Robert Gutierrez

Comp. for Poets

4/10/14

**Summery statement**

The Python program reads in an Anglo Saxon Corpus and it outputs the most unique words

**Summery**

Often time’s people have different styles of writing and tend to have different frequency of words. When looking at author’s writing it is possible to note subtle differences and words that a particular writer may like to use more frequently. The usage of this can come in helpful when validating authors work for their authenticity or perhaps check the time frame and the language most commonly used in that era. The problem is looking through multiple files and acknowledging each persons keen sense of style. With today’s technology it is much easier to tackle such problem by creating a program that reads in their work and compares it to something we know is stable such as other previous work or other text than we know its authenticity.

**Methods**

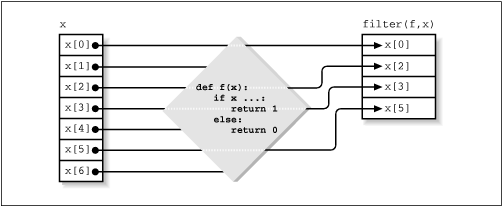
 The Python program is hard coded to take in the directory where the files are. Then specifies again to what directory to point. It declares a data path to the given work. Once established, the program begins to create a dictionary; A dictionary can be a considered a data structure that holds a “key” and a “value”. Given a traditional dictionary we search up a word or “key” in Python, and get a definition in this case a “value” in Python. This type of structure lays the foundation to our program, as it will store every word in the corpus (key) and will be assigned a number according to the times that it occurs “value”. In fig 1 we can see how the keys in the column x correspond to the cells in “filter (f,x)”, those would be considered values. In the middle there is a representation of code, which demonstrate that a series of functions can be done to a dictionary. This represents our function where we iterate through our code and check of all repeated words. Once the programmed has ran it creates a “csv” file that displays the top 10 most unique words that are used in the corpus. The “.csv” file shows the number of times it has occurred, name of Word, and Format of word. In addition the program creates a HTML file that displays similar output as the “csv "file, the HTML file only manages to display the top 3 most unique words as displayed in Table 2. This HTML file is also linked to a CSS file that holds all the styling of the HTML. There you can find the font, border-size, and background-color. Any styling modifications can be done in the CSS function where it creates and writes out to the css file.

Fig 1.

**Results**

****The Excel table in the left displays “frean” as the word of most occurrences, which happens to appear 90 times. It also displays “hrusan” as the 10th most used frequent word, which appears only 39 times. In Table 2 we have the table that is displayed in the HTML file. This Table displays the top 3 most unique words. As you can see there is an overlap with the “csv” file because they are rendering the same data.

Table 1. csv file top 10 results



Table 2. HTML page displaying top 3 results.

**Final Statement**

According to the program the highest occurrence is “frean” which happens to appear about 90 times. Then the second highest is “h&aelig;le&eth;a” which happens 68 times.