CSCI 682 - Topics in Artificial Intelligence:
Natural Language Processing
Assignment 1

Solutions to the questions on this assignment should be submitted via PDF and .py file to Canvas before **February 1st at 11:59 pm**. Make sure to justify your answers.

I highly encourage you to collaborate with one another. However, you must write up your own solutions **independently**. Feel free to communicate via Discord and to post questions on the appropriate forum in Canvas. Do not post solutions. Also, please include a list of the people you work with at the top of your submission.

Have fun!

Problems

1. (5 pts) How many words do you know? Carefully define any ambiguous terms and explain your reasoning carefully.
2. Consider the text of *Alice's Adventures in Wonderland* by Lewis Carroll available on Canvas as alice29.txt.
   (a) (5 pts) Perform a simple tokenization using regular expressions in Python. Separate sentences based on periods (.), exclamation points (!), or question marks (?) followed by one or more whitespace characters and add explicit sentence start and end tokens. For each sentence, separate words on whitespace and remove all punctuation. Include a code snippet in your PDF submission.
   (b) (5 pts) Report the (1) total number of tokens, (2) vocabulary size, and (3) the number of tokens that appear exactly once.
   (c) (5 pts) Rank the tokens with respect to frequency. Plot token frequency against rank on a log-log scale and describe your observations.
   (d) (10 pts) Use token counts to define a unigram language model and a bigram language model. Generate 5 sample sentences using each model and compare the results. Include code snippets in your PDF submission.
   (e) (5 pts) Extra Credit - Implement the Byte-Pair Encoding (BPE) algorithm at the character level in Python. Use this implementation to tokenize the corpus using 50 merges and repeat the analysis from above. Describe your results and include a code snippet in your PDF submission.