Group Design Document

**Question:** Can our system predict the typical star rating of a particular type (ethnicity, hours, other attributes) of business given its location in Pittsburgh?

**Data to use**:

* Business: Attributes and star rating serve as ground truth? Group businesses by location as well to calculate average stars in a location to serve as a baseline against type.
* Review + user: Overall sentiment analysis, evaluated by actual star ratings as compared to that user’s average star rating. Attributes parsed from features within the text, evaluated against business’s labeled attributes.
* Tips: Sentiment analysis, evaluated using business’s stars.

**Expected pipeline (this is kind of up in the air)**

* 2 systems? Sentiment and attributes?
* A training example is 1 review or 1 tip. Need to group by location.
* For the review, features can either be sentiment (“excellent”, “bad”) or attribute (“slow”, “good décor”) features. Probably going to be some overlap here.
* Train our systems to output a star rating based on a location and an attribute.

**What has been done so far:**

* (David) Parsed JSONs to remove a lot of unneeded fields and make them CSV. Business file reduced to contain only Pittsburgh businesses.
* (David) Reduced the latitude and longitude fields to specific neighborhoods in business file using Google Maps API.
* (Bogdan) Review preprocessing. Sentence boundary detection.
* (Tyler) Feature selection and extraction from reviews. Mix of sentiment and features
* (Tyler) Parsed reviews file to only contain reviews on Pittsburgh-based businesses

**What needs to be done:**

* Potentially consider focusing on restaurants instead of all businesses? Would keep theme of original idea and further reduce data. So parse to get just restaurant data. But may be reducing too much…
* (Xinhai) Sentiment analysis of reviews + tips. Use review stars as gold standard.
* Decide whether we want to map attributes from business and reviews. The attributes field for business likely doesn’t quite match up to the features we’ll see in reviews, but I think we might need to somehow treat it as a gold standard to evaluate that subsystem anyway.