**Introduction:**

Our analysis focuses on all pubs that have full bar in Wisconsin. Among these pubs our specific goals are to analyze what could be the top 5 important services that affect their ratings. We try to analyze both the review text and tip text from Yelp and the stars that customers rated on Yelp.

We hope to answer the following questions:

1. Will the atmosphere, music, light and temperature affect their ratings? If so, what kind of atmosphere are good for their business?
2. Do different kinds of alcohol drinks and brands of drinks affect rating? If so, what kind of drink or brand of drink do customer prefer?
3. Will opening hour influence their rating?
4. What extra services could be beneficial to pubs rating? For example, business-parking, valet, garage?

**Background Information:**

1. Discuss any relevant aspects about the data
2. Process of data cleaning
   1. Using R to handle the following cases
      * n’t to not; ‘d to would
      * he’s to he is, she’s to she is , i’m to i am，it’s to it is
      * change to lower case

output is review\_text.json file

* 1. using R to filter all pubs in Wisconsin by the following steps
     + business json file, filter with attribute alcohol==”full\_bar”and is.open==TRUE
     + retain all the reviews of pubs
     + replace reviews json file’s text with cleaned review\_text
     + retain all the tips of pubs
     + retain all the users of pubs
  2. tokenization
     + using tidyverse and tidytext library to tokenize review\_text
     + get the most frequent nouns and abjectives in review\_text

select the top 100 or top 0.01 nouns and adjectives

* 1. repeat steps 1~3 to tips text
  2. Sentiment analysis to create some new predictors
     + Analysis the sentiment of a review text, positive, neutral, or negative
     + Analysis the sentiment of a tip text, positive, neutral, or negative

**EDA/ statistics tests:**

* + - EDA plots of nouns, and adjectives
    - see if nouns like atmosphere are in top 20/30
    - see if abjectives like friendly, fresh, busy are in top 20/30
    - Tests to draw some conclusions that can answer our questions

**Key findings about business**

A multiple regression:

Output: stars

Explanatory variables: some key attributes of business(discrete) + opening hours + average grade of reviews’sentiment

**Recommendations for business:**

According to our analysis above.

**Limitations:**

Since we use bing, nrc, Afinn sentiment lexicons which are unigram based, our sentiment analysis cannot identify sarcasm and negated text.

A text the size of many paragraphs can often have positive and negative sentiment averaged out to about zero, while sentence-sized or paragraph-sized text often works better.