Posudek oponenta diplomové práce

Autor/autorka práce: **Jakub Morávka**

Název práce: **Klasifikace dokumentů s použitím hierarchické reprezentace**

Obsah práce

Cílem práce bylo prozkoumat vliv hierarchické reprezentace dokumentů na výsledky jejich klasifikace do více tříd. Pro klasifikaci jsou použity metody založené na neuronových sítích. Výsledky jsou ověřeny na vybraných standardních datových kolekcích v českém a anglickém jazyce.

Kvalita řešení a dosažených výsledků

Realizované řešení je plně funkční. Autor porovnal výsledky celkem dvou různých topologií neuronových sítí (MLP a CNN), provedl velké množství experimentů s velmi zajímavými výsledky.

Formální úroveň

Průvodní dokument (84 stran + přílohy) je vytvořen v systému LaTeX. Číslování ale končí překvapivě kap. Závěr a v kap. Literatura čísla stran chybí. Tato kap. by měla být rovněž číslovaná. Dokument je velmi rozsáhlý, obsahuje ale pouze zajímavé relevantní informace a ne „vycpávky". Práce je na velmi dobré jazykové úrovni, neobsahuje pravopisné chyby, jen několik zcela ojedinělých překlepů. Práce obsahuje některé drobné nepřesnosti, které ale odpovídají znalostem studenta této formy studia. Přiložené DVD má přehlednou strukturu. Kořen správně obsahuje readme soubor s popisem celého disku. Program v jazyce Python je přehledný a dobře strukturovaný.

Práce s literaturou

V práci je uvedeno celkem 88 odborných publikací, ze kterých student čerpal. Je zřejmé, že nebyly všechny přečteny kompletně, ale jen přímo související pasáže. Což je ale zcela v pořádku. Uvedený počet referencí považuji za velmi nadstandardní s ohledem na typ práce.

Splnění zadání

Zadání bylo splněno v plném rozsahu.

Dotazy / připomínky k práci:

- Na str. 49 uvádíte, že jste s ohledem na použité PC zvolil délku embeddins vektorů 150. Obvykle se ale používají delší vektory, zpravidla délky 300 prvků. Bylo by vhodnější změnit použitý HW např. za metacentrum s dostatečným výpočetním výkonem, protože kratší délka vektorů může být jedním z důvodů některých horších výsledků.

- Na str. 52 uvádíte, že jste w2v a fastText natrénoval na trénovací sadě českého korpusu CTDC v 2.0. Tento korpus je ale příliš malý na toto trénování a bylo by vhodné použít např. českou wikipedii. Proč jste tak neučinil?

- V sekci 8.2.5 stanovujete vhodný počet vět pro hierarchickou reprezentaci. Na základě experimentu bylo stanoveno 30 vět pro češtinu. Pro angličtinu jste vybral 20 vět a volbu zdůvodňujete stejným % dokumentů, které nejsou zkráceny. Toto zdůvodnění se mi zdá diskutabilní. Proč jste neprovedl stejný experiment pro anglický jazyk?

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

V Plzni 27.8.2018

doc. Ing. Pavel Král, Ph.D.

SOUHLASÍ S ORIGINÁLEM

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