COMP 472 Assignment #2

Connor Bode -- #6281060
March 9, 2014

Running the program

1. Run **drivers.Main**
2. When the program prompts **“Feed me text!!”**, type in a sentence
3. The program will output calculations and a final educated decision on the language of the text

Training sets are loaded in **drivers.Main**

Program Details

Character Set

The set of characters which will be processed by the application is **`/[a-z\s]*/`** (English lowercase letters from a to z as well as spaces). All other characters get ignored.

Input Handling

- Input is lowercased using **String.toLowerCase()**
- Unrecognized characters are not processed, but do not cause the program to halt.
- Occurrences of members of the following set are removed from the text: **["'", "\"", ":", ",", ".", "--", "!", "?"]**
- Occurrences of members of the following set are converted to spaces: **[System.getProperty("line.separator"), "-"]**

Sample Input

Given an the input **très bien** (which contains the unrecognized character **è**), the following set of bigrams would be created: **[(t,r), (b,i), (i,e), (e,n)]**

Languages

The application currently trains for English, French, and German. German was chosen as the third language because Deustch was the first link I saw on gutenberg.org.

There are some very weighty character sequences in German that cause the application to make many German predictions for short sentences. A good example is the sequence "je". In the training text I chose, there are 380 occurrences of "j" and 102 occurrences of "je".

References

German training text retrieved from: http://www.gutenberg.org/cache/epub/9181/pg9181.txt

Originality

I, Connor Bode, certify that this submission is my original work and meets the Faculty's Expectations of Originality.