PROJECT PROPOSAL

Reproduce A Paper: A cross-collection mixture model for comparative text mining

1. **What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.**

   Xi Luo (Captain): xiluo4

   Yuheng Zhang: yuhengz2

2. **Which paper have you chosen?**

   ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2004). ACM, New York, NY, USA, 743-748. DOI=10.1145/1014052.1014150

3. **Which programming language do you plan to use?**

   python

4. **Can you obtain the datasets used in the paper for evaluation?**

   The paper uses 2 datasets: war news, and laptop reviews. We are not able to obtain exactly the same datasets as those are used in the paper.

5. **If you answer "no" to Question 4, can you obtain a similar dataset (e.g. a more recent version of the same dataset, or another dataset that is similar in nature)?**

   For the war news, we will manually search and download 30 news articles from BBC or CNN for each of the two wars, published in one year span (May 2003 - April 2004 for Iraq war, Nov 2001 to Oct 2002 for Afghanistan war). This would be a very close proximation of the war news dataset that is used in the paper.

   For the laptop reviews, we choose 3 laptops from the Amazon.com best sellers in laptop computers: Acer Aspire 5 Slim Laptop, Apple MacBook Air, HP Chromebook 14-inch HD Laptop, and will manually download the top 40 reviews from Amazon.com. This would be comparable to the laptop reviews dataset that is used in the paper.

6. **If you answer "no" to Questions 4 & 5, how are you going to demonstrate that you have successfully reproduced the method introduced in the paper?**

   (Not applicable)