Project Progress Report

Yutong Lin
NetId: yutong21

Project Introduction

The project is to reproduce the paper "A Cross-Collection Mixture Model for Comparative Text Mining". The project is in its final phase so far.

The program is improved based on the PLSA model, which includes an expectation algorithm step, a maximization algorithm step and a likelihood calculation step. The new mixture model has more matrix to update. In the PLSA model, there is a document topic coverage matrix and a topic-specific language model matrix. In the new CCMM (Cross-Collection Mixture Model) has a collection-specific topic language model. It is designed as a 3D array.

I finished the code and tried running data to test its performance. The original paper used two datasets for experiments. The first is a news dataset including about 30 articles on Iraq War and 30 articles on Afghanistan War from BBC and CNN between 2003 to 2004. I wrote a scraper to scrape from BBC and CNN websites (scraper code in news_scraper.ipynb, scraped data in wars_news.txt). The second dataset is customer reviews on three types of laptop from Apple, Dell and IBM from epinion.com. However, the epinion.com website has been shut down, so I scrape 84 customer reviews from BestBuy.com on the M1 chip Macbook Air, Dell g5 fhd gaming laptop and Lenovo Yoga c940 (scraper code in reviews_scraper.ipynb, scraped data in laptop_reviews.txt).

The model performs well and could print the common theme model and the collection-specific theme model matrix in terminal by command "python model.py". The format is the same as the original paper. The column stands for different clusters, and the probability of the word decreases by row.

Remaining Challenge

The remaining challenge is to find a better set of parameters to pass in, that could make each cluster more distinct make the top 5 words more representative. At present, the clusters seem to be too much for the laptop review dataset, so each cluster is not very distinctive. Also, there remains a question in the maximization step, I shall contact the professor to check my implementation.