispělo k lepšímu pochopení.

Popis datových sad je částečně v kapitole 2 a pak je upřesněn v kapitole 6. Bylo by vhodnější mít informace na jednom místě.

V práci je nejednotné použití zkratek. Bylo by vhodné uvést význam při prvním použití, ne jen v seznamu zkratek.

Text práce obsahuje minimální množství překlepů a chyb.

Seznam použité literatury je rozsáhlý a zdroje jsou odpovídajícím způsobem citovány.

Splnění zadání

Zadání bylo splněno bez výhrad

Dotazy k práci

V práci [13] je prováděno vyhledávání na úrovni řádek textu. Na straně 46 je zmíněno porovnání s touto prací. Jsou výsledky opravdu porovnatelné?

Graf 6.9 na straně 35 ukazuje chování systému pro různé způsoby porovnání vektorů. Z grafu jednoznačně nevyplývá, že Euklidova vzdálenost funguje lépe. Naopak kosinová vzdálenost je po 9 epochách lepší. Zkoušel jste výpočet s větším počtem epoch?

Jaká sada byla použita pro trénování u výsledků uvedených v tabulce 6.2? Základní, nebo některá z rozšířených?

Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 16.5.2018

Ing. Ladislav Lenc, Ph.D.