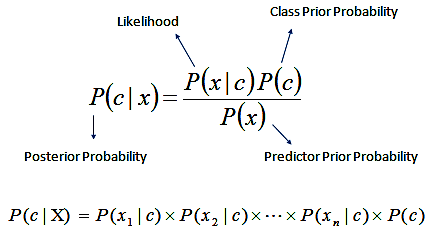
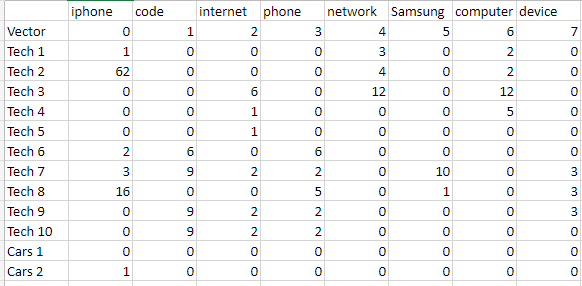
**Data Classification**



:<https://uc-r.github.io/naive_bayes> (Image)

Using the Naive Bayes classifier method we were able to augment our scraper to define the articles it read as one of our target classification topics, technology, automotive, or sports.

Figure 1



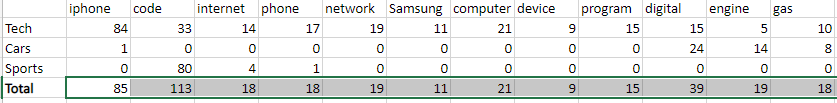
By selecting a set of words that would be common to articles of each perspective typing we were able to form the feature vectors mentioned previously and form a DTM.

(Above a subset of the data is shown, however some columns are left out due to space constraints, Fig 1.)

Using this DTM we were able to apply our test training set in order to extract data used in the Naive Bayes Classifier method

(Below a subset of the data is shown, however some columns are left out due to space constraints, Fig. 2)

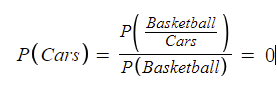
Figure 2

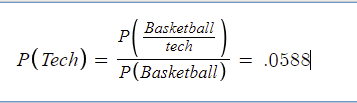


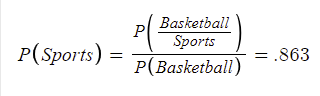
These numbers assist us in learning the Class prior probability, likelihood and predictor probability which will be used to classify additional articles.

Something interesting to note is that there are no occurrences of the target word “hockey” that is specified in the target vector. Hence, this term shall be ignored from any calculations.

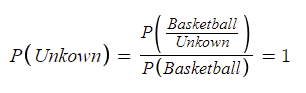
When adding the First Unknown article to be classified, the first real classification of the scraping tool, an article was selected and the only word it contained from the targets was the term “basketball”, which it included four times. This leaves our Naive Bayes equation as follows:







Now that the weight for each article has been calculated based upon the addition to the dataset (the unknown article), we can calculate the probability score, whichever it is closest to is going to be our classification.



As “basketball” is the only term to appear in the unknown article the program defines the probability score as 1, which is closest to the probability score of Sports, from our judgement with the data alone this seems to be a fitting score.