CSC 665 Assignment 4

Yongcheng Zhan 03/25/2017

I am going to implement the algorithm described in the following paper:

Tang, Jian, Meng Qu, and Qiaozhu Mei. "Pte: Predictive text embedding through large-scale heterogeneous text networks." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

My current research is to explore the patterns of e-cigarette use with the help of social media user-generated contents. I have done several preliminary research using topic modeling and text mining methods. However, they are not deep enough. I learnt idea of text embedding from my colleagues and found this paper pretty interesting. Generally speaking, this paper aimed to represent words into lower dimension vectors. Three networks: word-word, word-document, and word-label networks were proposed and the relationships in the networks were utilized to build word representations. I am trying to implement the algorithm in this paper by using Python. If possible, I would like to try to build one more network: word-user network, which incorporate one more layer of information in the social media. I can write a paper if this can be done.

This paper provides several datasets for testing. The results of text classification are also reported in the paper. If I can repeat the result, I can say this project succeed. Several compared algorithms were used in the paper, including: bag-of-words, skip-gram, PVDBOW, PVDM, LINE. These methods can be used to evaluate the performance of the PTE algorithm.

As for my own research, because a new layer, user layer, is added to the PTE model, I need to label several social media datasets myself. I am not sure whether I can finish this part by the end of the semester. But as least I can have a try on a small one, and provide some preliminary findings.