Ryan Pencak

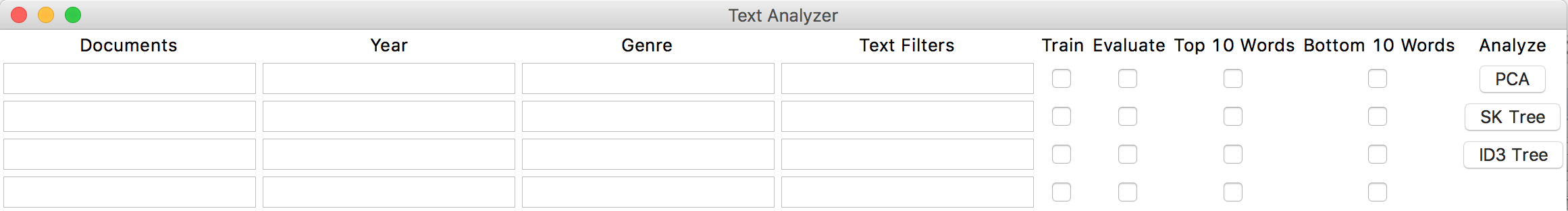
CSCI 204

Final Project Write Up

This project was designed around the concept of data mining and implements three different algorithms to analyze texts and compare them for authorship attribution. The program requires python3 to run, as well as modules numpy, pandas, matplotlib, and sklearn. To launch the program, access the project file folder using terminal and run the main file using python3. The main method file for this project is TextAnalysisMain.py.

The main method will prompt the user to enter the number of files he/she would like to analyze. The user should enter a real number that is greater than zero to represent the number of files that are to be analyzed. The main method will use this inputted number to set up and launch the user interface for the project.

The user interface will be displayed will the number of columns depending on the number of documents that are to be used. If four documents are to be analyzed, the user interface will be displayed as in Figure 1.



**Figure 1.** The user interface given that four files will be analyzed

The text entries for the documents column must all be filled in with valid documents in .txt form. The year and genre text entry columns are optional and can be used to assign year and genre attributes to the documents. The text filters text entry column is used to apply text filters to the documents. The available text filters can normalize white space, normalize case, strip null characters, strip numbers, and strip common words. These available filters can be applied by entering them as “normalizeWhitespace”, “normalizeCase”, “stripNullCharacters”, “stripNumbers”, and “stripCommonWords” respectively, without including the quotations in the entry. These filters can be entered as a list by typing them in the order that they are to be applied to the document, separated by a comma and no spaces anywhere in the string. For example, to apply the normalize white space filter, then the strip numbers filter, then the strip null characters filter, enter the following input text, again without the quotation marks: “normalizeWhitespace,stripNumbers,stripNullCharacters”. The check boxes for train and evaluate determine whether the document will be used to train the analysis method or be evaluated by the method. The check boxes for “Top 10 Words” and “Bottom 10 Words” can be checked if the analysis should be done including the top ten most frequent words or bottom ten most frequent words respectively. The last step to run the project is to click the button corresponding to the analysis algorithm that should be used to analyze the documents. The options are to use a PCA algorithm, an SK Tree Algorithm, or an ID3 algorithm.

Writing the code for this project has been a work in progress over the entirety of this semester. It was at times very comprehensive and at other times very abstract. I was able to implement data structures and programming concepts that I learned in class to complete the project or to make run times more efficient. In addition to learning how to program a linked list, a stack, or a heap, I was able to learn how these data structures can be used in different scenarios to make a method more effective or more efficient. I learned how to analyze the run time and choose the appropriate data type to accomplish a task. I also learned correct file management, as well as proper programming organization and commenting with docstring. This project has given me hands on experience with data structures and algorithms using python.

I struggled a lot with the more abstract concepts and algorithms that were to implemented in this project to analyze texts. These ideas were, for the most part, to be learned outside of class. I had a hard time filtering through online sources to find reliable instruction on creating an ID3 algorithm or completing an SK Tree or PCA algorithm. I also initially had a lot of trouble understanding how user interfaces worked in python. I had no previous experience coding a user interface and had to start from scratch, reading through the online tutorials for tkinter. I spent a lot of time researching the many methods and implementations of tkinter in python and determining exactly which ones I would need to utilize. Utilizing resources online in language documentation is an integral part of accomplishing any programming task. Another way I overcame these struggles was to reach out directly for help in office hours to get pointed in the right direction when I started going down the wrong path. The struggles I encountered getting this project to come together have allowed me to learn to become a more independent and efficient programmer. I am able to more efficiently solve problems I face in programming and debug code. I feel more confident that, given a project description, I could combine my knowledge from computer science courses as well as utilize online documentation to successfully put together a final project using correct file management and proper code organization.