ENGR 212 Programming Practice
Mini Project 4
December 13, 2016

In this mini project you are going to develop a Python program that will automatically estimate the "main topic" of a given set of papers accepted as part of the AAAI (*Association for the Advancement of Artificial Intelligence*) Conference on Artificial Intelligence, in 2014 (you will be working with the same dataset you used in Mini Project 3), based on the title, and abstracts of each article. Besides developing a simple GUI for your program, you will also be responsible for converting the dataset provided into the proper format appropriate for your solution.

As you would remember the format of the dataset for Mini Project 3 was;

title,authors,groups,keywords,topics,abstract

You should preprocess the original dataset file, and construct the dataset for your project appropriately. Please note that there is no single solution for this step. Every project may come up with a different format for the dataset they will use. As long as the dataset you will prepare enables the correct functioning of your program, it is acceptable. Pay special attention to the below listed points while constructing your dataset:

- Since there might be more than one topic listed for each paper, select the **first** topic listed as the "main topic" to be estimated.
- Use abbreviated topic titles (i.e., APP, NMLA, etc.) rather than full topic names for the classes you will use for the rest of the project.
- Make sure that, each entry for your dataset has its associated label to be used during training & testing.
- Submit your dataset file together with your project with an explanation of each of the individual fields in your dataset.

Your program will have a graphical user interface (GUI) which will look like as shown in Figure1. Details about how it should work are given below.

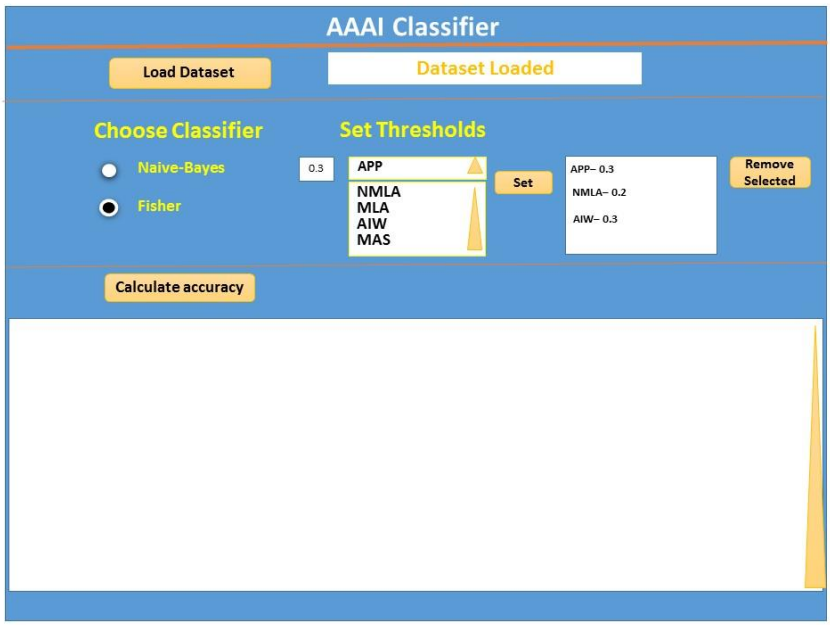

Figure 1

- Whenever the "Load Dataset" button is pressed, your program should load the dataset you prepared before from the appropriate place. The program should indicate that the data loading step is complete by the message "Dataset Loaded" in the status panel next to the "Load Dataset" button. Note that, this status should be displayed only when the loading of the data is completed. It should be empty beforehand.

- Then the user has the option to choose from the two classifiers available; either "Naïve-Bayes", or "Fisher".

- Next, optionally, the user may set thresholds for categories (i.e., topics) that will be considered during classification and estimation. Here, a listbox (initially empty) will be used to list the currently set threshold values for each topic. The user can set a threshold for a topic by selecting the topic from the combo box (located left to the listbox), and entering a value in the entry field next to the combo box. When the user clicks on the set button, the threshold should be properly set using the methods in the Python code for the classifier we developed during the lecture hours. The set threshold should be listed in the listbox as shown in Figure 1. The user should be able to select any previously set threshold in the listbox, and click on the "Remove Selected" button to remove the selected threshold from the listbox (this corresponds to setting the removed threshold to their default values, i.e., 1 for naïve-bayes, 0 for fisher classifier). If the user sets a new threshold for a topic for which the threshold is already set, the latest set threshold should be used and displayed in the listbox properly.

- Please note that the combo box should be populated automatically by your program upon loading the dataset (i.e., you should NOT hard-code any topic in your program beforehand!).

- When the user presses "Calculate Accuracy" button, your classifier shall, given a paper, try to determine the main topic it has been listed in. There are 396 papers in the original dataset; the selected classifier must use the randomly selected 300 articles for training, and the remaining 96 random articles for testing. Since you already know the main topics of each paper, calculate the accuracy of your classifier by comparing the classifier's predictions to the actual main topics present in the original dataset.

- Repeat the above step 4 times by generating different random test and training data sets, and compute the average accuracy of your classifier out of these 4 runs.

- Finally, write the average accuracy of your classifier on the output as shown below in Figure 2 (accuracy numbers here are just random numbers provided as example).

- Please note that you should list separate accuracy values for each different topic present in the dataset. That is, each topic should have a corresponding average accuracy value listed in the output frame.

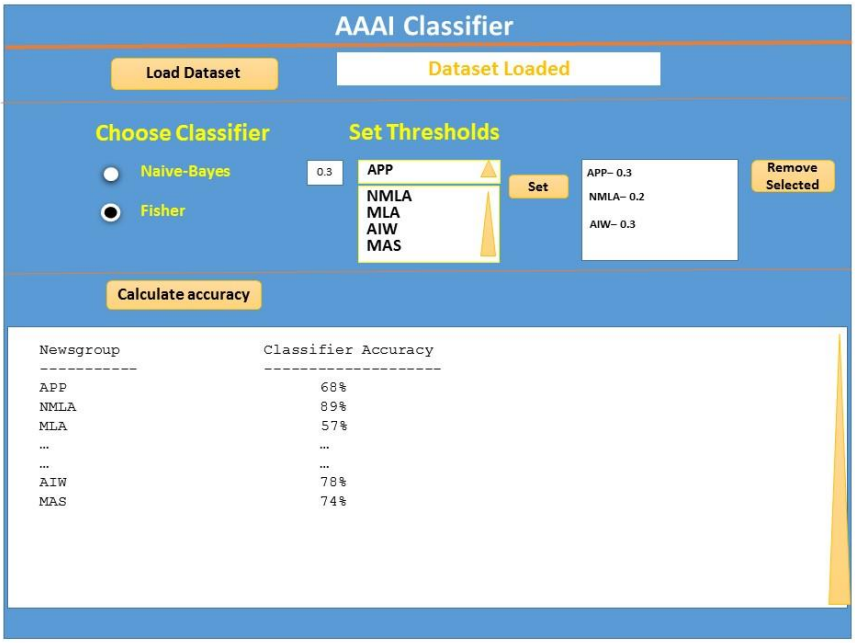

Figure 2

- The user should be able to make many experiments by changing the classifier and/or the threshold values, and recalculating the accuracy of the classifier he/she selected. He should not have to reload the dataset before each experiment session.
- Please feel free to use the codes that we covered in the class.

Warnings:

- **Do not** talk to your classmates on project topics when you are implementing your projects. **Do not** show or email your code to others. If you need help, talk to your TAs or myself, not to your classmates. If somebody asks you for help, explain them the lecture slides, but do not explain any project related topic or solution. Any similarity in your source codes will have **serious** consequences for both parties.
- Carefully read the project document, and pay special attention to sentences that involve "**should**", "**should not**", "**do not**", and other underlined/bold font statements.
- If you use code from a resource (web site, book, etc.), make sure that you reference those resource at the top of your source code file in the form of comments. You should give details of which part of your code is from what resource. Failing to do so **may result in** plagiarism investigation.
- Even if you work as a group of two students, each member of the team should know every line of the code well. Hence, it is **important** to understand all the details in your submitted code. You may be interviewed about any part of your code.

How and when do I submit my project? :

- Projects may be done individually or as a small group of two students (doing it individually is recommended). If you are doing it as a group, only **one** of the members should submit the project. File name will tell us group members (Please see the next item for details).
- Submit your own code in a **single** Python file (Do **not** include docclass.py that you import). Name it with your and your partner's first and last names (see below for naming).

- If your team members are Deniz Barış and Ahmet Çalışkan, then name your code file as deniz_baris_ahmet_caliskan.py (Do **not** use any Turkish characters in file name).
- If you are doing the project alone, then name it with your name and last name similar to the above naming scheme.

- Submit it online on LMS (Go to the Assignments Tab) by **15:00 on December 19, 2016**.

  Late Submission Policy:

  - -10%: Submissions between 17:01 – 18:00 on the due date
  - -20%: Submissions between 18:01 – midnight (00:00) on the due date
  - -30%: Submissions which are 24 hour late.
  - -50%: Submissions which are 48 hours late.
  - Submission more than 48 hours late will not be accepted.

Grading Criteria? :

Code Organization			Functionality							
Meaningful variable names (%5)	Classes and objects used (%5)	Sufficient commenting (%5)	Compiles % Runs? (10)	GUI Design (5)	Dataset preparation (15)	Loading Training Data & status update (20)	Selecting the classifier, and setting thresholds (20)	Automatically populating the combo box (15)	Accuracy Calculation (20)	Ability to make experiments without rerunning & reloading data (10)

- Interview evaluation (your grade from interview will be between 0 and 1, and it will be used as a coefficient to compute your final grade. For instance, if your initial grade was 80 before the interview, and your interview grade is 0.5, then your final grade will be 80*0.5 = 40). Not showing up for the interview appointment will **result in** grade 0.

Have further questions? :

- Contact your TA's during their office hours, and please do NOT hesitate to ask the instructor also (by e-mail or in-person), about anything related to the project.

Good Luck!