Preliminary Analysis

November 1, 2015

1 Distribution of correct answers

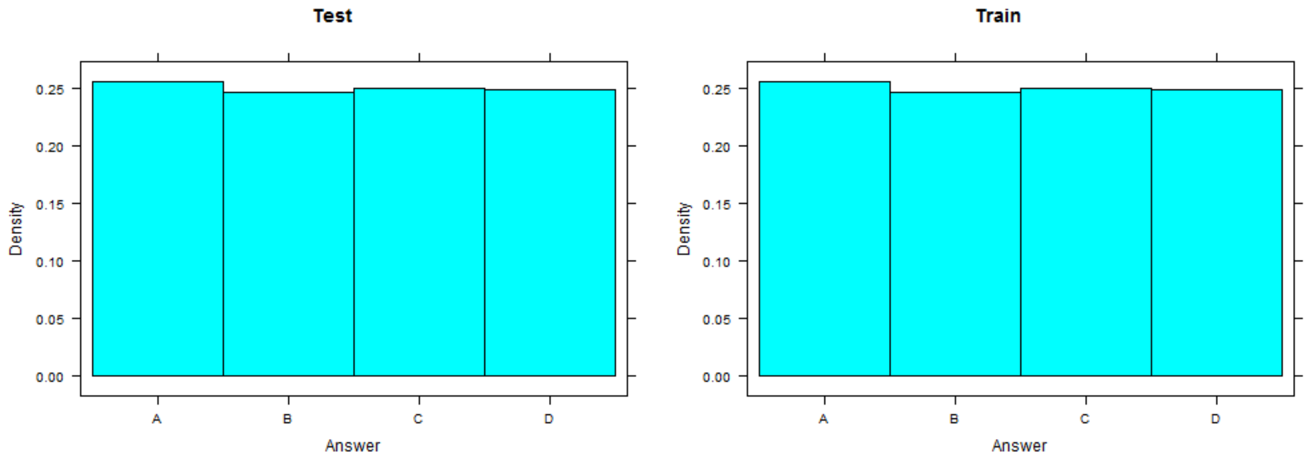
1.1 Code

```
## Test answer distribution
ans1 <- test$correctAnswer
table(ans1)/length(ans1)
# ans1
#         A         B         C         D
# 0.2556515 0.2457269 0.2501378 0.2484837

## Train answer distribution
ans2 <- train$correctAnswer
table(ans2)/length(ans2)
# ans2
#         A         B         C         D
# 0.2556515 0.2457269 0.2501378 0.2484837

library('lattice')
x1 <- histogram(~correctAnswer, data=test,
          type="density",
          xlab="Answer",
          main = "Test")
x2 <- histogram(~correctAnswer, data=train,
          type="density",
          xlab="Answer",
          main = "Train")
require(gridExtra)
grid.arrange(x1, x2, ncol=2)
```

1.2 Plots



2 Term Frequency

2.1 Code

```r
library('tm')

# Build corpus
train_data.corpus <- Corpus(VectorSource(train$question))

# make each letter lowercase
train_data.corpus <- tm_map(train_data.corpus, tolower)

# remove punctuation
train_data.corpus <- tm_map(train_data.corpus, removePunctuation)

# remove generic and custom stopwords
train_stopwords <- c(stopwords('english'))
train_data.corpus <- tm_map(train_data.corpus, removeWords, train_stopwords)

# build a term-document matrix
train_data.dtm <- TermDocumentMatrix(train_data.corpus)
train_data.dtm

# inspect most popular words (change lowfreq for lower bound)
findFreqTerms(train_data.dtm, lowfreq=50)
```

```
# Most Frequent (>=50) Terms
  [1] "according"    "acid"         "algorithm"    "along"        "also"         "another"
        "associated"
  [8] "body"         "called"       "can"          "carbon"       "catalyst"     "cause"
        "caused"
 [15] "causes"       "cell"         "cells"        "certain"      "chemical"     "class"
        "common"
 [22] "complex"      "compound"     "compounds"    "constant"     "contains"     "create"
        "created"
 [29] "density"      "derivative"   "derived"      "described"    "developed"    "
   discovered"   "disease"
 [36] "divided"      "due"          "effect"       "electron"     "element"      "elements
   "    "energy"
 [43] "entities"     "enzyme"       "equal"        "equals"       "equation"     "
   experiment"   "factor"
 [50] "field"        "first"        "form"         "formation"    "formed"       "forms"
        "formula"
 [57] "found"        "function"     "functional"   "functions"    "gene"         "given"
        "gives"
 [64] "group"        "high"         "include"      "involves"     "known"        "law"
        "light"
 [71] "like"         "magnetic"     "man"          "mass"         "material"     "may"
        "mechanism"
 [78] "metal"        "method"       "model"        "molecule"     "molecules"    "name"
        "named"
 [85] "namesake"     "negative"     "number"       "numbers"      "object"       "objects"
        "occur"
 [92] "occurs"       "often"        "one"          "ones"         "order"        "organ"
        "paper"
 [99] "part"         "particle"     "particles"    "pathway"      "phenomenon"   "phylum"
        "potential"
[106] "power"        "presence"     "pressure"     "problem"      "process"      "produce"
        "produced"
[113] "product"      "property"     "proportional" "proposed"     "protein"      "proteins
   "    "quantity"
[120] "quantum"      "reaction"     "region"       "related"      "result"       "results"
        "rule"
```

```
[127] "set"          "showed"       "solution"     "sometimes"    "space"        "square"
        "state"
[134] "states"       "step"         "structure"    "structures"   "study"        "
   substance"    "substances"
[141] "surface"      "syndrome"     "synthesis"    "system"       "systems"      "
   technique"    "temperature"
[148] "term"         "theorem"      "theory"       "three"        "time"         "times"
        "two"
[155] "type"         "types"        "use"          "used"         "uses"         "using"
        "value"
[162] "variety"      "version"      "via"          "water"        "whose"        "work"
```

3 Word Associations

```
## Example: Associations with the word "cells", lower bound on correlation 0.2
findAssocs(train_data.dtm, 'cells', 0.20)
             cells
glial         0.33
purkinje      0.32
schwann       0.27
amacrine      0.26
epithelial    0.23
# Note association score is percentage that term occurs with the search term (i.e. "glial"
     occurs with "cells" 33% of the time)
```

4 Clustering

4.1 Code

```
# http://www.statmethods.net/advstats/cluster.html (Robert I. Kabacof s  Cluster
    Analysis  )

# remove sparse terms = simpler cluster plot (need to think about this in terms of
    classifying question types)
train_data.dtm2 <- removeSparseTerms(train_data.dtm, sparse=0.95)

# convert the sparse term-document matrix to a standard data frame
train_data.df <- as.data.frame(inspect(train_data.dtm2))

# inspect dimensions of the data frame
nrow(train_data.df)
ncol(train_data.df)

train_data.df.scale <- scale(train_data.df)
d <- dist(train_data.df.scale, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram? (i.e. 1-gram)

groups <- cutree(fit, k=5) # cut tree into 5 clusters
# draw dendogram with red borders around the 5 clusters
rect.hclust(fit, k=5, border="red")
```

4.2 Plot

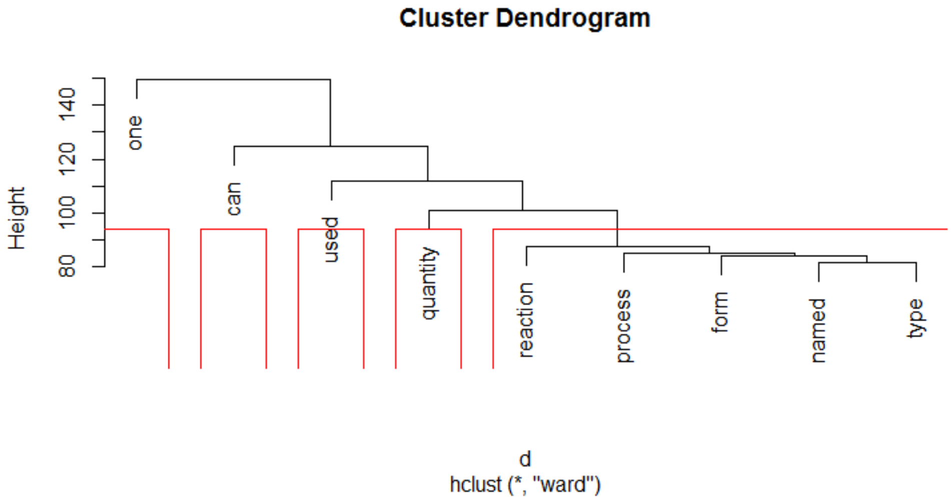


Figure 1: More popular terms are higher up, while more associated terms are closer together