STAT 511 Final Project (Initial Sketch)

Jim Curro, Eric Hare, Alex Shum

Apr. 15, 2013

Introduction

This is a sketch of our results thus far. We have made progress on our two biggest goals with the dataset. First, we've conducted an EDA on the data and discovered some interesting things, which is summarised in section two. Second, we've done some preprocessing on the review text data and have built a term document matrix. We are working on training a classifier using the term document matrix but are having a few issues with the size of the dataset. A summary of our progress on building a classifier is in section three.

Exploratory Data Analysis

Figure X displays the number of useful votes against the number of cool votes in blue, and the number of useful votes against the number of funny votes in red. It can be seen that while both cool and funny votes are great predictors of useful votes, the slope is larger for funny. In other words, we would expect that the reviews with a set number of funny votes would tend to have more useful votes than those with the same set number of cool votes.

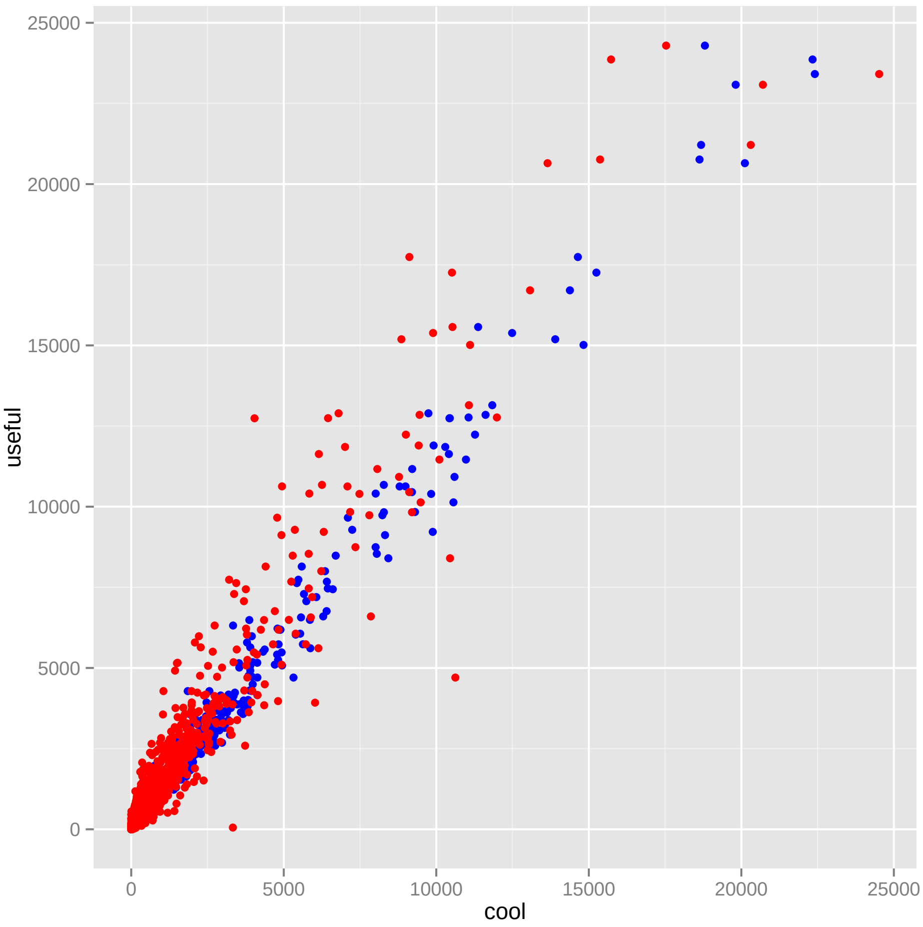

Table X shows the top 10 cities by number of checkins in the database. It also shows the total number of reviews for all businesses in that city, as well as the percentage of reviews over checkins. It can be seen that for the biggest cities in the database, there is a very uniform number of about .27 reviews per checkin.

city	reviews	checkins	percentage
Phoenix	91453.00	310640.00	0.29
Scottsdale	49042.00	202946.00	0.24
Tempe	27021.00	93446.00	0.29
Chandler	13944.00	48067.00	0.29
Mesa	9786.00	34328.00	0.29
Glendale	7017.00	26306.00	0.27
Gilbert	5803.00	22723.00	0.26
Peoria	2599.00	11361.00	0.23
Surprise	1274.00	5109.00	0.25
Avondale	1243.00	5053.00	0.25

Figure X shows boxplots of the number of checkins to each business by the average star rating of that business. It can be observed that businesses with about a four star rating tend to have the most checkins, but five star rating businesses actually have fewer. This may be either due to the fact that businesses with such a high rating have fewer total reviews, or they are more expensive and less likely to have a high number of customers.

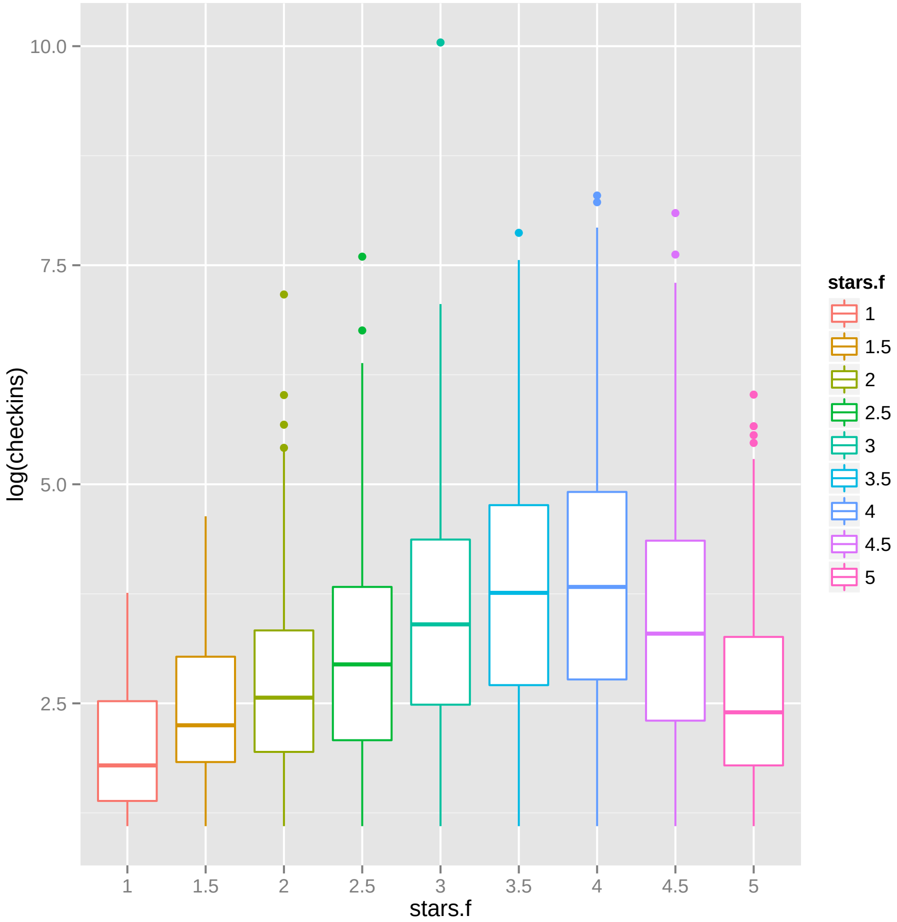

Table X displays the top ten users by the total number of useful votes divided by the total number of reviews for that user. We have selected only users who have made at least 100 reviews. It can be seen that...

Preprocessing and Classification

The review data provided by Yelp was provided without much preprocessing. Starting with the corpus we removed excess whitespace, put all words into lower case, removed all punctuation, removed numbers and removed stop words (stop words are common

funny	useful	cool	name	average_stars	review_count	good
20707.00	23080.00	19815.00	Hazel	4.06	814.00	28.35
13074.00	16707.00	14381.00	Katie	4.19	770.00	21.70
11070.00	13146.00	11836.00	Chris	3.77	624.00	21.07
9488.00	10134.00	10563.00	Robin	3.79	530.00	19.12
7481.00	10396.00	9832.00	Raider	4.24	546.00	19.04
1993.00	2586.00	2600.00	Clyde	3.90	143.00	18.08
5391.00	6063.00	5539.00	Marlon	3.92	346.00	17.52
2544.00	2631.00	2638.00	William	3.97	153.00	17.20
10451.00	8401.00	8429.00	Thomas	3.67	517.00	16.25
9003.00	12233.00	11270.00	Jennifer	4.08	764.00	16.01

words that do not provide much meaning such as ’the’, ’is’ and ’or’). Our final step of preprocessing is to stem the corpus (reducing words to their base form). Unfortunately, the stemming algorithm used in the TM package has some external dependencies on Weka and Java. For our term document matrix we only included words that appear in at least 10 reviews. The resulting matrix is a 27826 by 229907 sparse matrix. The first 10 x 10 block of the matrix can be seen below in Table X.