Jeff Levesque

Professor Gates

IST 565

Homework #8

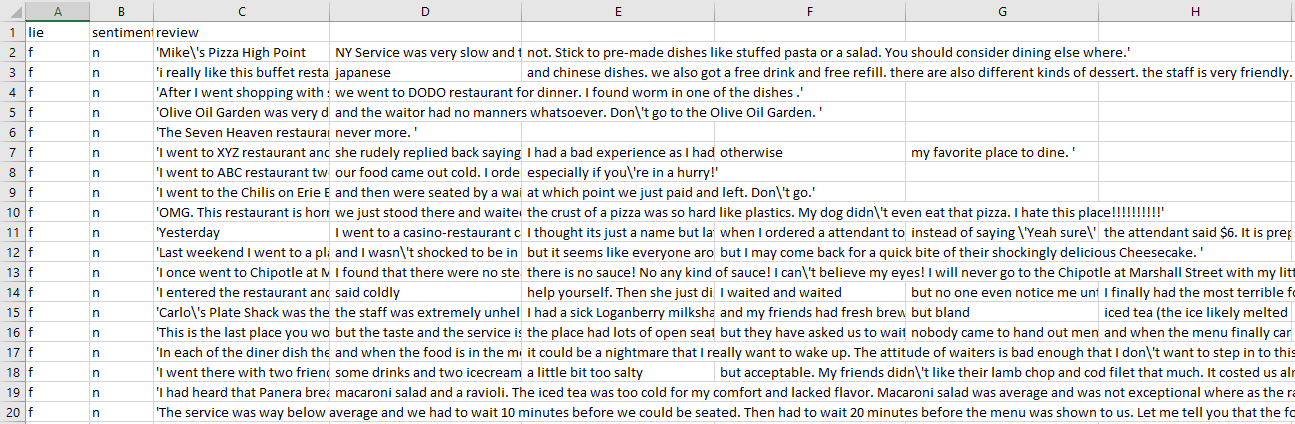
**Introduction**

**Analysis**

Data Preparation:

A single dataset was represented twice with two different file formats. The first was an arff file format, native to the weka application, and the second a csv file. This provided an option to use either weka, or an alternative csv reader. Since custom R code was written, the csv file was used.

When inspecting the csv dataset, it was immediately noticed there was 92 records, while the number of columns being non-standard. More generally, the number of columns varied in a given row, between three, to any 3+n columns. This was because the review could split into any number of columns:



Therefore, the 3+n columns needed to be collapsed into a single review column. This unique situation required a more involved method, then the simple read.csv. This meant that the column name needed to be read separately from successive rows. However,initial attempts to load the provided dataset, produced incorrect number of rows. More generally, the corresponding dataframe yielded only 45 entries:

> filepath = 'data/deception\_data\_converted\_final.csv'

> df.colnames = read.table(filepath, nrow=1, stringsAsFactors=FALSE, sep=',')

> df.full = read.table(filepath, skip=1, header=FALSE, sep='\n', fill=TRUE)

> count.fields(filepath, sep='\n', blank.lines.skip=FALSE)

[1] 1 NA NA NA 1 1 NA 1 NA 1 1 NA NA NA NA NA 1 NA NA NA NA 1 1 1 1 NA NA 1 1 NA NA 1 1 1 1 1

[37] 1 1 1 1 1 1 1 1 1 NA NA NA NA NA NA NA NA 1 NA NA NA 1 1 1 1 1 1 1 1 1 NA 1 NA NA NA NA

[73] NA 1 1 1 NA NA NA NA NA NA NA NA NA NA 1 1 1 NA NA 1 1

This required additional attributes to the read.table function:

## import dataset

filepath = 'data/deception\_data\_converted\_final.csv'

df.colnames = read.table(filepath, nrow=1, stringsAsFactors=FALSE, sep=',')

df.full = read.table(

filepath,

skip=1,

header=FALSE,

sep='\n',

quote = '',

comment.char = ''

)

Then, the corresponding single column dataframe was exploded into three columns:

out = stri\_split\_fixed(str = df.full[, c(1)], pattern = ',', n = 3)

df.split = as.data.frame(do.call(rbind, out))

colnames(df.split) = df.colnames

df.split$review = as.character(df.split$review)

Discontinued rcran module for multiclass svm had to be manually added.

**Results**

**Conclusions**