New features from the Address field

Wei Xu

April 12, 2017

```
suppressMessages(library("jsonlite"))
suppressMessages(library("dplyr"))
suppressMessages(library("tidyr"))
suppressMessages(library("plotly"))
suppressMessages(library("purrr"))
suppressMessages(library("RecordLinkage"))

lst.trainData <- fromJSON("../input/train.json")
vec.variables <- setdiff(names(lst.trainData), c("photos", "features"))
df.train <-map_at(lst.trainData, vec.variables, unlist) %>% tibble::as_tibble(.)
```

In this notebook, I'll try to create a new feature based on the similarity between the street address and the display address. In order to do that, I used a function that is based on the Levenshtein Distance. In this particular case I used a package called "RecordLinkage" that did the work for me.

```
vec.addressSimilarity <- levenshteinSim(
      tolower(df.train$street_address),tolower(df.train$display_address))
```

Here you can see some examples of how the data looks like with the distance function,

```
df.similaritySamples <- data.frame(
      street_address = tolower(df.train$street_address),
      display_address = tolower(df.train$display_address),
      distance = vec.addressSimilarity)
head(df.similaritySamples,10)
```

```
##              street_address     display_address  distance
## 1      145 borinquen place 145 borinquen place 1.0000000
## 2             230 east 44th            east 44th 0.6923077
## 3     405 east 56th street    east 56th street 0.8000000
## 4  792 metropolitan avenue metropolitan avenue 0.8260870
## 5     340 east 34th street    east 34th street 0.8000000
## 6     145 east 16th street    east 16th street 0.8000000
## 7     410 east 13th street    east 13th street 0.8000000
## 8         1661 york avenue         york avenue 0.6875000
## 9          346 e 19 street         e 19 street 0.7333333
## 10         94 hicks street        hicks street 0.8000000
```

Finally, I decided to create a dummy variable based on this new feature and analyze the differences in the interest ratio for each group: - Group 1: Distance >= 0.5 - Group 2: Distance < 0.5

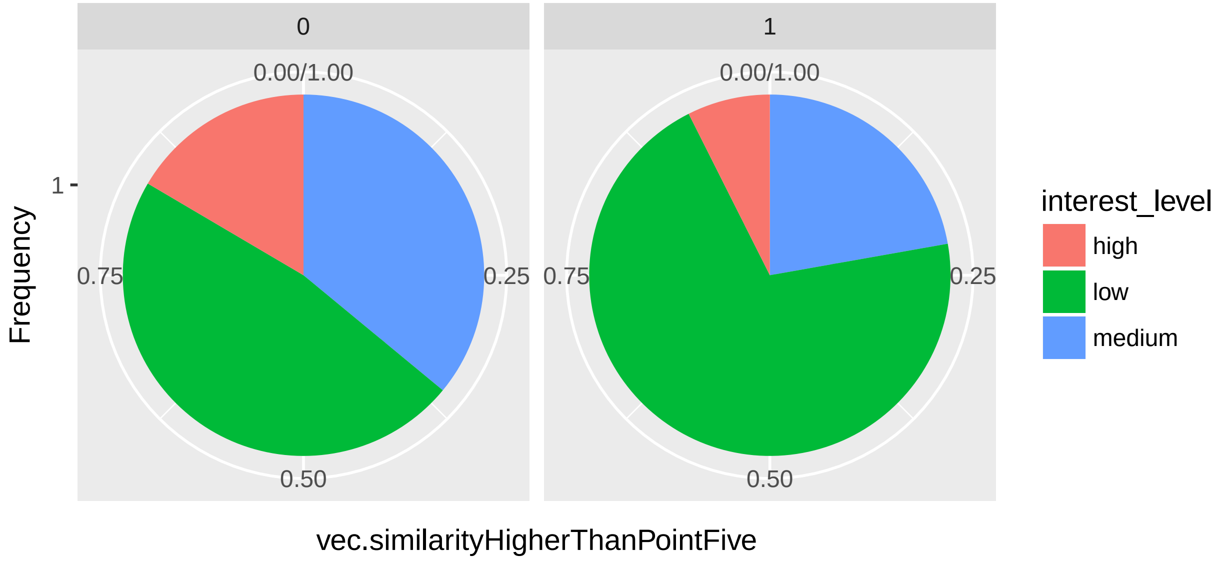
```
vec.similarityHigherThanPointFive <- ifelse(vec.addressSimilarity >=0.5,1,0)
df.train <- data.frame(df.train, vec.similarityHigherThanPointFive)
df.groupOne <- subset(df.train, vec.similarityHigherThanPointFive == 1)
df.groupTwo <- subset(df.train, vec.similarityHigherThanPointFive == 0)
```

pie chart view of the distribution of interest levels

```
df <- select(df.train, c(vec.similarityHigherThanPointFive, interest_level)) %>% drop_na()
df_tb <- as.data.frame(table(df))
vec_tb <- as.data.frame(table(df[,1]))
colnames(vec_tb) <- c("vec.similarityHigherThanPointFive", "Freq")
df_tb <- merge(df_tb, vec_tb, by = "vec.similarityHigherThanPointFive")
df_tb
```

```
##   vec.similarityHigherThanPointFive interest_level Freq.x Freq.y
## 1                                 0           high    322   1948
## 2                                 0            low    925   1948
## 3                                 0         medium    701   1948
## 4                                 1           high   3515  47395
## 5                                 1            low  33354  47395
## 6                                 1         medium  10526  47395
```

```
df_tb <- mutate(df_tb, Freq = Freq.x/Freq.y) %>% select(c(1,2,5))
bp = ggplot(df_tb, aes(x = factor(1), y = Freq, fill = interest_level))
bp = bp + geom_bar(width = 1, stat = "identity" )
bp = bp + facet_grid(facets = . ~ vec.similarityHigherThanPointFive)
bp = bp + coord_polar(theta = "y")
bp + ylab("vec.similarityHigherThanPointFive") +
      xlab("Frequency") +
      labs(fill="interest_level")
```



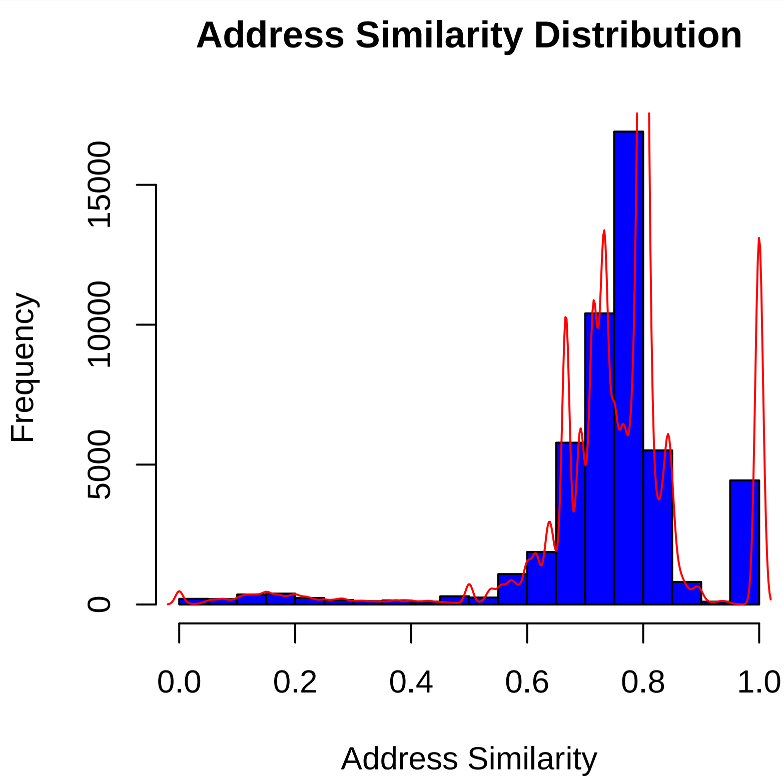
Chi Square Independence Test

```
df.chiSquareTest <- data.frame(interest_level = df.train$interest_level,
            group = vec.similarityHigherThanPointFive)
chisq.test(df.chiSquareTest$group, df.chiSquareTest$interest_level)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df.chiSquareTest$group and df.chiSquareTest$interest_level
## X-squared = 497.04, df = 2, p-value < 2.2e-16
```

Address Similarity Distribution

```
df.hist <- data.frame(interest_level = df.train$interest_level,
                      address_similarity = vec.addressSimilarity)
df.hist[is.na(df.hist$address_similarity),c("address_similarity")] <-
      mean(df.hist$address_similarity,na.rm = T)
hist <- hist(df.hist$address_similarity,
             col = "blue",
             xlab = "Address Similarity",
             main = "Address Similarity Distribution")
num.multiplier <- hist$counts / hist$density
df.density <- density(df.hist$address_similarity)
df.density$y <- df.density$y * num.multiplier[1]
lines(df.density, col = "red")
```



In my opinion, there seems to be a relation that indicates that the larger the difference is, the more interested it gets. However, most of the values have shown to have a high similarity between both fields.

UPDATE: After I ran a Chi Square an Independence Test we could clearly see that the address similarity is related to the interest level. (thanks to @saikiranputta suggestion)