**DRT IST718 Final Project Checkpoint 1 (DRT = Debasis, Rich, TJ)**

Online user reviews of products and services play an important part in the marketing, advertising and success of businesses today.

The goal of our project is to use the Yelp dataset from the Yelp Dataset Challenge to understand customer experience for a subset of businesses (tbd) and see if we can predict ways that the user/purchaser experience can be improved.

Our objectives are:

1. Integrate social data and business data to the base Yelp data
2. Determine whether we can predict star ratings from the corresponding textual reviews
3. Determine profitability of a business based on the ratings
4. Geographically visualize how demanding customers are

The data we will use is derived from various sources:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SubjectArea** | **SourcDataset Source** | **Records** | **Attributes** | **Obtainable** | **Notes** |
| Business | Yelp Challenge Dataset | 192609 | 13+ | Y | attribute counts varies per business |
| checkin | Yelp Challenge Dataset | 161950 | 2 | Y | Date attribute has multiple checkins for a business |
| photoix | Yelp Challenge Dataset | 200000 | 4 | Y |  |
| tip | Yelp Challenge Dataset | 1223094 | 5 | Y |  |
| review | Yelp Challenge Dataset | tbd | 8 | Y |  |
| user | Yelp Challenge Dataset | tbd |  | Y |  |
|  | Instagram | ? | ? | Maybe | integrate instagram reviews per business |
| business | DBHoovers | ? | ? | Maybe | integrate business information |
| tweets | Twitter | ? | ? | Maybe | integrate tweets per business |

Technically, we anticipate using a combination of Python and R programming for this work. Python for the heavier data processing and integration processes (likely from within Watson), and R for the geographic visualization (the feature set appears richer, as well as easier to use).

Our project team members (and roles) are:

* Debasis Chatterjee Geo-architect and visualizer
* Richard Paterson Storyboard and Developer
* Tajudeen Abdulazeez Architect and integrator

This project is important to our team as it will give us a deeper understanding of an online customer review process, and it will allow us to determine whether we can identify predictor variables that drive review responses.