Data Science Capstone Project Report

Gert De Geyter

Sunday, November 22, 2015

Title

This is the final report written for the Capstone project for the Coursera Data Science Specialization. In this project we use the data made public for the Yelp Dataset Challenge. The main focus of this report is based on identifying and clustering groups of people with similar interests in food and environment. Additionally we will also investigate if some overlap exists between different groups (for example a "vegan" and "hispters" group could partially overlap).

Introduction

In this project we try to answer the question if there are clearly separated groups of people and their preferences of both dietary restrictions and the environment. For example a related question would be "Do vegetarians have a preferences for hipster places?". The reason why this question could be interesting is for better customer profiling. For example, in the above case vegetarian restaurants could target potential customers more accurately, increasing their publicity with minimal cost. Not only would they be able to address new clients, they could also better suit the needs and wishes of the existing customers. In this way they could build up the returning customer base which is vital for any restaurant to produce steady cash flows and remain open for business.

Data merging and aggregating

To answer these questions we use a combination of three different data sets from the Yelp Dataset Challenge:

1. yelp_academic_dataset_user.json
2. yelp_academic_dataset_review.json
3. yelp_academic_dataset_business.json

Figure 1 shows a schematic overview of how we link the different data sets in order to get all of the necessary data combined before we move to the analysis. As this project is specifically focused on food, we first narrow down to the reviews for business who are clearly based on preparing food (and not for example specific food shops). From the different categories of businesses we have manually selected all relevant ones to narrow down the selection. As this data set includes the environment (essentially the 'Ambience') and 'Dietary Restrictions' it is important to take this information and link it to a specific user.

The next step consists of merging the reviews with the food selected business. We need this merging for two reasons:

1. Not all the reviews were written on a business we are interested in
2. The reviews contain the link between businessID and userID

Once we have merged the business and reviews data set we get rid of all the information we don't need for our analysis, like the text of the review it self. Doing so we create lighter data frame which should be easier to handle in the upcoming steps. The next thing we have to do is then aggregate this frame to have one

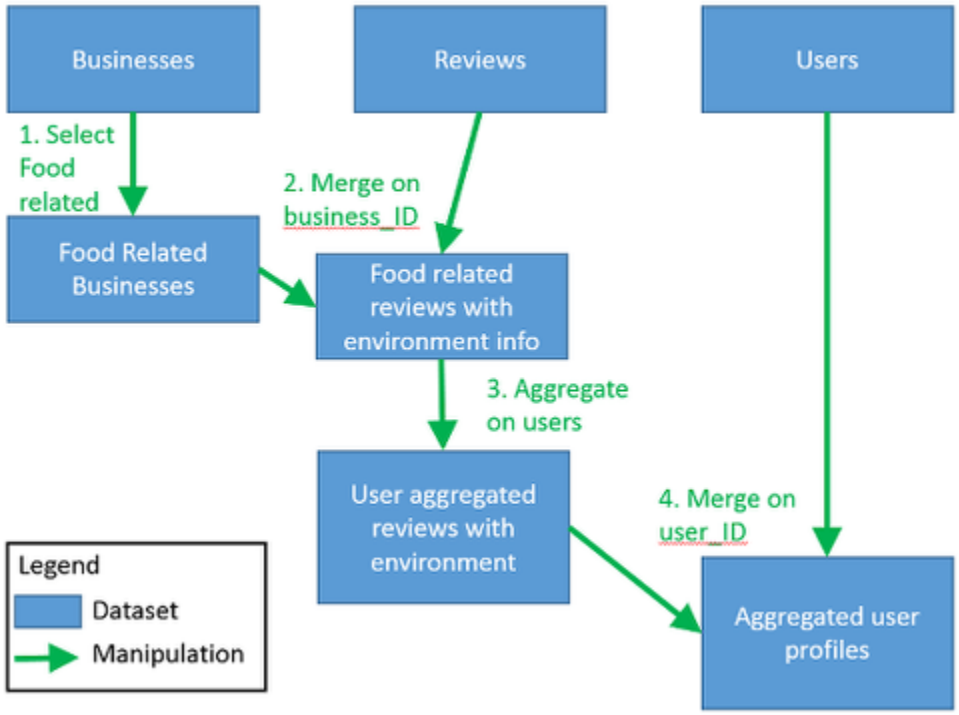

Figure 1: Schematic overview of merging and linking of datasets

line per unique user as this data frame still contains multiple lines per review. However, what we need in the end is a specific 'user-profile' so therefore we need to combine the several reviews into one. The final step remaining is to merge the newly created dataframe of relevant reviews with the dataset containing the users information. This can be easily done by merging using the userid. Each of these steps is explained in more detail in the next chapter. Once the final frame is created we check if certain ambience and dietary restrictions have tendancy to occur more or less frequently with others. However before we can do this it is important to make the necessary corrections on numbers as not all of them have the same occurancy in the final data set. This will be discussed in the 'Results' chapter. # Methods and Data In this section we dig a bit deeper into what manipulations we have to done reach to the 'user-profiles'. To keep this report light and within 5 page limit not all steps are included here. However, if not all steps are perfectly clear are the reader is interested in finding out specific details, feel free to look at the code posted on my GitHub.

We first start by loading the necessary packages, reading in the data files and saving it so it can be loaded faster afterwards.

```
if(!require('BBmisc')){
    install.packages('BBmisc',dep=TRUE)
}
# load packages
suppressPackageStartupMessages(library('BBmisc'))
pkgs <- c('jsonlite','plyr','plyr','stringr','doParallel','ff','ffbase',
          'corrplot','rjson','Hmisc')
suppressAll(lib(pkgs)); rm(pkgs)
registerDoParallel(cores=4)
# .. cut part of the code here. see GitHub
```

Next we make the first selection 'Select Food related' on the business file to select only the relevant food preparation business (i.e. any 'sort' of restaurant). The file 'FoodStuff.csv' contains a list of manually selected

relevant categories.

```
load(paste0(getwd(),'/Capstone_Quiz.RData'))
user <- dat[['user']]
bus <- dat[['business']]
reviews <- dat[['review']]
FoodCat<- read.csv("FoodStuff.csv",sep = ";", dec = ".")
foodcat <- as.vector(FoodCat$Cat)
isFoodRelated <- sapply(1:nrow(bus), function(x) sum(foodcat %in% unlist(bus$categories[x])) >0)
bus.selec <- bus[isFoodRelated,]
```

The second step, 'Merge on business_ID', starts by making the 'ambience' and dietary restriction columns from boolean to numeric. This will be important in step 3. We first only select the columns we need from the reviews file. Next, we filter on those reviews which were written on relevant businesses. Finally we merge the reviews with the businesses file using the businessID.

```
rev.sm <- reviews[, which(names(reviews) %in% c("user_id", "stars", "business_id"))]
busids <- rev.sm$business_id
pos.bus<- match(busids,bus.narrowed$buID)
userids <- rev.sm$user_id
df.positions <- data.frame(pos.bus = pos.bus, busID= busids, userids=userids)
df.pos.comp <- na.omit(df.positions)
bus.narrowed$busID <- bus.narrowed$buID
df.bu <- merge(df.pos.comp, bus.narrowed, by="busID")
```

The third step consists of aggregating the multiple reviews per user into one. This can be done by simply stacking the ambience and dietary restrictions after converting them to a numeric value. So for example, if a user has written two reviews on a business categorized as 'trendy', the rows in the data frame will consists of one user only but the column 'trendy' will now contain '2' instead of 'TRUE'. The 4 step is then just to merge this data frame with the user data frame using the userid.

```
df.bu2 <- df.bu[, -which(names(df.bu) %in% c("busID", "pos.bus","buID"))]
df.bu2 <- na.omit(df.bu2)
df.bu2$nrev <- rep(1,nrow(df.bu2))
df.bu3 <- ddply(df.bu2,"userids",numcolwise(sum))
names(df.bu3)[names(df.bu3) == 'userids'] <- 'user_id'
df.bu4 <- merge(user,df.bu3, by="user_id")
df.final <- df.bu4
```

Results

Now that we have created a final data frame containing all the information let's have a look at the frequency of all the different ambiences and dietary restrictions.

From Figure 2 one could assume that there are lot of vegetarians among the users. However, it is important to notice that we have created these profiles based on business traits. For example, a non-vegetarian could easily have a hamburger in a place that also serves vegetarian dishes but obviously that does not make him a vegetarian. Therefore, before we can draw any conclusions on which traits occur more (or less) frequently together we must weight this by the absolute occurence. This is shown in the following step. First we must correct the 'user-profile' by weighing the number of reviews he/she has written on a certain type of restaurant by the total number of reviews the user has written. Next, this is also correct by the total amount of businesses having this specific trait.

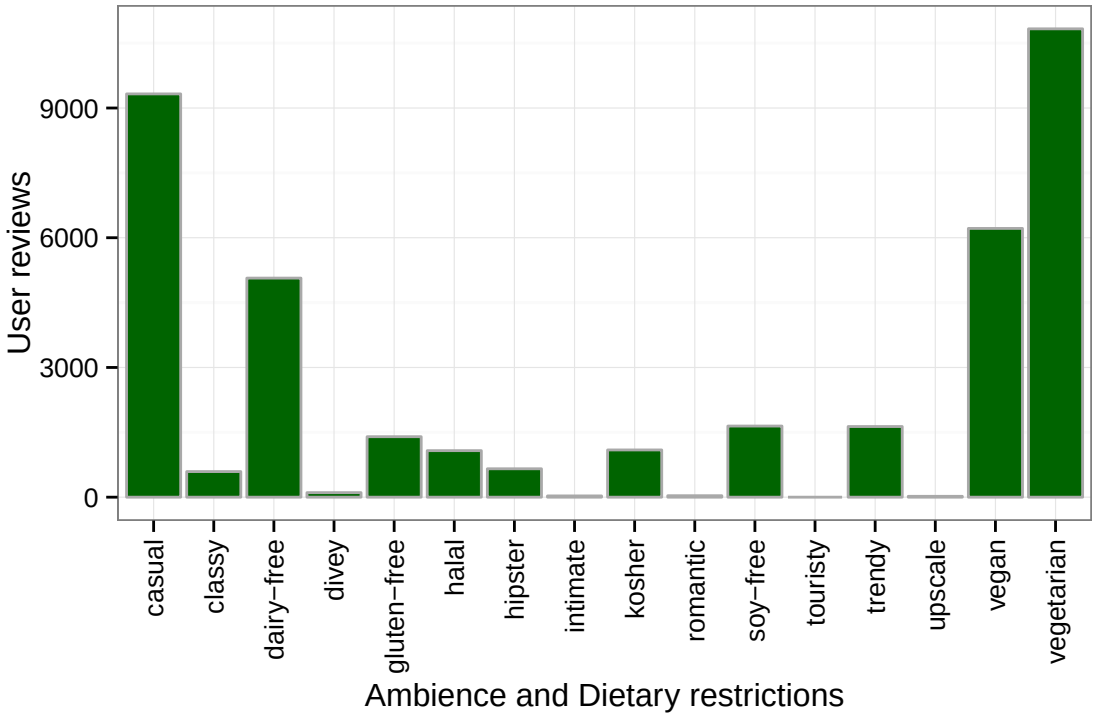

Figure 2: Frequency of Ambience and Dieterary restrictions among user profiles

```
for(i in 1:length(envDiet)){
    df.final[[envDiet[i]]] <- df.final[[envDiet[i]]]/df.final$nrev
    df.final[[envDiet[i]]] <- df.final[[envDiet[i]]]/sum(df.final[[envDiet[i]]])
}
```

Now that we have the corrected/weighted occurences we can look at how frequently they appear together. This can be done by calculating the correlation matrix among all traits. This matrix will indicate a value between -1 to +1 indicating how likely or unlikely two traits will pair up. This matrix can be seen in Figure 3 and will be discussed in the next chapter.

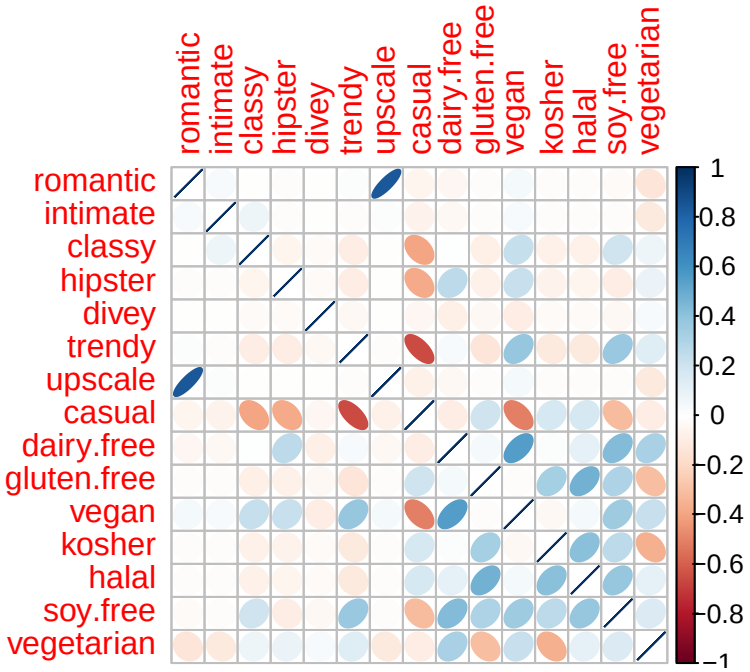

Figure 3: Weighted correlation among different ambiences and dieterary restrictions

Discussion

Before we begin this dicussion it is important to notice *Correlation is not causation*. Therefore, one should **ALWAYS** we very careful when drawing conclusions from them. Because of this, and because of the smaller scale of this project, all the conclusions here will be expressed very conservativaly and should also be regarded as such. As a final remark, we will speak of f.e. 'romantic users' while it is important to remember that these are always trait of businesses were users go. However we use this to keep the discussion light and not overly verbose.

When looking at Figure 3 we notice a few things:

1. Not suprisingly there is a high tendency for dietary restrictions to pair up. this is most likely partly due to an overlap in dietary restrictions (f.e. all vegans are vegetarians and dairy.free) and because of business specifically focussing more on several dietary restrictions.
2. Romantic users tend to be more upscale. This might not be so suprising as most people won't take their dates out on a cheap, fast-food restaurant.
3. Trendy people are more likely to be vegan and eat soy free.
4. Casual users are less likely to be classy, hispters or trendy. Interestingly they have slightly higher tendency towards eating gluten free while being less likely to be vegan or vegetarian

These are just a few first remarks that can be made. Because of the richness of this Yelp dataset, much more questions could be answered. We believe that this is just a potential start of investigating many tendencies among users and could easily be extendend. For example, not just looking at dietary restrictions but at specific cuisines like "Italian" and the business ambience. Having this information could be an important tool for some businesses to adjust more to their clients needs. For example, using conclusion 2 from this work, upscale restaurants could focus specifically on advertising romantic dinners for two. This is just a small example of the amount of information that could be locked in this data set alone so imagine what could be learned if it was enriched with other data like geospatial information of social network interests.