

Hodnocení oponenta diplomové práce

Autor/autorka práce: **Lukáš Hain**

Název práce: **Experimenty s Apache Mahout**

Cílem práce bylo prostudování knihovny pro strojové učení Apache Mahout, která využívá metody MapReduce pro paralelizaci výpočtů, a její následné využití při řešení tří běžných úloh strojového učení.

Text práce je srozumitelný, logicky strukturovaný. Po formální stránce má text několik chyb. Citace jsou uváděny bez mezery mezi slovem, např. slovo[1] místo slovo [1]. Proměnné popisované v textu by měly být sázeny jiným typem písma, aby je bylo možné odlišit od zbytku textu. Autor špatně užívá čárku před nebo (i v případech, kdy mohou nastat obě situace). Autor používá (a počesťuje) některá anglická slova, která se nedají považovat za běžné termíny nebo se běžně užívá jejich překlad (canopa/canopu, nody, labely). Text občas nenavazuje, protože autor předpokládal jiné umístění obrázku či tabulky (např. str. 29). Zásadním proviněním (po formální stránce) je kapitola 3.3 (str. 17), která neobsahuje žádný text, pouze z textu neodkázanou tabulku. Některé vzorce jsou nevhodně vysázené (např. funkce `argmin` na str. 10).

Po obsahové stránce text dostatečně pokrývá řešenou problematiku, ale k některým částem mám výhrady. Kapitola 2.3.3 chybně popisuje lineární regresi jako klasifikační algoritmus. V kapitole 7 je zřejmé, že student ne zcela pochopil iterativní trénování u metody logistické regrese. Kapitola 2.2 by mohla být zpracována lépe. Výhrady mám k popisu vzorců, kde často chybí popis všech parametrů. V některých případech nesouhlasí čísla indexů.

Výsledky na jednotlivých úlohách ukazují, že student zvládl dané úlohy vyřešit. V některých případech by ale řešení mohlo být o něco lepší (z pohledu kvality výstupu).

Zdrojové kódy jsou komentované (i když ne stoprocentně). Množství kódu odpovídá povaze práce. V textu práce nejsou uvedeny příkazy (nebo parametry) pro spuštění aplikací. Programy běží a vypisované výsledky se zdají být adekvátní (jejich ověření je netriviální).

Student splnil všechny body zadání. Nad rámec zadání jsou v práci řešeny čtyři úlohy namísto tří.

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

Dotazy k obhajobě:

1. Zkoušel jste měřit urychlení při použití SingleMachine a MapReduce (kapitola 3.3)? Je SingleMachine jednovláknový?
2. Testoval jste i Apache Spark? Jaký byl rozdíl v rychlosti?

V Plzni 1. 6. 2015


Ing. Michal Konkol

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

②

