Predicting Strengths and Weaknesses of a Business from Yelp Data

Stephen Dimig

November 12, 2015

# Title

Predicting Strengths and Weaknesses of a Business from Yelp Data

## Introduction

This report focuses on using samples of of review data from the Yelp Dataset Challenge data set to efficiently predict the strengths and weaknesses of a business. Text mining encompasses a vast field of theoretical approaches and methods with one thing in common: text as input information [1]. The tm package in R provides a framework that allows text mining and natural language processing techniques to be applied to datasets in a convenient way. We will leverage the tm package to analyze review data from the Yelp Dataset Challenge data set to determine the part of speech of each word in a review and determine the most frequently used nouns associated with positive reviews of a business (defined as strengths) and negative reviews of a business (defined as weaknesses).

### Glosary of Terms

|  |  |
| --- | --- |
| **Term** | **Definition** |
| tm | A text mining package for R that supports natural language processing of texts |
| NLP | Natural Language Processing |
| Favorable Review | A review with a rating of 3.5 or above |
| Unfavorable Review | A review with a rating of 2.5 or below |
| Similarity Score | The number of words in the top 25 strengths or weaknesses in the sampled data that are also present in the full data |
| Stopword | A commonly used word that is filtered out before processing review text |
| Strength | A high frequency noun in the favorable reviews of a business that is not a stopword |
| Weakness | A high frequency noun in the unfavorable reviews of a business that is not a stopword |
| Popular | A user that reviews a business that has many friends |
| Connected | A user that reviews a business that also has many friends that reviewed the same business |
| Maven | A user that reviewed a business that also has many friends that reviewed the same business AND those friends gave that business the exact same star rating |
| Boosted | An attempt was made to boost the similarity scores by predicting which sampling technique was best for the given business. |
| Random | Randomly selected reviews about a business used as a baseline |

So, why not just crunch all of the data (or a high percentage of it) to determine the strengths and weaknesses? Why try to predict or sample at all? As wonderful as the tm package is, it is very slow. Examining a large number of reviews to find the strengths of a business can take a very long time. A sample size of up to 500 reviews was used in this analysis to determine the true strengths and weaknesses and processing it could take up to an hour. The goal here is to determine a sampling technique tha can perform better than random so that you can pick a smaller number of samples (75 was used) to make a reasonably accurate prediction. The top 25 strengths and weaknesses idetified from the full data were then compared against the top 25 strengths and weaknesses identified by the sampled data to determine the similarity.

|  |  |  |
| --- | --- | --- |
| **Scenario** | **Size** | **Time (in s)** |
| Strengths - Full | 500 | 616.56 |
| Strengths - Sample | 75 | 18.3 |

## Methods and Data

### Examined Sampling Techniques

A **Maven** is a trusted expert in a particular field, who seeks to pass knowledge on to others. The word maven comes from Hebrew, meaning "one who understands", based on an accumulation of knowledge. I was introduced to the concept of a maven through Malcolm Gladwell's book "The Tipping Point". To find the mavens in this data we examined the friends of a user that reviewed a business to see if they gave the business the exact same star rating. The potential maven user got a point for each friend that gave the same rating. The mavens were then sorted and the top 75 reviews were used for sampling.

A **Connected** user is similar to a maven, but they are scored only on the number of friends that also reviewed the same business with no weight given to the star rating. The connected users were sorted by number of connections and the top 75 reviews were used for sampling.

A **Popular** user is one who reviewed a business that has a large number of friends. The friends may or may not have also reviewed the business. The popular users were sorted by number of friends and the top 75 reviews were used for sampling.

A **Boosted** sample tries to predict the technique that will yield the best results using machine learning techniques and apply the best one to the business.

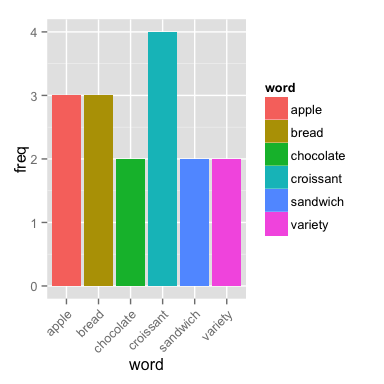
For control purposes, a **Random** sample of 75 reviews of the business taken was also used as a baseline.

### Use of tm Package for Part of Speech Tagging

Strengths and weaknesses are extracted from the review text in the following manner:

* All of the review texts are concatenated and collapsed into a single text
* All characters are converted to lower case
* Common english stop words are removed
* The category of the business is removed to keep it from showing up as a strength or weakness (ie; reviews of a Taco joint will mention Taco a lot)
* Each word is tagged with a part of speech
* All non-nouns are removed
* The word list is sorted in descending order by frequency

The result looks like this graphically.



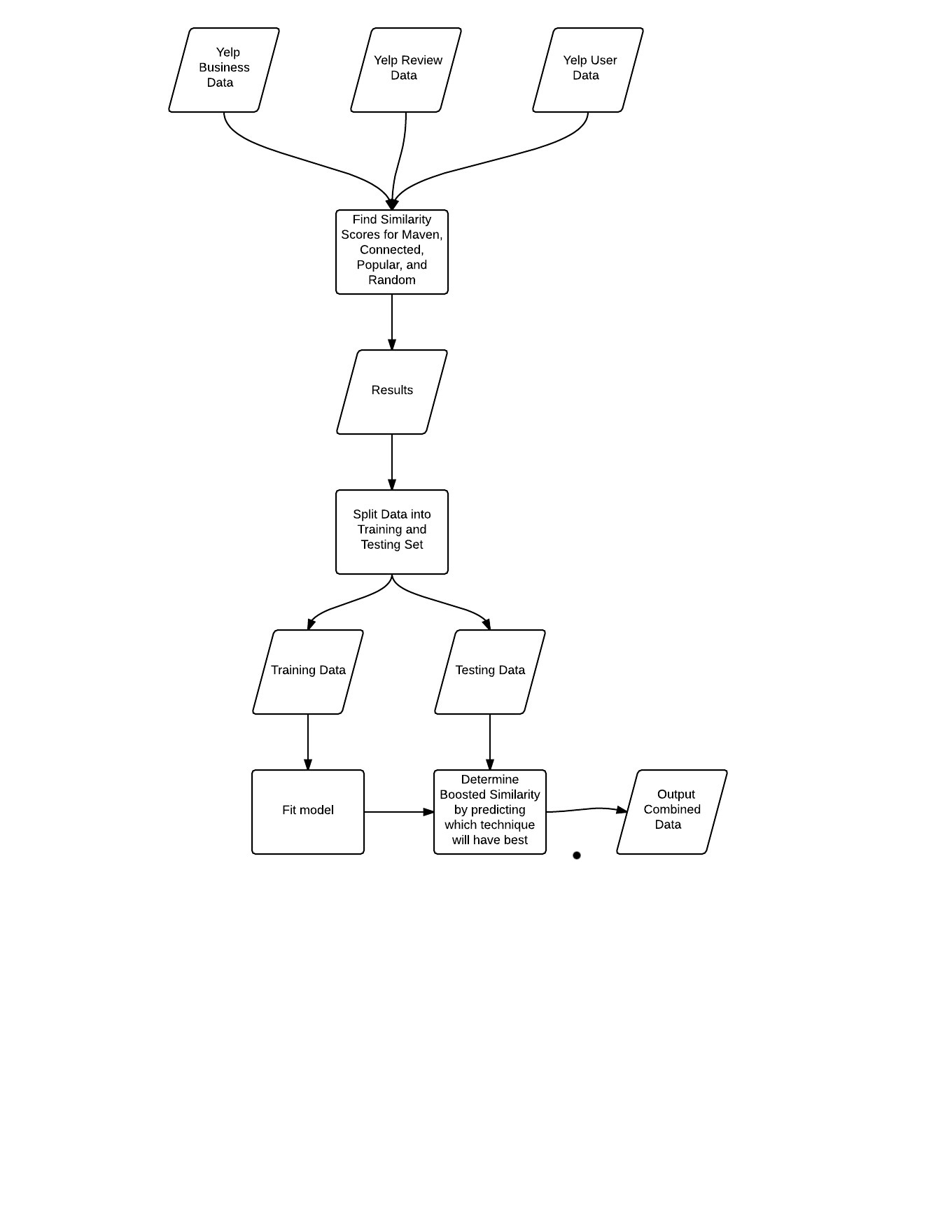
### Similarity Scores

The similarity of a full run and an sampled run was determined very simply by determining how many nouns from the top 25 of the sampled text are in the top 25 words of the full text.

similarity <- function(x, y) {  
 count <- 0  
 names\_x <- names(x)  
 count <- sum(sapply(names(y), function(z) ifelse(z %in% names\_x, 1, 0)))  
 100 \* (count / length(y))  
}

### Basic Processing

The following procedure was used to test how well a sampling method performed. Random sampling is the control group.



### Modeling for Boosted Sampling

Boosted sampling is performed due to a strange anamoly in the data where certain sampling techniques work well on certain data but perform only as well or below random sampling as a whole.

|  |  |  |
| --- | --- | --- |
| **Technique** | **Rating** | **Mean** |
| Maven - Strength | >= 3.5 | 77.2222222 |
| Conn - Strength | >= 3.5 | 75.8888889 |
| Pop - Strength | >= 3.5 | 70.4444444 |
| Random - Strength | >= 3.5 | 72.2222222 |
| Maven - Strength | All | 71.0322581 |
| Conn - Strength | All | 72.3870968 |
| Pop - Strength | All | 68.4516129 |
| Random - Strength | All | 70.2580645 |
| Maven - Strength | <= 3.0 | 62.4615385 |
| Conn - Strength | <= 3.0 | 67.5384615 |
| Pop - Strength | <= 3.0 | 65.6923077 |
| Random - Strength | <= 3.0 | 67.5384615 |

As you can see the maven technique performs much better than random when the overall when the rating is over 3.5, but not any better overall, and worse when the start rating is less than 3.0. The following is an example of how the modeling was done using the RandomForest method.

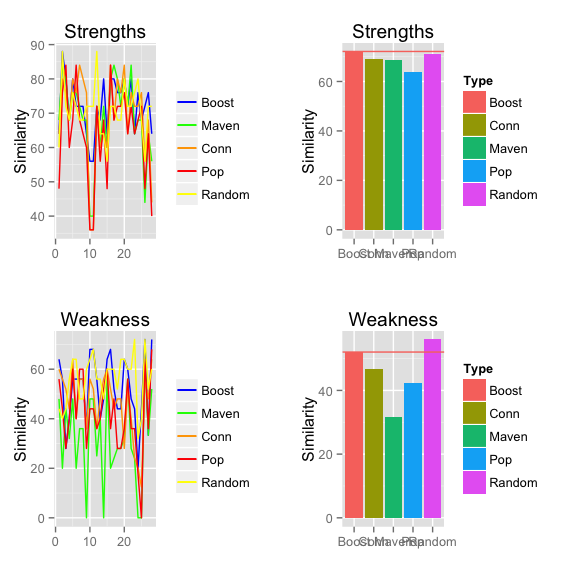
create\_models <- function() {   
 trc <- trainControl(method="cv",number=5)  
 training <- read.csv("training.csv", stringsAsFactors=FALSE)  
 good.mavenFit <- train(good.maven ~ rating + review.count +   
 good.review.count +   
 good.maven.count +   
 good.conn.count,  
 data=training,method="rf",  
 trc,  
 prox=TRUE,allowParallel=TRUE)  
   
  
   
 bad.mavenFit <- train(bad.maven ~ rating + review.count +   
 bad.review.count +   
 bad.maven.count +   
 bad.conn.count,  
 data=training,method="rf",  
 trc,  
 prox=TRUE,allowParallel=TRUE)  
   
 # Models are also trained for Connected, Popular, and Random   
 # Sampling as well but not shown here.  
}

## Results

Dissappointingly, the results show that no sampling techniques performs significantly better than random sampling over a representative mix of businesses. This is in spite of the fact that both the Maven and Connectected techniques perform significantly better on businesses that were rated at 3.5 or above.

|  |  |  |
| --- | --- | --- |
| **Technique** | **Scenario** | **Mean** |
| Boosted | Strength | 72.1428571 |
| Maven | Strength | 68.5714286 |
| Conn | Strength | 69 |
| Pop | Strength | 64 |
| Random | Strength | 71 |
| Boosted | Weakness | 52.047619 |
| Maven | Weakness | 31.797619 |
| Conn | Weakness | 46.7321429 |
| Pop | Weakness | 42.2016807 |
| Random | Weakness | 56 |

The Boosted technique definitely shows promise and I think it could be tuned to outperform random sampling by a higher margin for strengths. The weaknesses are more variable, but even there boosting shows proise.



## Discussion

If I had to do this project over again, I would batch process the full data set with something like Hadoop to determine the top strengths and weaknesses of each business. R did not perform well performance wise on the computationally expensive data mining tasks. I had to run a task for over a day just to get the results for 50 businesses. This left ar less time to explore and tune the different sampling techniques. The results were mixed due to this even though the Maven and Connectected techniques showed great promise on certain subsets of the data. The Boosted approach could also be tuned slightly better with more runs and more features so that it could better predict which technique to use. Even though the results were not what I expected, the challenge of working on the data set was fun and I can definitely use the experince on future projects.