Diego Millan – NetID: diegom3

CS410: Text Information Systems - UIUC

Project Proposal

Reproducing a Paper: Contextual 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.**

   Will work individually on the project.

   Captain: Diego Millan

   NetID: diegom3

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

   The paper selected is one of the contextual text mining subtopic options, specifically the following:

   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?**

   I plan on using Python.

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

   The paper references 2 datasets, the first one is about war news that compares the Iraq and Afghanistan war. The second dataset compares 3 laptop model reviews.

   I am not able to obtain the exact same datasets used on 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)?**

   The paper references the BBC and CNN websites as the source of the news articles, also mentioning how many articles from each were selected and the time span (30 articles starting 1 year before the paper publication for the Iraq war and 26 articles on a 1 year span starting November 2001 for the Afghanistan war). Unfortunately, there is no way to know the exact articles used but a similar sample can be obtained from the same sources (will be using Google news to obtain random articles on the specified time span).

   Regarding the laptop dataset, the review source website (epinions.com) no longer exists which makes it impossible to get the same dataset. Since laptop reviews are readily available in numerous other sites (amazon.com could be a good replacement), my plan is to use more recent data to replace this dataset. I will be comparing a newer model of each of the 3 laptop brands referenced on the paper (Apple, Dell, and IBM).

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?**

   Even though the exact same datasets cannot be obtained, the alternatives should be close enough to get comparable results to the ones concluded on the paper.