Diego Millan – NetID: diegom3
CS410: Text Information Systems - UIUC

Project Progress Report

Reproducing a Paper: A Cross-Collection Mixture Model for Comparative Text Mining

1) Which tasks have been completed?

- Similar test data has been compiled for both experiments described in the paper. I'm using about ~30 news articles per event (Afghanistan and Iraq war) from BBC and CNN news and a combined total of 250 reviews of 3 recent model of laptops (Lenovo, Dell and Apple) from Amazon.com.
- Data has been roughly curated. Major spelling errors were removed.
- I'm using the MP3 skeleton code as base since it is very similar to the paper structure-wise and can help me and reviewers follow along the code.
- The following functions have been added or implemented:
    - Init variables – Added needed variables
    - Build corpus – Builds whole corpus and individual collections. Cleans data of punctuation and digits
    - Build vocabulary – Same as MP3
    - Build term matrix – Added term matrix per individual collection
    - Build background model – This is a new function
    - Random initialization of parameters with normalization
    - Expectation step – First implementation
    - Maximization step – First implementation
- Currently writing the clustered words to 2 different text files, one for the top ten words per topic on "common.txt" and top 10 words per collection per topic on "specific.txt"

2) Which tasks are pending?

- Implement log likelihood function
- Check for errors in algorithm (see question 3, challenges faced)
- Interpret and report results in a friendly manner
- Tune Lambda B and C parameters (background and Collections "weights")
- Clean code

3) Are you facing any challenges?

In the EM updating formulas presented on the paper, I have not figured out one operation circled in the image below. The formula states to sum the terms across all d's in Ci across all Cm (collections). If my understanding is correct, this is wrong as that would pool all the documents in the corpus together and lose the focus to a specific collection ending up with only "k" themes. My current implementation only sums across all d's in Ci and normalizes based on this sum across words. This results in "k" times "number of collections" specific theme models where all probabilities sum to 1 within each of them. Will reach out to TA if I get stuck debugging.

$$
\begin{aligned}
p(z_{d,C_i,w} = j) &= \frac{\pi_{d,j}^{(n)}(\lambda_C p^{(n)}(w|\theta_j) + (1-\lambda_C)p^{(n)}(w|\theta_{j,i}))}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)}(\lambda_C p^{(n)}(w|\theta_{j'}) + (1-\lambda_C)p^{(n)}(w|\theta_{j',i}))} \\
p(z_{d,C_i,w} = B) &= \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{k} \pi_{d,j}^{(n)}(\lambda_C p^{(n)}(w|\theta_j) + (1-\lambda_C)p^{(n)}(w|\theta_{j,i}))} \\
p(z_{d,C_i,j,w} = C) &= \frac{\lambda_C p^{(n)}(w|\theta_j)}{\lambda_C p^{(n)}(w|\theta_j) + (1-\lambda_C)p^{(n)}(w|\theta_{j,i})} \\
\pi_{d,j}^{(n+1)} &= \frac{\sum_{w \in V} c(w,d) p(z_{d,C_i,w} = j)}{\sum_{j'} \sum_{w \in V} c(w,d) p(z_{d,C_i,w} = j')} \\
p^{(n+1)}(w|\theta_j) &= \frac{\sum_{i=1}^{m} \sum_{d \in C_i} c(w,d)(1 - p(z_{d,C_i,w} = B)) p(z_{d,C_i,w} = j) p(z_{d,C_i,j,w} = C)}{\sum_{w' \in V} \sum_{i=1}^{m} \sum_{d \in C_i} c(w',d)(1 - p(z_{d,C_i,w'} = B)) p(z_{d,C_i,w'} = j) p(z_{d,C_i,j,w'} = C)} \\
p^{(n+1)}(w|\theta_{j,i}) &= \frac{\sum_{i=1}^{m} \sum_{d \in C_i} c(w,d)(1 - p(z_{d,C_i,w} = B)) p(z_{d,C_i,w} = j)(1 - p(z_{d,C_i,j,w} = C))}{\sum_{w' \in V} \sum_{i=1}^{m} \sum_{d \in C_i} c(w',d)(1 - p(z_{d,C_i,w'} = B)) p(z_{d,C_i,w'} = j)(1 - p(z_{d,C_i,j,w'} = C))}
\end{aligned}
$$

Figure 3: EM updating formulas for the cross-collection mixture model