FINAL REPORT INTERACTIVE LEARNER

This is the format for your third and final report on the Interactive Learner Assignment.

You must fill in this format and submit it together with src code and readme file in a zip file.

Using this form helps you and the assessment team to check if you have done what is needed.

The questions below cover all three parts of the assignment.

Section 2 and 3 are about part B (and the optional bonus assignment).

Section 4 is about part A (design) and C (performance).

Section 5 is about part C.

This document contains also information about the main assessment criteria for computing the final mark for this assignment.

Submit this document on BB before the deadline in a zip file together with the other deliverables (see below).

Put additional information at the end of this report only (for example figures and tables, or any comments that you want to communicate).

This report does not replace your deliverables of parts A and B. But you have now the opportunity to improve those parts –based on feedback you received on those parts- by reporting those parts in this report. In case you do this final delivery will be assessed, not the earlier version.

Note that this is the final deliverable. There is no repair round.

I.　Final Report for Interactive Learner

1. The developers

Group Number: 48

Names:　　Tjeerd Jan Heeringa, s1497324

　　　　　Joshua van Kleef, s1385801

2. The Classifier: type of NBC and performance on data sets

Which type did you implement and what is the accuracy on blogs and mails?

Multinomial:　　**YES**　~~NO~~

Blogs:　　　　　72 %

Mails:　　　　　98 %

Binomial:　　~~YES~~ **NO**

Blogs:　　　　　N/A

Mails:　　　　　N/A

Does your classifier work for any number of class values?　**YES** ~~NO~~

If you computed additional performance measures or confusion matrices put the results in the appendix at the end of this report.

You should at least implement one variant of the NBC correctly and provide accuracies for at least one of the given corpora. Performance should exceed base line performance.

If your classifier only works for problems with 2 class values your work does not meet one of the assignment's requirements.

3. The Vocabulary: feature/word selection

What did you implement and test?

Text normalization:　**YES** ~~NO~~

Regarding word filtering (feature selection):

Stopwords removed: ~~YES~~ **NO**

Filter words based on number of occurrences

Rare words removed: ~~YES~~ **NO**

Words that occur very often removed: ~~YES~~ **NO**

Other feature selection methods implemented: (Fill in here what method)

....

You can add test results at the end of this report.

Smart feature selection is NOT required. But some form of selection may be required to improve runtime or performance of your classifier.

For the BONUS you have to implement and test CHI SQUARE.

Provide test tables at the end and deliver the list of distinguishing words and their chi^2 values in a file chi-words.txt.

4. The Interactive Learner: the iterative strategy

A ``session'' with the IL refers to the process from starting up the IL system to finishing the program.

These questions concern the learning cycle.

- 1) Does the interactive learner (IL) **only** store the new information (documents and classes based on feedback by the user during a session) for an update at a later session? (without updating the classifier during a session) **YES** ~~NO~~
- 2) Does the Interactive Learner update the classifier during a session?   **YES** ~~NO~~
- 3) Is the Vocabulary updated every time when a document is given a corrected class by the user during a session?   **YES** ~~NO~~
- 4) Are the probability tables updated every time when a document (or a number of documents) is given a corrected class during a session?   **YES** ~~NO~~
- 5) Can the user add new classes during a session?   **YES** ~~NO~~
- 6) Does the IL work for classifiers with any number of classes? **YES** ~~NO~~

You may add a flow diagram, and other diagrams to illustrate architecture and the interactive process in the appendix.

Submit a small demo corpus to show that your IL ``learns'' during a session.**

Assessment:

If you (truthfully) answer question 1 with YES and your answer to question 2 is NO you did not implement an IL as required.

If you answer question 6) with NO you do not meet a requirement of the assignment.

5. The User Interface and the User Instructions

GUI: **YES**   ~~NO~~ (**if yes provide picture at end of report**)

TUI: ~~YES~~ **NO**

Submit a README file that gives information about :

- o **System requirements (tested) for compiling and running the code**
- o **How to install (including where data files should be stored)**
- o **How to run**
- o **What types of interactions the the TUI or GUI offers**

Submit your src code.

Submit your small test corpus.

Do not submit the large mails or blogs corpora!!

Put everything in a zip file with a name that identifies your group.

Appendix A GUI

Multinomal Naive Bayesian Classifier Group 48

Directories

Choose Train Directory   Choose Test Directory

Operations:

Train Classifier

Load Pretrained Classifier

Test the speed

Test the accuracy on the test set

Test on single file:

Load