Posudek oponenta bakalářské práce

Autor/autorka práce: **Petr Hlaváč**

Název práce: **Klasifikace textových dokumentů pomocí neuronových sítí**

Obsah práce

Cílem práce bylo prostudovat vliv předzpracování, tj. tokenizace, lemmatizace a stemmingu na výsledky klasifikace dokumentů. 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 sedmi různých topologií neuronových sítí, provedl velké množství experimentů s překvapivými výsledky. Experimenty ukázaly, že vícevrstvý perceptron, tj. základní síť, dosahuje nejlepších výsledků prakticky ve všech případech.

K samotnému popisu metod / řešení mám následující připomínky:

- Kapitola 2 obsahuje jen dvě metody reprezentace dokumentů. Metod ale existuje celá řada (n-gramy, tf-idf, ...), proto by bylo vhodné je doplnit.

- Pravidlo pro volbu počtů neuronů v jednotlivých vrstvách MLP (sekce. 7.1) je velmi přibližné a nejde obecně použít tak, aby byla dosažena optimální struktura sítě.

- Kapitola 8 obsahuje popis jen jedné knihovny pro implementaci neuronových sítí a to knihovny Keras s backendem TensorFlow. Knihoven ale existuje celá řada, viz výčet na začátku kapitoly. Knihovny z výčtu by měly být alespoň stručně popsány v textu. Popis možností knihovny Keras by měl být podrobnější. Tato kapitola dále obsahuje popis Metacentra, který není z pohledu „obecného“ čtenáře podstatný. Keras (s TensorFlow) je možné spustit na libovolném PC (technologie CUDA velmi žádoucí).

Formální úroveň

Průvodní dokument (39 stran + přílohy) je vytvořen v systému LaTeX. Má přehlednou strukturu. Očekával bych ale podrobnější popis některých pasáží, konkrétně viz předchozí sekce. Dokument je na velmi dobré jazykové úrovni, neobsahuje pravopisné chyby, jen několik málo překlepů. Některé anglické termíny by bylo vhodné přeložit do češtiny (např. „multi-class“ nebo „multi-label“). Práce obsahuje některé nepřesnosti, které ale odpovídají znalostem studenta bakalářského studia. Přílohy by mely být číslovány jiným způsobem, než samotný text práce. Obrázky 3.1 a 3.2 ani Listing 3.1 nejsou popsány v textu. Pokud je použita v popisu obrázku reference, tečka by měla být uvedena až na konci popisku. Přiložené DVD má přehlednou strukturu. Kořen správně obsahuje popis celého disku.

Práce s literaturou

Student nastudoval celkem 24 odborných publikací, které jsou většinou on-line a v anglickém jazyce. Tento počet považuji za nadstandardní s ohledem na typ práce. Bohužel je formát některých referencí chybný.

Splnění zadání

Zadání bylo splněno s připomínkami (viz výše).

Dotazy k práci

- Vysvětlete odkud se berou v Obr. 6.2 hodnoty w_1, w_2... a x_1, x_2 ...?

- Z textu není jasné, kolik skrytých vrstev má Vaše MLP implementace. Jde o dvě skryté vrstvy (viz Obr. 7.1)?

- Z textu práce nejsou jasné přesné konfigurace dalších neuronových sítí. Prosím vysvětlete.

- Proč myslíte, že vícevrstvý perceptron překonal všechny ostatní metody?

- Jakým způsobem jste vytvářel seznam nevýznamových slov? Slova „článek“, „zprávy“ nebo “strana“ významová jsou.

Vzhledem k připomínkám uvedeným výše navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 17.8.2018 doc. Ing. Pavel Král, Ph.D.