Project Report

Our project aims to reproduce the results in the paper by Zhai and coworkers, “A cross-collection mixture model for comparative text mining”. In this paper, the author developed an efficient text mining model, the Cross-Collection Mixture Model (CCMix), using the Expectation-Maximization (EM) algorithm for Comparative Text Mining (CTM). The simple mixture model was originally used as a naïve solution to CTM. This method treats the multiple collections as one single collection and performs clustering. The CCMix Model, however, uses latent common themes as well as a potentially different set of collection-specific themes for each collection. These component models directly correspond to all the information we are interested in discovering.

In the paper, the authors used new articles from BBC and CNN as datasets to test this model. The comparison was taken between results from news on the Iraq War and news on the Afghanistan War. Unfortunately, we are unable to obtain the same dataset for this project. Instead, we chose the COVID-19 pandemic and the SARS outbreak in 2003 as the comparing events. We extracted 35 news articles from CNN, BBC, and HUFFPOST to form our dataset.

To build the model, we used an open-source code on GitHub as the basis, with added features on theme clustering to let it behave as similarly in the paper.

The paper suggested that the CCMix model can help reveal many interesting common aspects between the Iraq War and the Afghanistan War that the SimpMix model failed to do. The results presented by SimpMix were less meaningful, making it hard for people to extract useful information from the common words. In contrast, we were able to gather useful information from the CCMix results. For example, from cluster 5 we knew that the news about both events mentioned the diplomatic role played by the United Nations, and from cluster 4 we knew that

there were Monday briefings by an official spokesman of a political administration during both wars.

Our attempt in comparing the SimpMix and CCMix model also led to meaningful results. We set $\lambda_B = 0{:}95$ for SimpMix and set $\lambda_b = 0{:}9$, $\lambda_c = 0{:}25$ for CCMix; in both cases, the number of clusters is fixed to 5. These parameters are exactly the same as in the paper. The results are tabulated below:

SimpMix

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Common Theme words	coffee	people	students	sars	sars
	students	rules	vaccine	china	china
	people	wales	firms	disease	health
	workers	lockdown	vaccines	april	outbreak
	universities	covid	pandemic	travel	people
	university	government	people	health	virus
	health	restrictions	support	chinese	kong
	beans	england	governments	hong	hong

CCMIX

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Common Theme words	firms	sars	people	health	people
	vaccine	china	rules	virus	covid
	vaccines	disease	lockdown	people	dr
	animals	coffee	wales	kong	hospital
	governments	beijing	restrictions	students	south
	christmas	workers	government	hong	health
	prices	officials	support	china	infection
	pandemic	authorities	university	canada	patients

Although there were some interesting themes from the SimpMix results (for instance, cluster 2 mentioned the lockdown policies in Britain, and cluster 3 indicates the government's effort in inventing vaccines), we could not find *common* themes to both events. It appears that

cluster 1 to 3 solely described themes on COVID-19 and cluster 4 and 5 only covered the SARS pandemic.

The CCMix results gave us more information on the similarities between SARS and COVID-19. In cluster 1, “vaccine”, “firms” and “animals” suggest that there had been attempts for animal trials on vaccines in both cases. In cluster 4, we were able to tell that both pandemics involved an outburst of infected cases in Hong Kong, which allowed the events to draw global attention. The common theme captured in cluster 5 simply told us that a lot of patients went to the hospitals because of the diseases. The term “dr” also implied that there are large portions of interviews and advice of doctors in the news articles.

Overall, our reproduction work using CCmix model gave us similar results as in the paper. However, it was evident that there were aspects that our model failed to do. For example, words such as “coffee” and “beans” that were unrelated to our topic appear in the common theme words section in our results. One possible explanation is that we falsely collected words in the advertisement into our dataset. Since our new articles mainly came from BBC News, the article websites may be displaying the same advertisement on coffee which let “coffee” being the most popular theme word.

Lacking information on specific events was another drawback of our research. In the paper, the authors were able to gather detailed information on the themes (for example, the name of the spokesman, the name of the commander, etc.). Our results, however, only contained general aspects of the events. We were unable to figure out what exactly happened by just looking at the common theme words. This might be caused by the inherent difference of our dataset compared to the one in the paper. As the geographic scale and timescale are much larger in the case of the pandemic than the war, the focus on the news articles tends to be diverse.

Therefore, it can be hard for a word that is only important to a specific event to stand out in the common theme words list in our case. One possible solution we can think of is to set a lower λ_b value to let uncommon words more favored that set a higher λ_c to make the collection more collection-specific.