CS410 Fall 2020 Course Project Progress Report

Author: Thiago Seuaciuc-Osorio
E-mail: thiagos2@illinois.edu
NetID: thiagos2

Team: Individual

Programming Language: Python

Topic: Reproducing the following paper on *Latent Aspect Rating Analysis*
Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In Proceedings of ACM KDD 2011, pp. 618-626. DOI=10.1145/2020408.2020505

Completed Tasks

The following initial tasks have been completed:

- Reviewed paper, identified tasks and planned approach.
- Obtained both relevant datasets (hotel and MP3 reviews).
- Extracted and reviewed both datasets to understand format and accompanying metadata.

Future Tasks

- Fully ingest data and implement same pre-processing steps listed in the paper. After this step, it is expected that the dataset statistics will resemble that reported on Table 1 in the subject paper.
  - ➢ *Expected timeframe*: complete by December 2
- Implement posterior inference and model estimation steps (sections 4.1 and 4.2 of subject paper).
  - ➢ *Expected timeframe:* complete by December 7
- Apply the topical Latent Aspect Rating Analysis Model (LARAM) to both datasets (hotel and MP3 reviews) and calculate selected relevant metrics for assessment. Only the experimental results related to the subject method (LARAM) will be reproduced; experimental results obtained with other methods for comparison and reported in the paper (such as LDA, sLDA, LRR, etc) will <u>not</u> be reproduced. Specifically, the following results will be reproduced:
  - The LARAM results shown in table 2
  - The LARAM results shown in table 4
  - The LARAM results shown in figure 3

  Two sets of results will not be reproduced:

  - The results in table 3, since they depend on “ground truth” generated by a LDA model.
  - The results in table 5, from the experiment where all reviews are concatenated and which was performed only to allow comparison with a Bootstrap+LRR model, since it is a degeneration of the more general case and similar metrics are already computed as part of table 4.

  ➢ *Expected timeframe:* complete by December 10

- Prepare code documentation and presentation.
    - *Expected timeframe:* complete by December 14

Challenges

Following the initial review and planning for the project, the following challenges were encountered:

- Table 1 in the subject paper lists 2,232 different hotels reviewed. However, in reviewing the available dataset listed in the provided reference, review files for only 1,850 hotels were initially found. This does not pose a problem for the execution of the project, but it may lead to slightly different results than obtained in the subject paper.
- The model equations described in section 4 of the subject paper are not fully clear or complete, as some are just referenced in other sources. These sources have been retrieved, but further study and analysis are necessary to fully determine the equations defining the LARAM model implementation.