@Starbucks

Text Analysis to support Starbucks' Twitter Campaign

Kamakshi Nittala

8/28/2016

Introduction

Client: Starbucks, excellent social media strategy

Focus of this project: Twitter

Hypothesis: SB's twitter popularity is driven by it content

Important Questions:

- What are the types of topics covered on SB's twitter channel?
- What are the common characteristics of these latent topics?
- Do they trigger an information spread (In this case- retweet count)?

The Dataset

Extracted using the twitter API for R: twitteR

To read about the API and understand the twitter authentication process in detail:
https://dev.twitter.com/rest/public/search

Dataset contains 84090 observations and 16 variables

Data Preparation

Text mining: important aspect aspect of data preparation.

Text data highly unstructured.

Processing of data using R's plyr and dplyr packages

Resultant data contains 84090 observations and 6 variables

Topic Modelling & Exploratory Data Analysis

Start with some more text processing to transform the texts into structured formats to be actually computed with. Using tm package create a document-term matrix and text corpus for further processing

Word cloud:

“Coffee” and “starbucks” are the two most important words.

Terms like “macchiato”, “caramel”, “coconut” are not as exposed

Dominated by the presence of regular words like “morning”,

“milk” and “tea”.

Topic Modelling (contd..)

Latent Dirichlet Allocation: Facilitates the automatic discovery of themes in a collection of documents. LDA attempts to model how a document was 'generated' by assuming that a document is a mixture of different topics, and assuming that each word is 'generated' by one of the topics. Our model: 6 Topics (k = 6) , 5 Terms under each topic (the first terms under each latent topic), Total number of terms: 30

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
starbucks	coffee	coffee	milk	coffee	coffee
iced	starbucks	starbucks	iced	free	starbucks
coffee	mocha	time	coconut	tea	morning
card	time	store	going	thank	milk
caramel	ever	work	macchiato	drinks	iced

Topic Modelling (contd..)

Moving forward, in order to accurately analyze the popularity quotient of each of these terms, it is important for us to first name each term

Topic 1	CaramelFix
Topic 2	MochaMania
Topic 3	StoreTime
Topic 4	CoconutEspresso
Topic 5	FreeDrinks
Topic 6	MorningMix

Feature Engineering

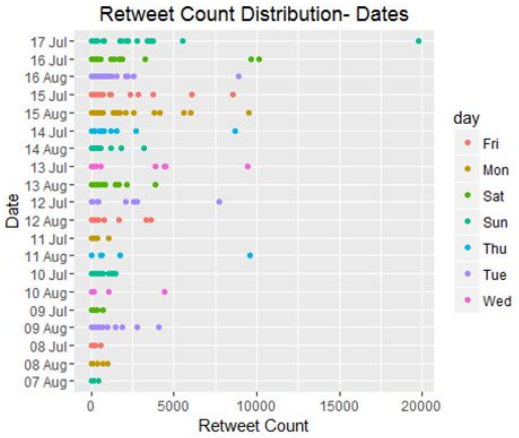
One of the main limitations in twitter analytics is that each retweet is identified as a separate text. This will have an adverse effect on our regression analysis - so remove all the duplicate text. The resultant dataset contains 40,512 unique observations

Perform feature engineering on our dataset to create a statistical model from it. Meaning, our dataset has 40,512 unique observations or texts and each text will be a mixture of each of the thirty terms- in order to accurately determine that, we add binary features for each term and insert those binary features as separate columns. Additionally we parse the 'created' column to form 11 extra columns. This feature helps us to identify if there a particular day, or time in the day that the featured terms attract traffic. The resultant dataset has 47 new columns with 41 new features.

Exploratory Data Analysis- 1

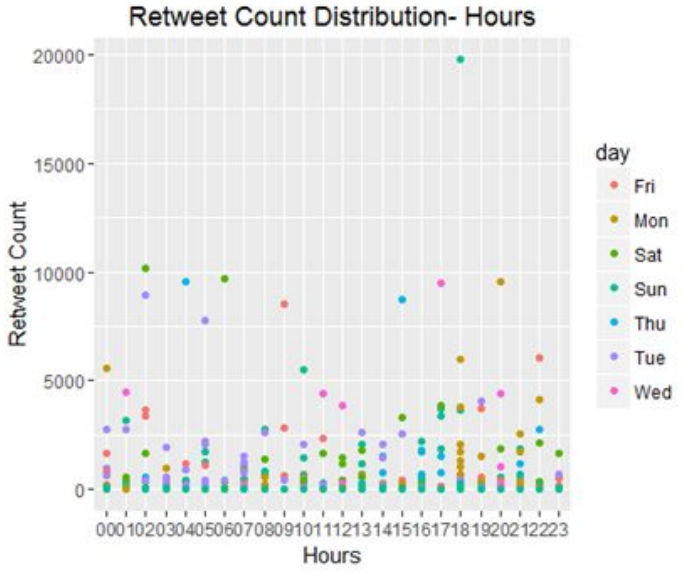
The plot gives us an overview of the dates where the retweetCount is highest - 7/17/2016. A quick look at the data to identify the text, shows this: *"Ordered my drink @Starbucks Asked the barista if she wanted my name. She winked and said. "We gotcha" #JodieFoster"*

This is a very interesting observation as it creates a possibility for us to develop this analysis towards social network analysis. The plot is also successful is giving us the comparable data between the same dates in each month and it can be observed that a large number of tweets have a retweet count of 0 to less than 5000 followed by the range of 5000-1000



Exploratory Data Analysis- 2

It is also important to know the time at which there was maximum information spread and for that, we create a similar plot using the 'hour' attribute. If we exclude the one outlier, we can notice that most of the information spread took place either in the ***Latenight-Earlymorning*** or **Evening** periods.

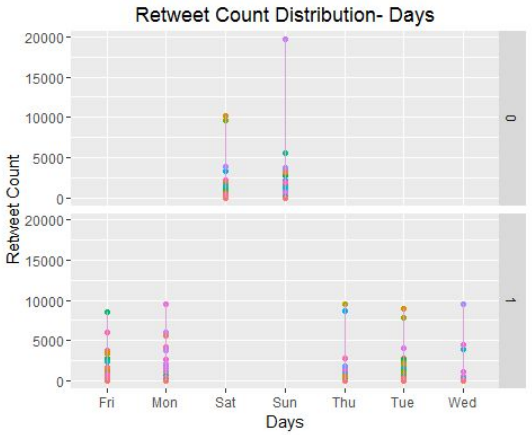


Exploratory Data Analysis- 3

Another interesting visualization could be the information spread between weekdays and weekends.
While the plot shows the numbers, we can a quick calculation to know the total number of retweet counts for weekdays and weekends to generally compare the two.

	wday	totalretweetCount	num
1:	0	126427	11609
2:	1	230985	28903

This means that we have 28903 weekday observations with a total retweet count of 230985 and 11609 weekend observations with a total of 126427. The data for wend is noticeably biased due to the one outlier with a retweet count of '19774'. Subtracting that from the original sum gives us '106653'



Linear Regression

Idea: To know if the 'retweetCount' has any kind of relationship with our new binary features or any other relevant attribute in the dataset representing content and time.

We will have to split our data into a 80:20 sample (training:test), then, build the model on the 80% sample and use the model thus built to predict the dependent variable on test data.

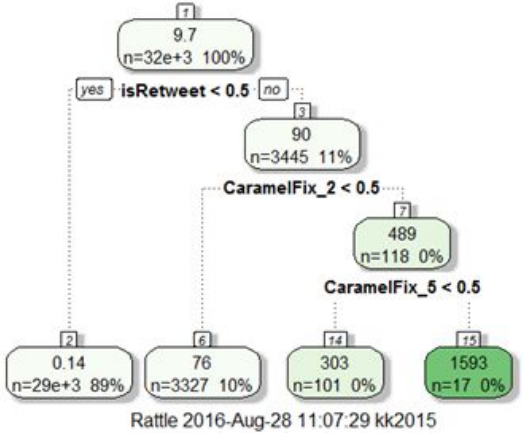
Adjusted R-squared on Training Data: 0.01485: Predicted R-squared on Test Data: 0.02965

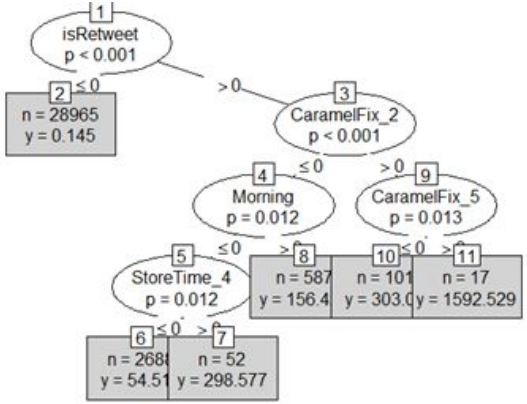
Since our r-squared value is closer to 0% we can infer that our model does not clearly explain the variability of the response data.

Trees (with rpart and CART package)

Given the deficiency of our linear model, it is highly unlikely that we can improve the outcome with single regression trees. Two widely used implementations for single regression trees in R are **rpart** and **party**. The rpart package makes splits based on the CART methodology using the rpart function, whereas the party makes splits based on the conditional inference framework using the ctree function. Both rpart and ctree functions use the formula method.

Predicted R-squared with rpart package: 0.150209. Predicted R-squared with party package: 0.1396499



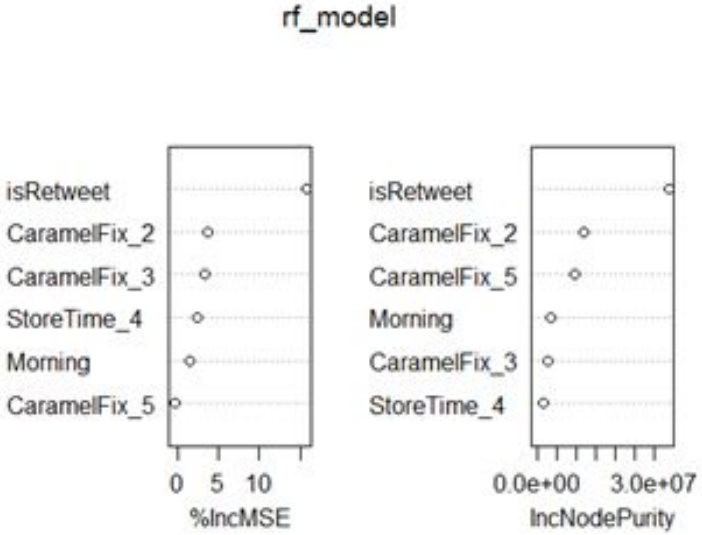


Random Forest

Designed to improve the prediction accuracy of CART methodology and works by building a large number of CART trees. Unfortunately this makes the method less interpretable.

Predicted R-squared: 0.1237538.

Dot chart of variable importance as measured by Random Forest:



Conclusion

In this analysis, we looked at a very significant feature of text analysis: Topic Modelling. We started with data extraction using the twitter API for R followed by simple data tidying techniques and a twitter-LDA model designed for short texts. We settled with in-depth analysis of data using linear regression, single trees and random forest.

While the hypothesis was interesting, given the nature of our text data, our regression and tree models were deficient and did not help us prove our hypothesis.

In conclusion, our analysis revealed some interesting findings.

While twitter can be a good source of entity-oriented topics that have low coverage on traditional news media, we find that such content may not be solely responsible for Starbucks' Twitter stardom.

This very interestingly directs us towards the argument that it is not just content but also distribution that matters for brand popularity and, by this means, we can contemplate towards the less popular research approach of "content-based social network analysis" for Starbucks.

Alternatively, keeping this analysis as the basis we can compare the influence of Starbucks' twitter content with other social or traditional news media to know if the distributions of topic categories and types differ in Twitter and in other media.

Limitations and Future Work

Our text data was highly unstructured and did not always comply with the rules of topic modelling and regression analysis.

This brings us to the second limitation. In the extracted data each retweet is identified as a separate text. This had an adverse effect on our regression analysis. To avoid this, we had to remove all the duplicate text and this in turn resulted in a poor regression model.

In order to accurately determine the information spread, it is important for us to have more data and additional attributes such as total followers/friends, geolocations for the different screen names in the dataset. In addition to the memory issues, we had challenges with the
Twitter API's inconsistency, due to which we were unable to collect this more potentially useful information.

Owing to the availability of more data, the learned topic models can definitely complement a more complicated social network analysis.
Our initial results show that topic models are able to obtain meaningful categories from unsupervised data and show promise in revealing network-like statistics to support Starbucks' social media popularity.

Thank you!