

COMP348

Document Processing and the Semantic Web

Assignment I, Part A

Mark Jaskolski [40969150]

21-MAR-2008 14:30

The contents of this are the works of Mark Jaskolski, except where explicitly noted as otherwise.

1 Introduction

The world is full of data, in raw, natural forms, and at times it is important for us to be able to take this data and derive useful information from it. Many methods and tools exist for doing this, one such tool is the science of statistics. For our particular purpose, we are using statistics (in a very rudimentary way) to compare the outputs of two systems (text categorizers), and determine if the difference in the output accuracy is statistically significant. Which is to say that the results are either significantly better/worse than the control set of results, or the results are close enough to the control set that the difference is probably just a result of randomness in the data.

2 System Description

The AdAge system uses sophisticated categorizing algorithms to determine the age group of a blog author, based on the contents of the blog. The results of this have been presented as part of this assignment in the form of a set of control data (the model), which is the correctly categorized results, and the data generated by the system.

3 Experiment and Results

The make-up of the data given is as follows:

	Correctly categorized		Incorrectly categorized		Total	
Teenager	306	30.6%	238	23.8%	544	54.4%
Adult	254	25.4%	202	20.2%	456	45.6%
Total	560	56.0%	440	44.0%	1000	100.0%

The testing process basically consisted of taking this data, and determining if the analysis algorithm was any good. This was achieved by calculating a baseline efficiency level and using it as a benchmark for comparing the results of the AdAge system.

To begin with, the accuracy of the AdAge software was calculated using the following equation:

$$\text{Accuracy} = \text{CorrectlyCategorizedItems} / \text{TotalItems}$$

$$\text{Accuracy} = 560 / 1000 = 0.56 (56\%)$$

Then, the baseline accuracy was calculated by taking the group (teenager or adult) with the largest number of items in the control data, and using that to derive an accuracy value.

$$\text{Baseline} = \text{GroupTotal} / \text{TotalItems}$$

$$\text{Baseline} = 544 / 1000 = 0.544 (54.4\%)$$

Finally, a the statistical significance of the two results is calculated using a proportional z-test. This is done using this formula:

$$\text{Difference} = (\text{Accuracy} - \text{BaseLine}) / \text{sqrt}(\text{Baseline} * (1 - \text{BaseLine}) / \text{TotalItems})$$

$$\text{Difference} = (0.56 - 0.544) / \text{sqrt}(0.544 * (1 - 0.544) / 1000) = 1.01586994446$$

From this result, we can see that the difference between results is not statistically significant, since the difference value is less than 1.96.

4 Conclusion

This experiment shows that (with the given data), the classifier within the AdAge software is only marginally better than the baseline. However, due the difference between results is not statistically significant, and therefore can be argued that any difference which does appear is simply a result of random chance, instead of the algorithms efficiency.

Source Code – assignment.py

```
# COMP348 Assignment - Part I
# By Mark Jaskolski (mark.jaskolski@students.mq.edu.au)
# 40969150

from array import array
from math import sqrt

def test():
    loadFile("40969150")

def testYenchi():
    loadFile("40964787")

def testSample():
    loadFile("12345675")

def loadFile(studentID):
    GEEZER_MARKER = "-1"
    KIDDIE_MARKER = "+1"

    GEEZER = 0
    KIDDIE = 3
    RESULT = 6

    REAL = 0
    FAKE = 1
    TOTAL = 2

    stats = array('L', [0,0,0,0,0,0,0,0])

    modelFile = open("model-"+studentID+".txt", "r")
    systemFile = open("system-"+studentID+".txt", "r")

    modelLine = modelFile.readline().strip()
    systemLine = systemFile.readline().strip()
    while modelLine != "" and systemLine != "":
        if systemLine == GEEZER_MARKER or systemLine == KIDDIE_MARKER:
            if systemLine == GEEZER_MARKER:
                group = GEEZER
            elif systemLine == KIDDIE_MARKER:
                group = KIDDIE

            if modelLine == GEEZER_MARKER or modelLine == KIDDIE_MARKER:
                if systemLine == modelLine:
                    result = REAL
                    stats[group+TOTAL] += 1
                else:
                    result = FAKE
                    if group == GEEZER:
                        stats[KIDDIE+TOTAL] += 1
                    else:
                        stats[GEEZER+TOTAL] += 1

                    stats[group+result] += 1
                    stats[RESULT+result] += 1
                    stats[RESULT+TOTAL] += 1
                else:
                    print "WTF?! " + systemLine + " in model!"
            else:
                print "WTF?! " + systemLine + " in system!"

            modelLine = modelFile.readline().strip()
            systemLine = systemFile.readline().strip()

    if modelLine != "" or systemLine != "":
        print "WTF?! MODEL=" + modelLine + " SYSTEM=" + systemLine + "!"

    modelFile.close()
    systemFile.close()

    print "DATA BREAKDOWN [" + studentID + "]:"
    print "      real  fake  total"
    print "geezer   " + str(stats[GEEZER+REAL]).rjust(4) + " " + str(stats[KIDDIE+FAKE]).rjust(4) + " " +
str(stats[GEEZER+TOTAL]).rjust(4)
    print "kiddie   " + str(stats[KIDDIE+REAL]).rjust(4) + " " + str(stats[GEEZER+FAKE]).rjust(4) + " " +
str(stats[KIDDIE+TOTAL]).rjust(4)
    print "total   " + str(stats[RESULT+REAL]).rjust(4) + " " + str(stats[RESULT+FAKE]).rjust(4) + " " +
str(stats[RESULT+TOTAL]).rjust(4)

    accuracy = float(stats[RESULT+REAL])/float(stats[RESULT+TOTAL])

    if stats[GEEZER+TOTAL] >= stats[KIDDIE+TOTAL]:
        baseline = float(stats[GEEZER+TOTAL])/float(stats[RESULT+TOTAL])
    else:
        baseline = float(stats[KIDDIE+TOTAL])/float(stats[RESULT+TOTAL])

    difference = (accuracy-baseline)/sqrt(baseline*(1-baseline)/float(stats[RESULT+TOTAL]))

    print
    print "accuracy  " + str(accuracy)

    print
    print "baseline  " + str(baseline)

    print
    print "difference " + str(difference)
    if difference > 1.96:
        print "The difference is statistically significant."
    else:
        print "The difference is not statistically significant."
```