**Capstone Project 2: Sentiment Analysis of the Yelp reviews**

|  |  |
| --- | --- |
| * What is the problem you want to solve? | * Sentiment analysis of the Yelp customer review * Prediction of the star rating based on the review * Recurring Themes * Try to Tag/Entities with sentiment |
| * Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise? | * Client could be all the e-commerce companies starting with Yelp |
| * What data are you using? How will you acquire the data? | DataSet:  1)Dataset is directly downloaded from  **https://www.yelp.com/dataset** |
| * Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution. | Simple Analysis of the dataset:   * Which type of data we are working- Understanding the data * The current loan status including paid/unpaid/as per credit policy etc. * We have missing values? Dropping features that have more than 30% missing data= fraction of data missing (Histogram) * Correcting the format of the reviews like removing extra spaces, symbols etc.   Pre-processing:   * Drop the feature if it is not useful. * View summary statistics * Explore the languages of the reviews * How many unique entries we have? * What's the distribution of the STARS allowed to the business * Relationship between length of review and the rating * Number of reviews v/s average star ratings * Most common businesses * Most common tag for a particular business/restaurent * And a lot of other questions that will raise through the exploration;   In depth Analysis:     * We will try to train our model as per positive and negative sentiment about the business based on the review.It would make sense to associate 4- and 5-star reviews with a positive sentiment and 1- and 2-star reviews with a negative sentiment. * 3-star reviews would be neutral, but for simplicity purposes, we will first attempt to predict the positive and negative sentiment, and we will revisit neutral after that. This is because our goal is to train a model to recognize positive or negative language, and 3-star reviews are likely to contain both. |
| * What are your deliverables? Typically, this includes code, a paper, or a slide deck. | Code, Project Report (pdf) |