FINAL REPORT INTERACTIVE LEARNER

Module 6 Artificial Intelligence Practical Assignment 2016-2017

Gerwin Puttenstein, s1487779

I. Final Report for Interactive Learner

1. The developers

Note: the practice assignment has to be made by pairs of 2 students that did not do the assignment in a previous year. Students that make this assignment for the second time have to make the assignment individually. They are allowed to use their own work of a previous attempt.

Group Number:

Names: *Gerwin Puttenstein*

Programming language used to implement the system (Python or Java): *Java*

IDE used to develop the system (i.e. Eclipse, …): *IntelliJ 2016.3.1*

Did you write the system code completely yourself? YES ~~NO~~

If not indicate which parts did you copy from others (include reference)

NA

Work submitted must be your own creation. We test your code for plagiarism.

If you use substantial parts of code written by others without acknowledgment/reference you failed for this assignment.

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

Fill in which type of NB classifier you implemented and what the accuracy is on the blogs and mail data sets. Give vocabulary size for the best performance of your classifier.

Multinomial: YES ~~NO~~

Vocabulary size: Blog: 415 unique words Mail: 2572 unique words

Blogs: accuracy: 40correct 10 incorrect 80% baseline: Female/Male *(50%)*
There were 25 female blogs and 25 male blogs for testing

Mails: accuracy: 287 correct 4 incorrect 98,6% baseline: Ham *(83%)*
There were 242 ham mails and 49 spam mails for testing

Binomial: ~~YES~~ NO

Vocabulary size:

Blogs: accuracy: baseline:

Mails: accuracy: baseline:

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 an 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 will be graded lower than when it works for any finite number of class values.

3. The Vocabulary: feature/word selection

What did you implement and test?

Text normalization: YES ~~NO~~

Regarding word filtering (feature selection, i.e. what words are included in the vocabulary):

Stopwords removed: YES ~~NO~~

Filter words based on number of occurrences
Yes, I create a mapping of word-count combinations

Rare words removed: YES ~~NO~~

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

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

………*Chi square*……………………………………………………………………………………

Chi square has to be implemented as feature selection method.

Provide test tables showing accuracy for different sizes of the vocabulary 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 Interactive Learner refers to the process from starting up the IL system to closing the system. A system requirement is that the IL learns during such a session.

The following questions concern the learning cycle of the Interactive Learner.

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 Does your system support the creation of a completely new annotated document corpus ? YES ~~NO~~

6 Does the IL work for classifiers with any number of classes? YES ~~NO~~

You may add diagrams to illustrate architecture and the interactive process in an appendix.

Your IL must allow to re-train the classifier during a session after the user provided new annotated documents.

You must be able to demonstrate your system with a small demo corpus to show that your IL ``learns’’ during a session.

I used the blog corpus to test and demonstrate that the interactive is working and is learning. I found that this corpus was small enough to use for testing and demonstration.

5. The User Interface and the User Instructions

GUI: YES ~~NO~~ (**if yes provide picture(s) at end of report**)
The GUI is created with the build in GUI builder from IntelliJ. That is what the GUI.form file is used for.

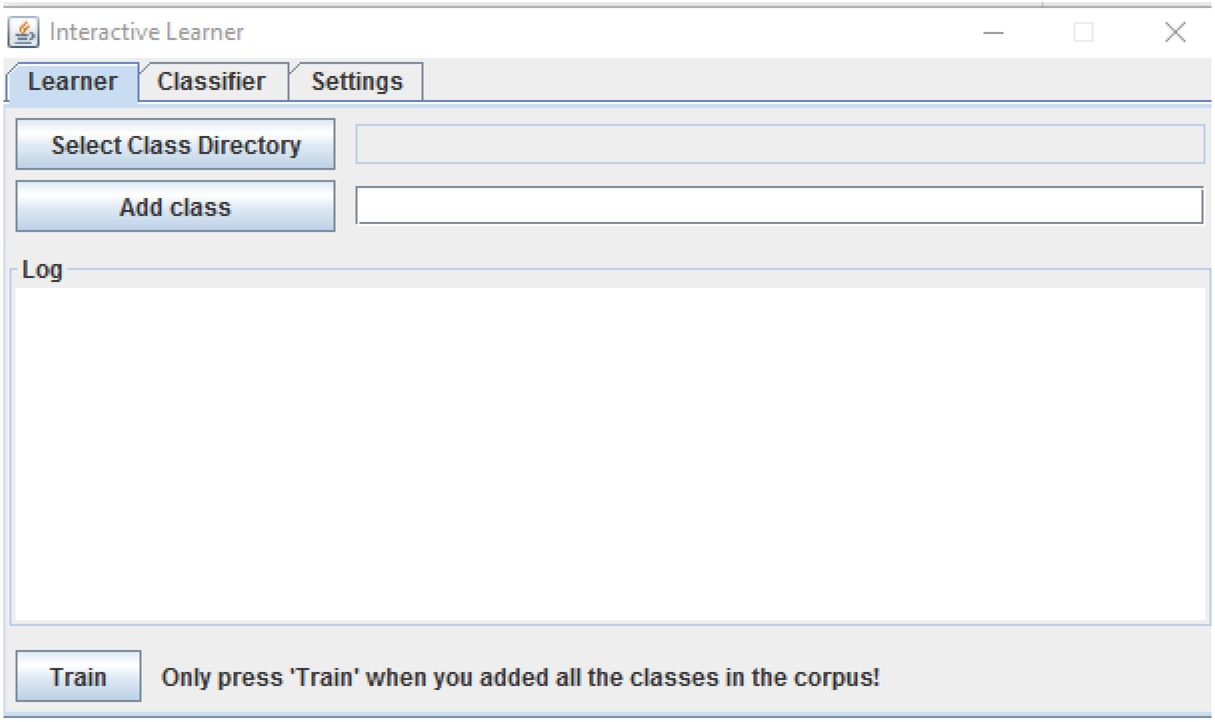
TUI: ~~YES~~ NO

You must provide a GUI that satisfies basic user requirements of human computer interaction,

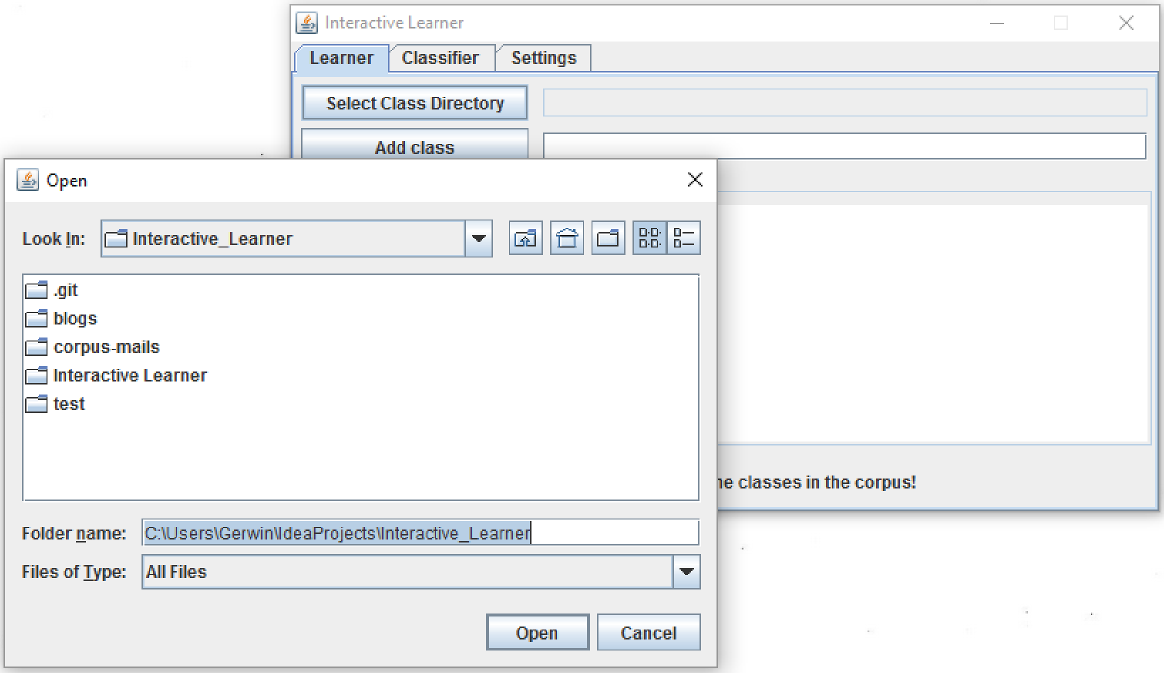
i.e. it should be clear for new users what can or should be done at each stage in the process.

APPENDICES GO HERE (please order them per section)

GUI:

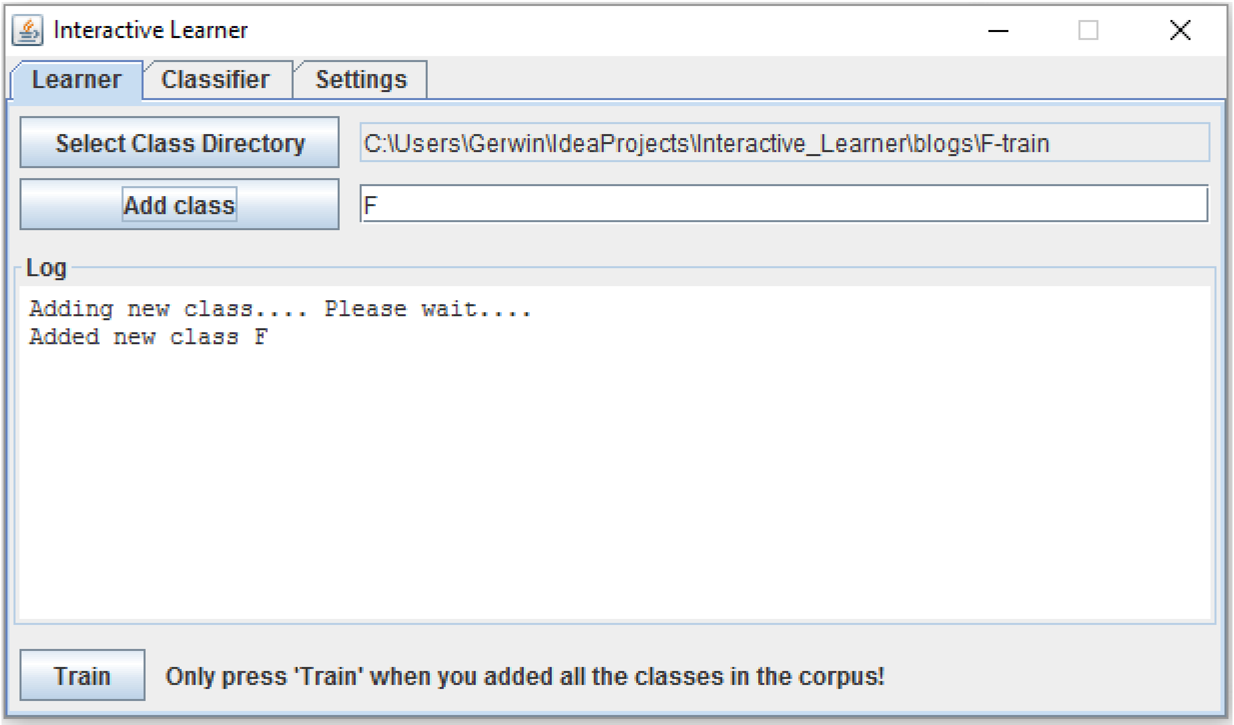


Only press 'Train' when you added all the classes in the corpus!



Folder name: C:\Users\Gerwin\IdeaProjects\Interactive_Learner

Files of Type: All Files

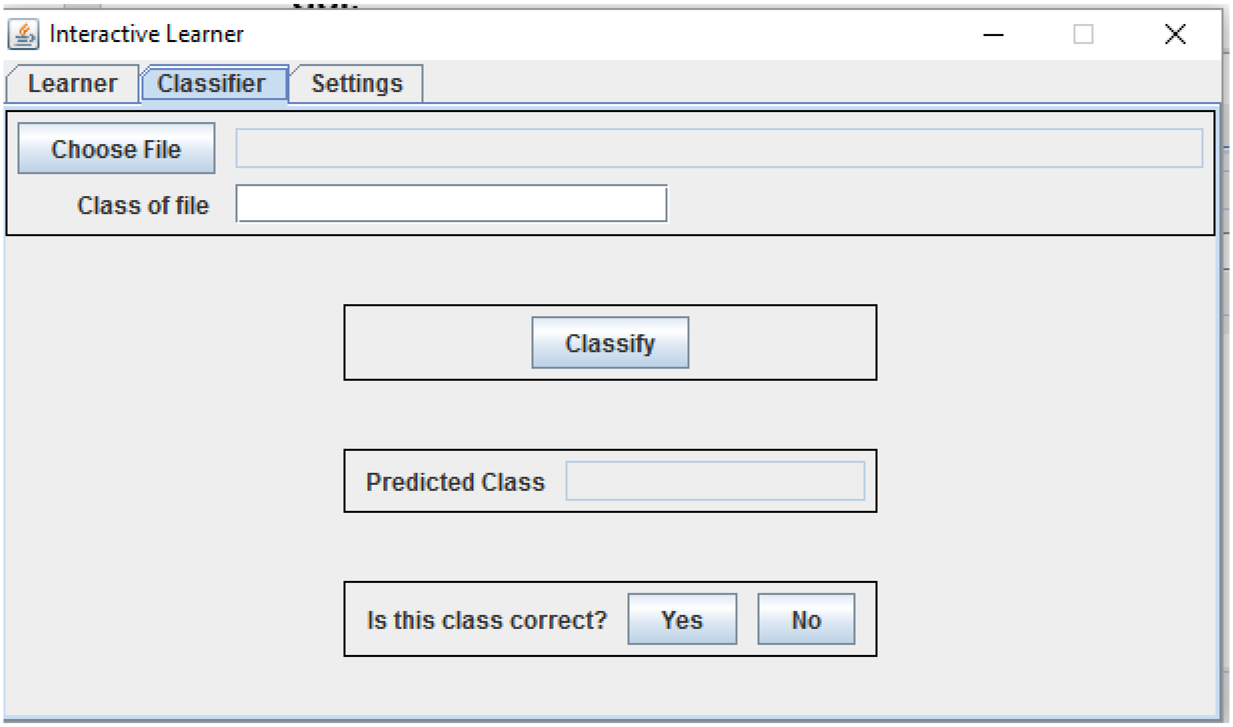


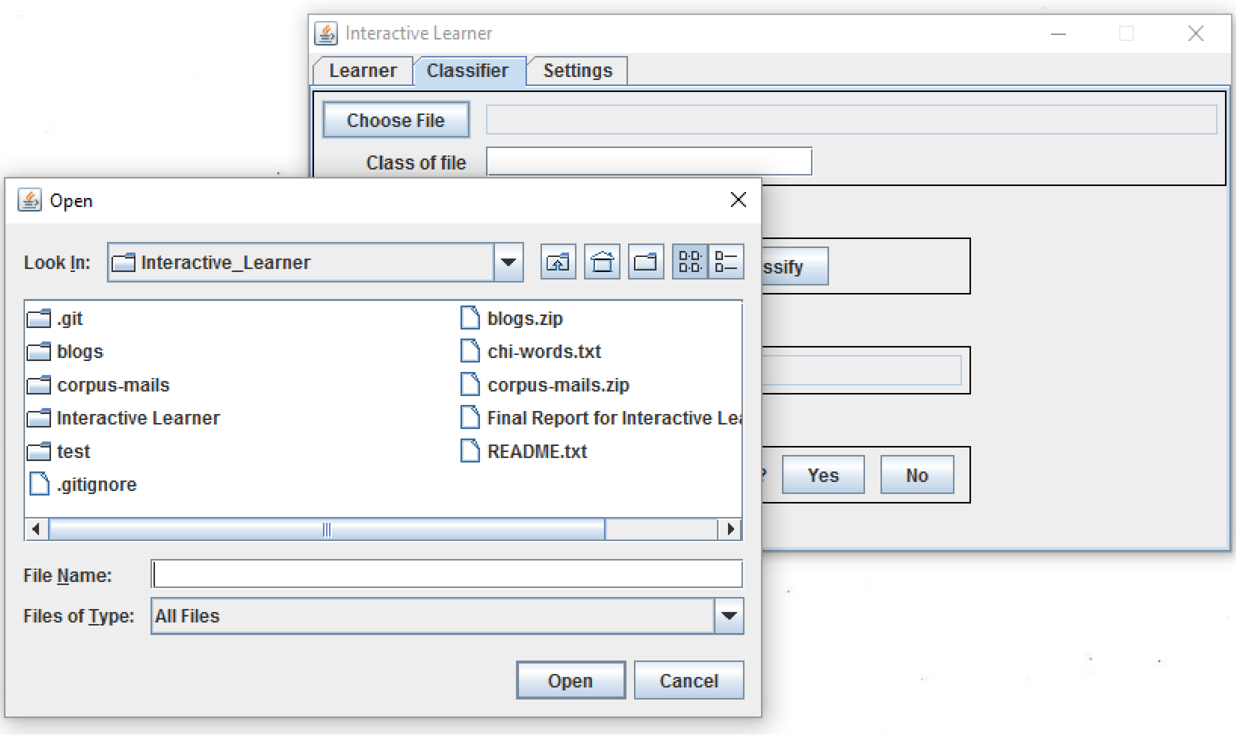
C:\Users\Gerwin\IdeaProjects\Interactive_Learner\blogs\F-train

F

```
Adding new class.... Please wait....
Added new class F
```

Only press 'Train' when you added all the classes in the corpus!

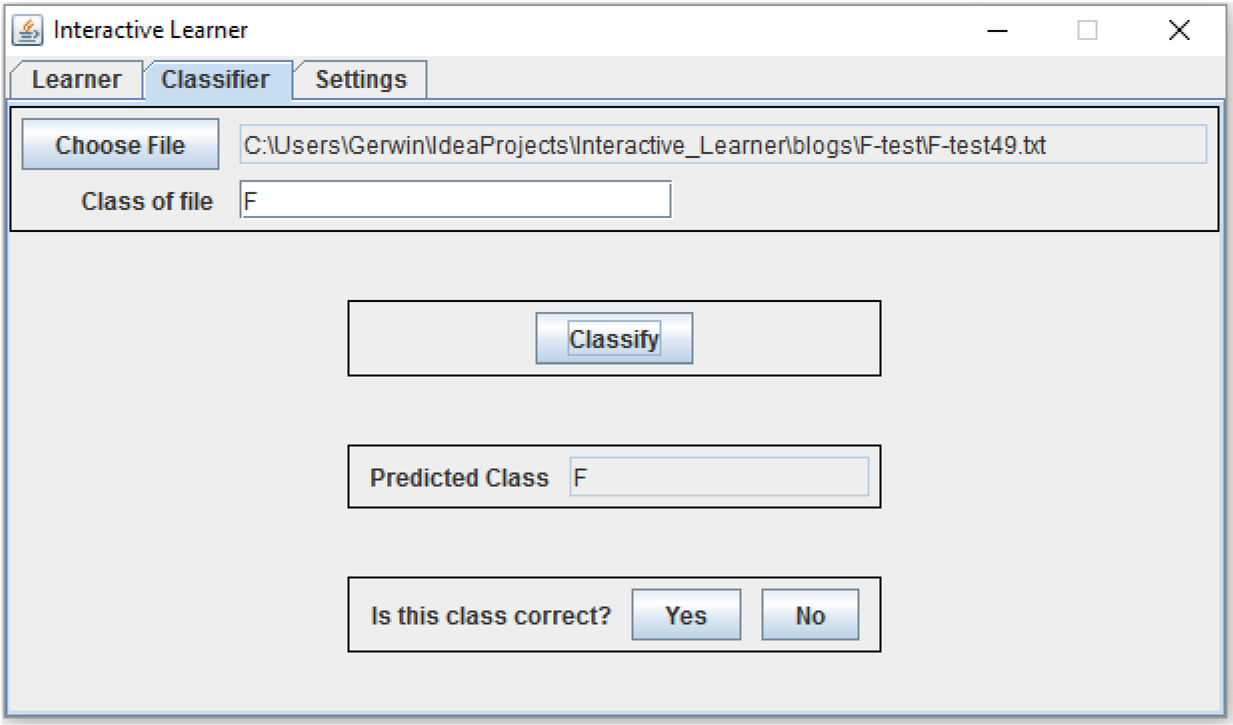




Look In: Interactive_Learner

File Name:

Files of Type: All Files

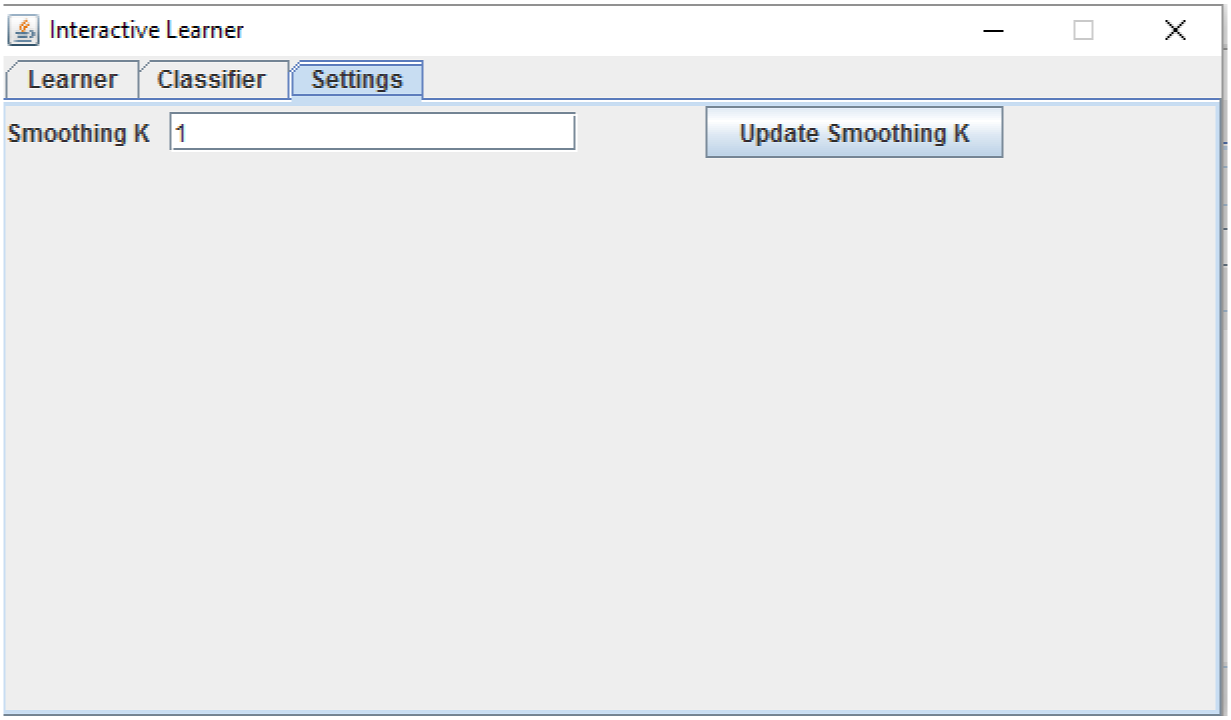


Choose File C:\Users\Gerwin\IdeaProjects\Interactive_Learner\blogs\F-test\F-test49.txt

Class of file F

Predicted Class F

Is this class correct?



Smoothing K 1