Progress Report

Ed Pureza (epureza2@illinois.edu)
Dansi Qian (dansiq2@illinois.edu)
Joe Everton (everton2@illinois.edu)

Change of Scope

We initially planned to reproduce Latent Aspect Rating Analysis without Aspect Keyword Supervision. We read through the paper, discussed among ourselves, and documented our understandings here. We had an hour-long conversation with Prof. ChengXiang Zhai (one of the authors of the paper) and email correspondence with Prof. Hongning Want (the main author), and eventually decided that none of us had the substantial math background required to reproduce the paper. Instead, we decided to reproduce Latent Dirichlet Allocation (https://dl.acm.org/doi/pdf/10.5555/944919.944937 and http://times.cs.uiuc.edu/course/510f18/notes/lda-survey.pdf), which was referenced by the original paper and also described briefly in week 9 of the course.

Which tasks have we completed?

- We have read both papers for Latent Dirichlet Allocation (LDA) and documented our understanding here. There is no analytical solution to the E-M algorithm in LDA. Instead, the optimization can be done using variational inference or Gibbs Sampling.
- We have implemented LDA using Gibbs Sampling (an initial version without learning rate, and a subsequent version that applies a learning rate across iterations).
- We have implemented pre-processing for text-classification datasets, the LDA model training on term frequencies, the inference of new documents using trained model (note that LDA is generative), and compared the classification accuracy of Support Vector Machines (SVM) using term frequencies and using topic weights. There were no significant differences in accuracy between the two for the text message spam filter dataset (short documents) or for the fake news detection dataset (long documents).

Which tasks are pending?

- We are working on the variational inference version of LDA (https://github.com/purecod3/CourseProject/blob/main/lda_var_inf.py).
- We plan to compare the accuracy of SVM classification using the variational inference version of LDA with the baseline (SVM using term frequencies).

What challenges do we face?

- There do not seem to be canonical ways to implement either the variational inference or the Gibbs sampling for LDA. There are some resources on implementation details but they are not necessarily correct or perform well. As a result, we need to compare multiple resources based on our understanding of the algorithm, and potential implement multiple versions and compare their results.
- There is a significant amount of hyperparameter tuning (number of topics, learning rate (and decay) for Gibbs Sampling, stop criterion definition and threshold for the variational inference, etc.). There is very little literature about how the hyperparameters should be set or tuned for LDA in practice so it would require trail and errors on our side.