Project 4

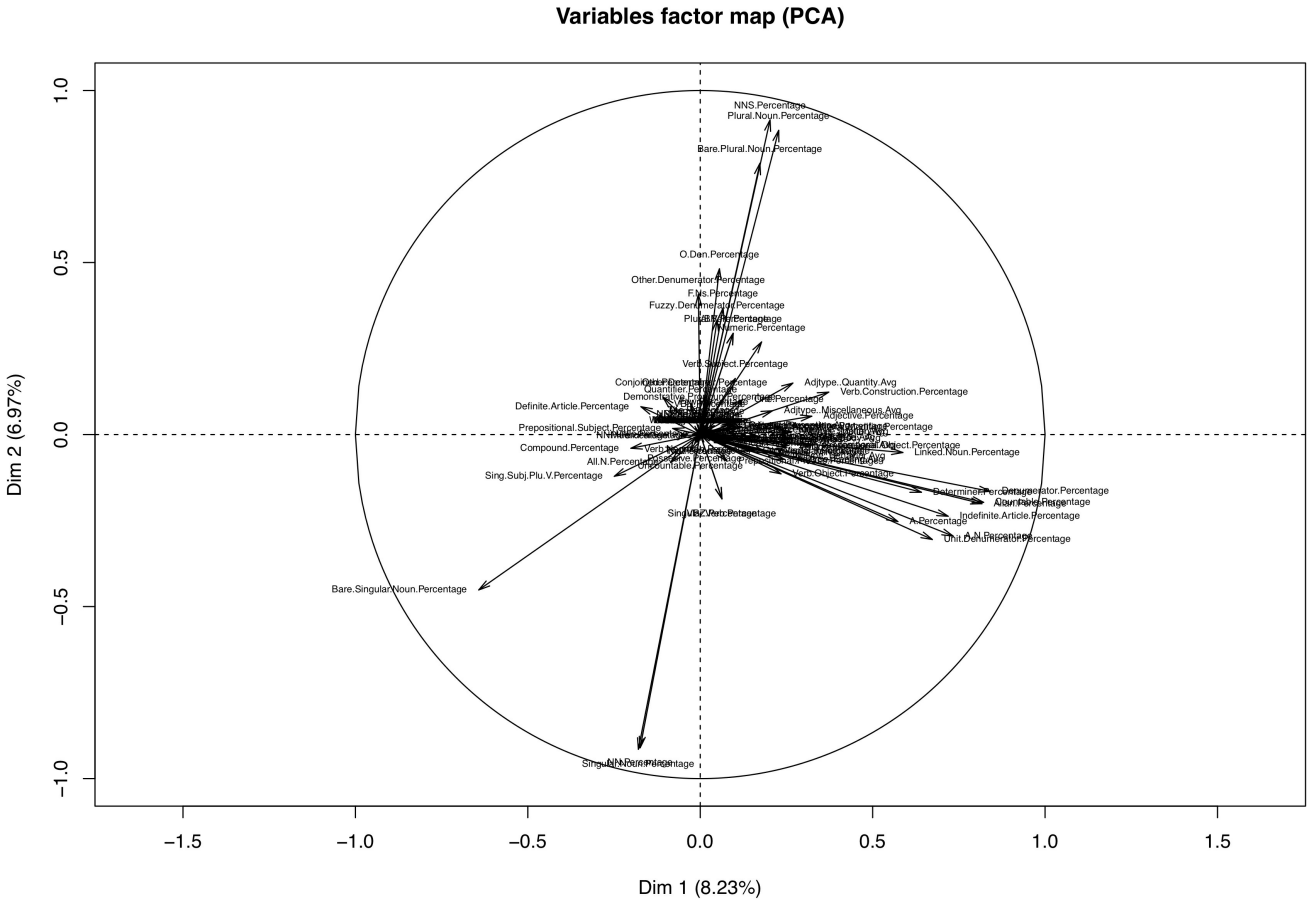
Liqi Zhu & Kiran Hu

Random Forest Modeling

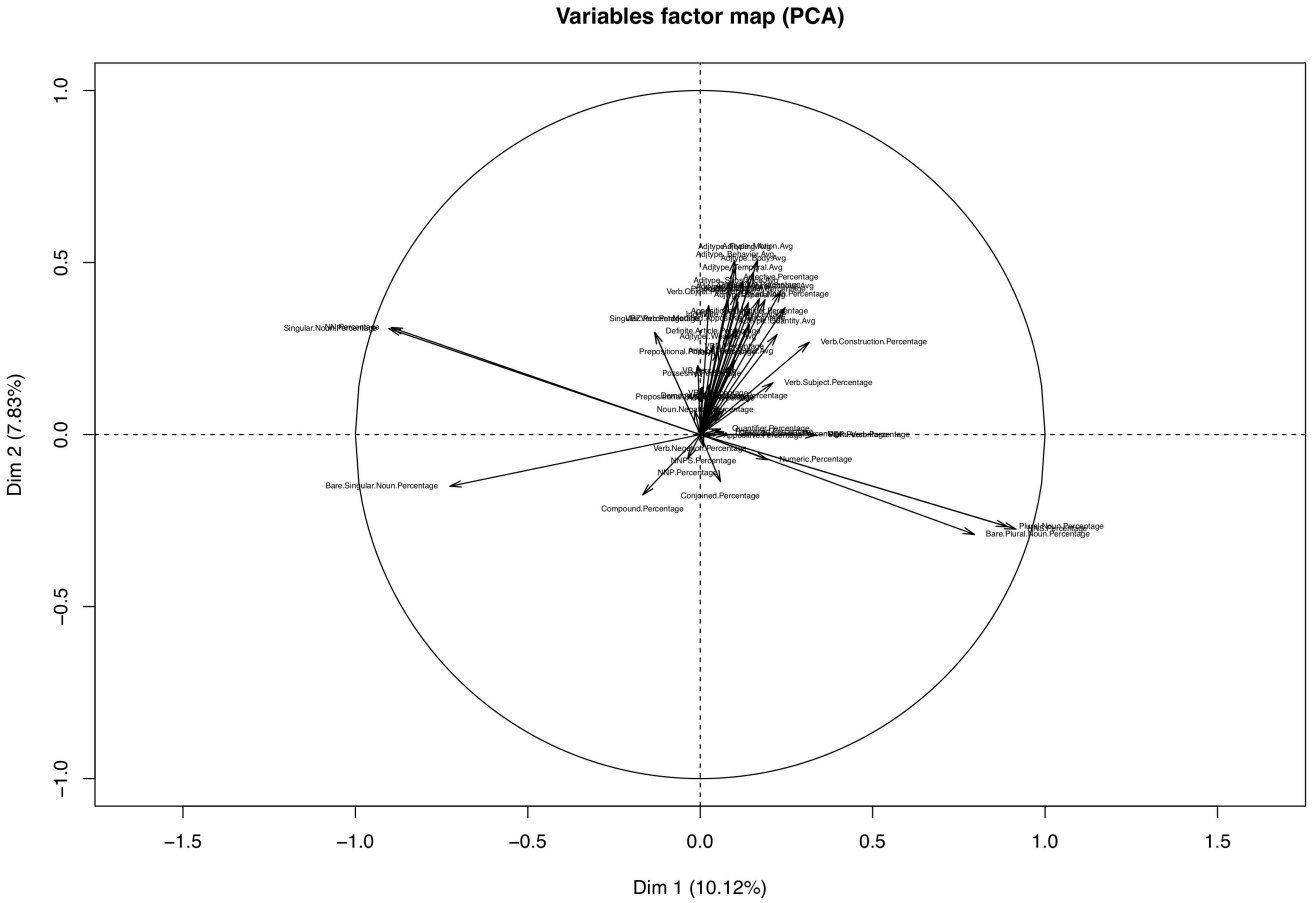
The purpose of this part is to explore how to classify one noun with these variables, and we are trying to build a model that is able to determine any noun with given properties.

With all those variables, the first thing we decided about modeling is we have to observe the correlations and quality of different variables. The first part is accomplished with the help of Variable factor map. By default, the algorithm creates a plot with individuals. The variables are represented as vectors pointing away from the origin. The angles between the vectors and the axes indicate how strongly the variables are correlated with the dimensions. The smaller the angle, the stronger the correlation. If two vectors point to almost the same direction, this means that the corresponding variables are highly correlated and therefore may represent the same underlying theoretical construct. The length of the vectors reflects how much variation in the variable is captured by this low-dimensional display, with the maximum length of 1 (limited by the circle). In other words, the length represents the quality of the representation of a variable on the plane.

```
setwd("/Users/ajkoma/Desktop/U of R/Linguistic/Project 4")
data <- read.csv('noun_distributional_information.csv')
ncol(data)
colname <- colnames(data)
library(FactoMineR)
library(randomForest)
library(party)
library(stringr)
library(Rling)
library(tidyverse)
library(factoextra)
a <- str_extract_all(string = colname,pattern='.*Percentage|.*Avg$',sim
plify = TRUE)
b <- which(a == '')
a <- a[-b]
colpercent <- subset(data,select = a)
col.pca <- PCA(colpercent)
plot(col.pca, choix = "var", cex = 0.5)

```



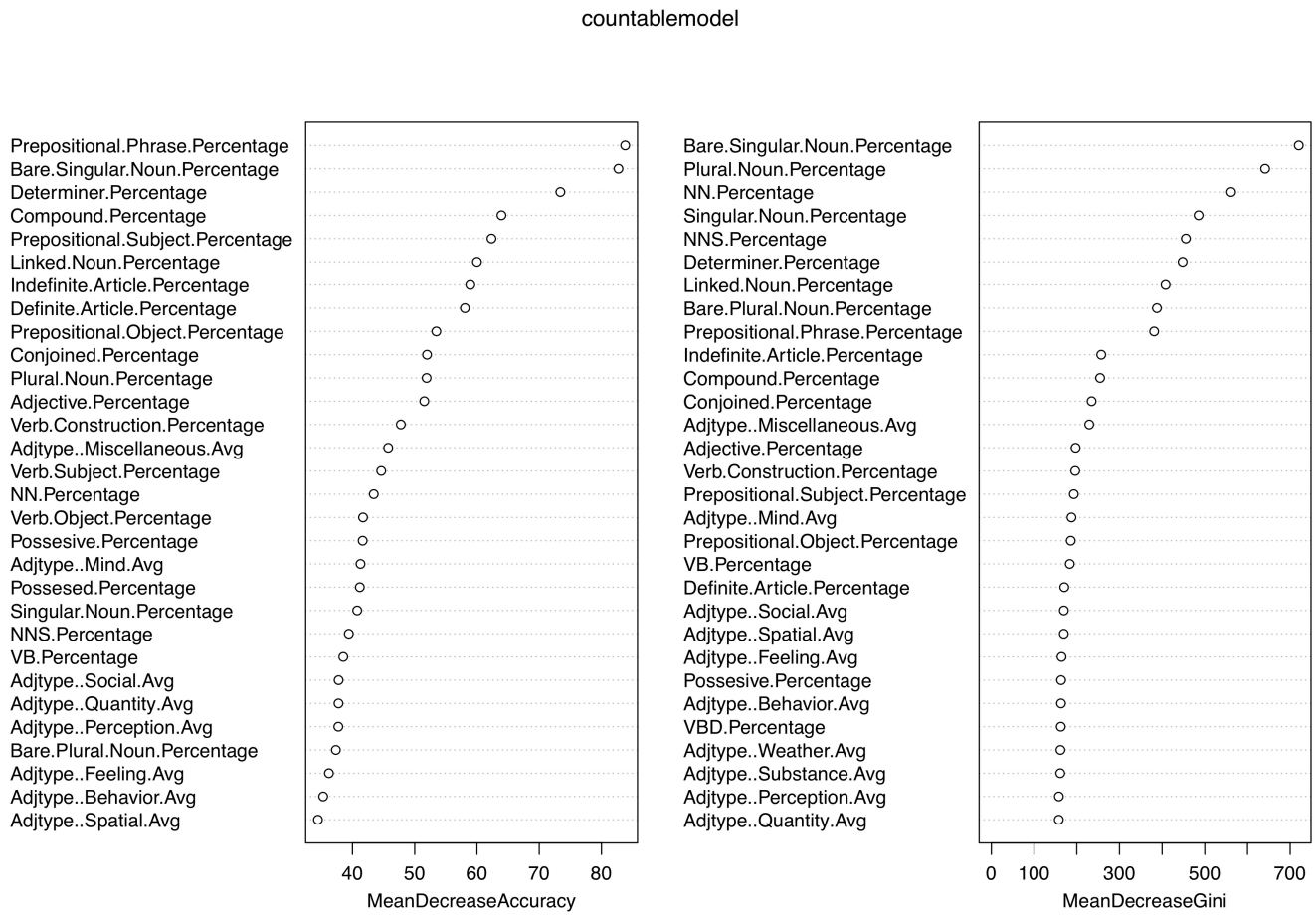

```r

colnames(colpercent)
newtag <- paste(data$countable,data$uncountable)
colpercent$countable <- factor(newtag)
singletable <- paste(data$singularia,data$pluralia)
singletable
colpercent$singulara <- factor(singletable)

colpercent <- colpercent[-c(55:104)]
col.pca <- subset(colpercent,select = -c(countable,singulara))
col.pca <- PCA(col.pca)
plot(col.pca, choix = "var", cex = 0.4)
```



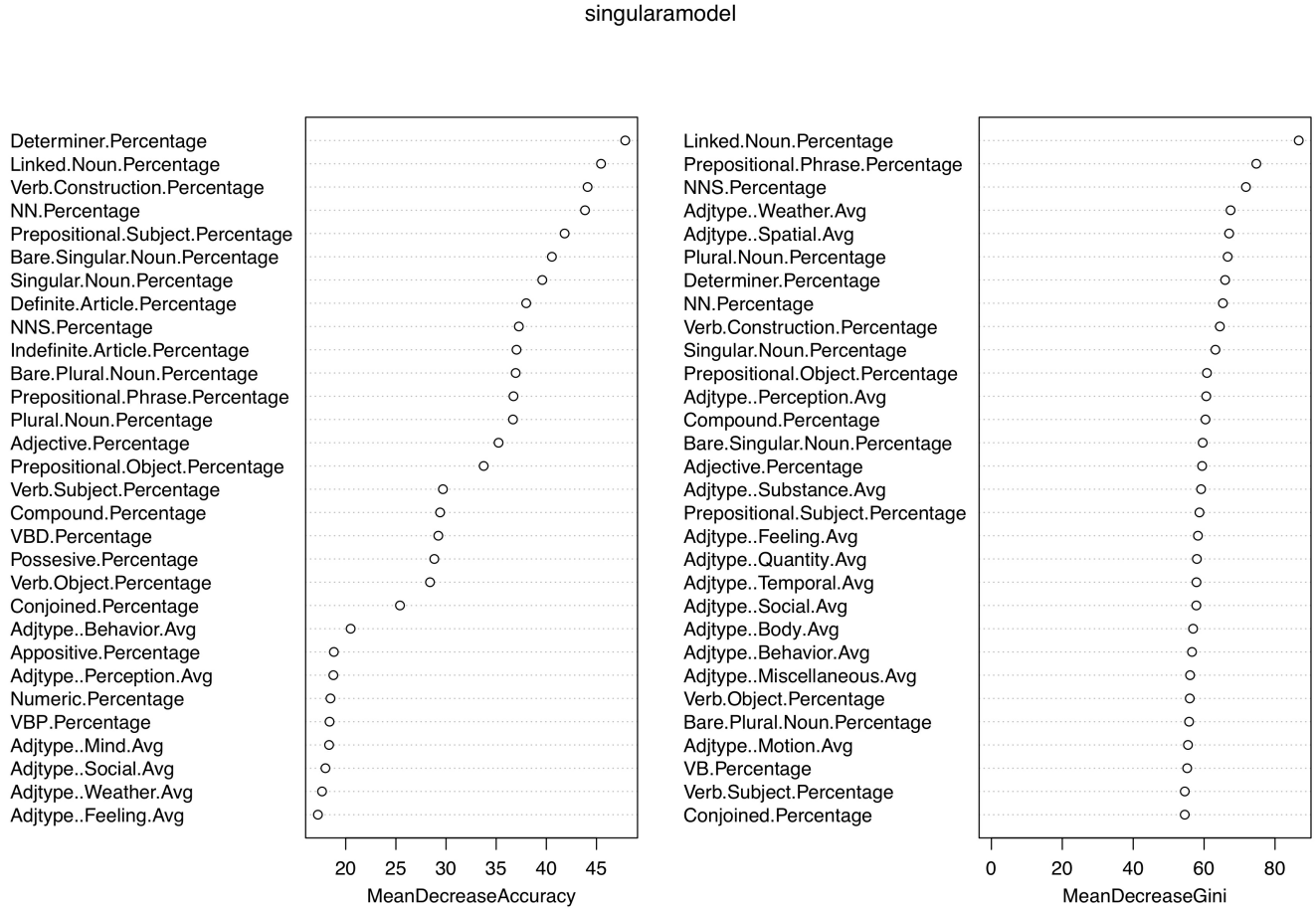
As showed above, two dimentions are those components generated by algorithm that contribute the most variances. The first chart involves many noises variables that should be deleted according to professor's tip. And with maps, we observed that there are fewer variables we should consider. The factor map is useful when we are observing the factors that may be involved in our modeling.
However, in order to observe the influence of each variable on Countable or Singularia, we have to consider the importance of each variable.

```
1.
2.  colnames(colpercent)
3.  set.seed(35)
4.  colpercent_ <-colpercent[c(1:54)]
5.  colpercent_$tag <- paste(colpercent$countable,colpercent$singulara)
6.  colpercent_$tag <- factor(colpercent_$tag)
7.  countable.model <- subset(colpercent,select=-c(singulara))
8.  single.model <- subset(colpercent,select=-c(countable))
9.  countablemodel <- randomForest(countable ~ .,
    data=countable.model,importance = TRUE, proximity = FALSE, ntree = 1000
    )
10. singularamodel <- randomForest(singulara ~ .,
    data=single.model,importance = TRUE, proximity = FALSE, ntree = 1000)
11. mixedmodel <- randomForest(tag ~ ., data=colpercent_,importance = TRUE,
```

```
    proximity = FALSE, ntree = 1000)
print(countablemodel)
plot(countablemodel)
plot(mixedmodel)
importance(countablemodel,type=1)
importance(singularamodel,type=1)
importance(mixedmodel,type =1)
varImpPlot(countablemodel)
varImpPlot(singularamodel)
varImpPlot(mixedmodel)
```

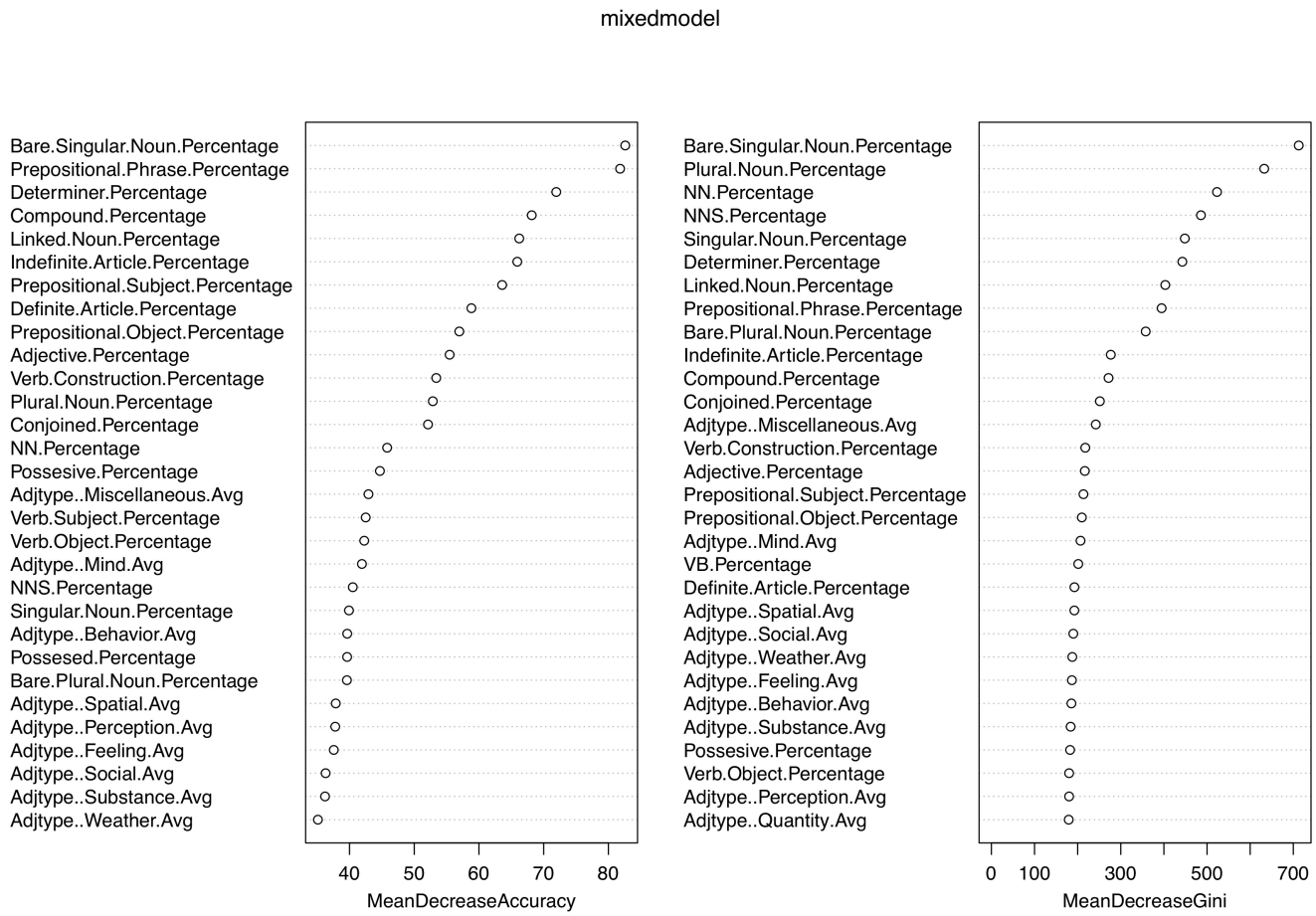


This chart is the importance ranking of variables that contributed to the Random Forest model. The model is built with correspond to 4-level **factor(paste(countable,uncountable))**.
Gini Index are reference for local tree spliting, and based on Mean Decrease Accuracy, we observed the identifiers that contributes most to the determination of countable or uncountable for a noun.



This chart is the importance ranking of variables with the most importance that contributed to the Random Forest model. The model is built with correspond to 4-level **factor(paste(singularia,pluralia))**.

Based on Mean Decrease Accuracy, we observed the identifiers that contributes most to the determination of singularia or pluralia for a noun.



This chart is the importance ranking of variables that contributed to the Random Forest model. The model is built with correspond to 14-level **factor(paste(countable,uncountable,singularia,pluralia))**.
Gini Index are reference for local tree spliting, and based on Mean Decrease Accuracy, we observed the identifiers that contributes most to the defenitiveness for a noun.

As a conclusion, with the variables in chart, we can built a classification model with less variable.
As we found, Random Forest model using all variables results in high error rate for those "N N"" Y Y"corresponding values. And with those models, one maybe able to determine the classification of rest corresponding values with an average error rate of around 10%. And reducing the variables according to importance doesn't improve the accuracy. We are considering if this is related to the data point in dataset where percentage are calculated by different bases.
And we also found that the mixed model with 14-level factor is not reliable according to error rate.
Within the given time, this might be what we can do. For futher study, we would like to split the dataframe into smaller ones for study of different clusters and we may build radom forest model with less variables to observe the different accuracy and influence each cluster have in modeling.
And we are aiming at improving the error rate of classification model, we are also considering other algorithms such as Xgboost, we wish we have more time.