CS 591/791 – Fall 2012, Assignment 3

Analysis due at 10PM on Wednesday September 26

Your third assignment is to run a regression analysis on a large-scale e-commerce dataset. For this assignment, we're going to study TripAdvisor hotel reviews, in particular the review dataset collected for the paper "Latent Aspect Rating Analysis on Review Text Data: a Rating Regression Approach", by H. Wang, Y. Lu, and C. Zhai, in KDD '10. You can read the paper if you're curious, but for this assignment, we're mostly interested in the dataset.

Public Sage notebooks have been a little flaky lately, but I'd still recommend doing this with Sage (which is easy to install on your own laptop or desktop).

Problem Statement: The authors of the paper above have collected about 108,891 reviews for 1,850 hotels, and have processed the reviews into data files that provide data in summary form. First download the data from `http://sifaka.cs.uiuc.edu/~wang296/Data/index.html`.

There is a lot of material and data here, but for the basic assignment, I want you to investigate how overall rating scores for the hotels relate to "aspect" scores, which are scores that TripAdvisor users can provide for features such as "Cleanliness". Which of these aspects play the most significant roles in determining the overall room score? To conduct this analysis, you'll have to take a look at the various README files and work with the `Vector_shLDA_1999.dat` datafile. I'd recommend a regression analysis.

Now, drill in a little deeper, along a direction of your own choice. For example, you could take a look at how term frequencies within the reviews are correlated with rating score. This need not be, but certainly could be, regression-based.

Submission: Submit a short (two page) writeup describing the results of your analysis by e-mail, by no later than 10PM on Wednesday. A summary table of the identified regression coefficients accompanied by text is sufficient for the first part. It's up to you to decide how to present your other findings, but be brief.

If you used Sage, please share your notebook with me as well. Click the Share button from your document and add my user name: "John_Byers".