- The problem is focused on Text Mining
- You will have to tree main tasks:
  - Pre-processing - data is not in a readable format for R and Weka
  - Choose the best set of attributes for your learning algorithms (Attribute Selection)
  - Execute and evaluate learning algorithms with these attributes
- You are free to use any learning algorithms available in RWeka, Weka and R

- You must convert the dataset into a readable format for R (csv, xml, etc)
- The following code assumes you have done it:

```
library(tm)
data<-read.csv("projdataset.csv")
corpus<-Corpus(DataframeSource(data.frame(data[, 3])))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, function(x)
              removeWords(x, stopwords("english")))
tdm <- TermDocumentMatrix(corpus)
s.matrix(tdm)
```

```
                Docs
Terms            1 2 3
  0001           0 0 1
  004            0 0 1
  005            0 0 1
  013            0 0 1
  1000           0 1 0
  110            1 0 0
  115            1 0 0
  118            1 0 0
  135            1 0 0
...
```

- If you do not wish to use R for pre-processing, you may find useful use Weka/KEA
  `http://www.nzdl.org/Kea/index.html`