Posudek oponenta bakalářské práce

Název práce: Heuristiky v dolování dat z grafů pomocí vnoření uzlů
Autorka: Adeliia Gataullina
Vedoucí práce: Ing. Matej Mojzeš, Ph.D.

Předložená práce se zabývá problematikou využití heuristik pro nalezení optimálních hodnot parametrů algoritmu *node2vec* aplikovaného na úlohu shlukování na grafech a patří do oblasti vytěžování znalostí z dat. Kromě úvodu a závěrečného shrnutí je rozdělena do čtyř kapitol. V prvních třech kapitolách autorka postupně popisuje princip fungování neuronových sítí, reprezentaci grafů příznakovými vektory pomocí algoritmu node2vec, tento samotný algoritmus, a nakonec využití heuristik pro optimální nastavení jeho parametrů, což je originální přínos této práce. V praktické části v kapitole 4 autorka aplikuje postup navržený v teoretické části na datové sadě z reálného světa. Výsledky shlukové analýzy s použitím navržené metody jsou vyhodnoceny oproti očekávaným shlukům z datové sady.

Předložená práce je přehledně a logicky strukturována. Odkazovaný zdrojový kód v jazyce Python prokazuje schopnost autorky používat spektrum nástrojů, se kterými se studenti na fakultě nesetkají, avšak v praxi jsou de facto standardem pro datovou analýzu.

Po formální stránce práce bohužel trpí některými nedostatky. Pochopení textu ztěžuje nedůsledná práce s odbornými pojmy. Některé z nich nejsou v práci vůbec formálně definovány, jako například *rand index*, který autorka používá pro vyhodnocení navrhované metody. Jiné pojmy jsou na druhou stranu použity ve více významech, někdy velmi nestandardních. Například pojem *účelová funkce* (*objective function*) je v druhém odstavci sekce 3.1. evidentně použit hned dvakrát s odlišnými významy. Pro zasazení práce do kontextu, snazší interpretaci výsledků a zhodnocení přínosu navržené metody mi též v textu chybí shrnutí související práce a porovnání výsledků se standardně používanými metodami.

Při obhajobě práce navrhuji zodpovězení následujících otázek:

1) V textu práce opakovaně zmiňujete, že byl použit *rand index*, avšak z jiného místa v textu a ze zdrojových kódů se zdá, že používáte *adjusted rand index*. Vysvětlete rozdíl mezi těmito dvěma metrikami (ideálně na konkrétním příkladu) a uveďte na pravou míru, kterou variantu v práci používáte.
2) Pro shlukovou analýzu používáte algoritmus *k-means*, jenž požaduje jako jeden ze vstupních parametrů počet výsledných shluků. Vysvětlete, proč pro hodnotu tohoto parametru rovněž neprovádíte optimalizaci pomocí heuristik.

I přes popsané výhrady formálního charakteru mám za to, že cíle vytyčené v úvodu práce byly naplněny a předložená práce splňuje požadavky kvalifikační práce na bakalářské úrovni. Proto ji hodnotím známkou **D (uspokojivě)**.

V Praze dne 20. 8. 2020

Ing. František Blachowicz