CS 591 – Fall 2015, Assignment 3

Analysis due at 10PM on Wednesday, October 7

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. The Sage Math Cloud has been a little slow lately, but I'd still recommend doing your analysis there.

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 slide deck and/or Sage worksheet describing the results of your analysis by no later than 10PM on Wednesday. A regression formula plus summary table of the identified regression coefficients plus short text description is sufficient for the first part. It's up to you to decide how to present your other findings, but keep it brief.

If you used Sage, please share your Sage worksheet with me by making it publicly viewable in the settings and send me the link. Anything else can be e-mailed as an attachment. I'll select a few presentations to start class on Thursday.