Names: Lana Abdelmohsen, Robert Helck, Casey Lishko, Alex Quezada

What was easy about this assignment?

- Since the tasks assigned built upon each other, implementing Singletons was pretty simple after implementing AllTokens. The major difference between the two authorship attribution systems is that Singleton averages the unigram probabilities of unique words, rather than all the tokens in the test file. As a result, only a few changes were required to our AllTokens algorithm to ensure only unique tokens counted towards the vocabulary size of the test file.

What was challenging about this assignment, or parts that you couldn't get working correctly?

- The hardest part of this assignment was determining what kinds of data structures we would need to organize our data in an efficient manner. Our group decided to rely a lot on Dictionaries due to their quick index accessing for retrieval and storage. The unigram probabilities of the entire imbd62.txt file were stored in one dictionary. Our group also decided to use a two-dimensional dictionary to simulate a matrix-like structure that would allow us to access and store information in almost constant time. Unigram models for all 62 authors were stored in a two-dimensional dictionary.

What did you like about this assignment?

- After hours of debugging, it was rewarding to see test files get closely matched up with the true author. The amount of extra work we did, like outputting a list of where the correct attribution for each file ranks out of 62, gave us an idea of how effective our systems were overall (not just for "33913.txt" and "70535.txt"). Although we changed this output to report the geometric mean, we deepened our understanding of Singleton and AllTokens by exploring other ways to display results.

What did you dislike about this assignment?

- We did not know the kind of rankings that a "good" attribution system would produce, so we were constantly questioning the correctness of our calculations. We now understand why AllTokens had low confidence in 70535's authorship for "70535.txt."

How did your team function? Include details regarding what each team member contributed, how the team communicated with each other, and how team software development & design was accomplished.

- Our team worked very well together. Everyone contributed ideas that were crucial to the final implementation of the program. We were able to bounce ideas off each other and use each member's strengths to advance, elevate, and test the final product. Our primary form of communication revolved around text and Zoom, while our code development was done on VSCode using Liveshare.

- Lana: Created and wrote code for T3 to develop the unigram probabilities for Add-1 and developed the AllTokens and Singletons system for T4, helped with testing and debugging in implementations of T2 and T3 and the early implementation of T4. Checked over code documentation, D4, D5, and edited the README.
- Casey: Generated random numbers for the test set (T2). Produced rankings for 33913.txt and 70535.txt (T5) and additional output that helped the group see the overall strength of the attribution systems. Assisted others in debugging incorrect geometric mean calculations (T4). Wrote D3 and the README.
- Robert: Helped to populate train files with files, helped to implement nested dictionaries for T3 and T4, debugged the code which implements the Singleton method, helped implement the AllTokens system. Helped to write comments in the code.
- Alex: Helped implement 2-D dictionaries to store and retrieve information and debug any issues that came along with this data structure for T3 and T4. Also helped debug many other issues throughout the program. Wrote D4 and contributed to documentation.

What did you learn from this assignment?

- The strength of an authorship attribution system depends largely on the sort of input it is dealing with. AllTokens was a poor association mechanism for 70535 with its test file. Its train contained repeated instances of tokens that may have ended up in other authors' test files but not its own, which we suspect led the system to rank several authors before the actual one. It did very well for 33913. Singleton attributed test files to writers based on their distinct tokens and scored both true authors highly. The difference between attribution systems is the underlying assumption that guides how we implement them. Perhaps we could design an even better system by thinking of something it could consider other than all tokens or one-time appearances.