CS4248 Assignment 1:

Regexes and Language Models

By A0235954W (Please change this as appropriate)

*This is a sample writeup.pdf file, to illustrate the expected format. You may choose to use this source file but need not to. Just follow the instructions required in the Assignment 1 instructions.*

1. **Declaration of Original Work**.

By entering my Student ID below, I certify that I completed my assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, I am allowed to discuss the problems and solutions in this assignment but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify my answers as per the Pokémon Go rule.

Signed, A0235954W

2. **References**.

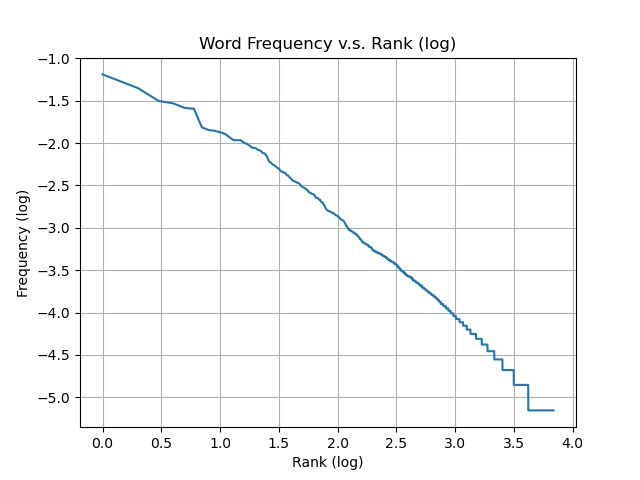
I give credit where credit is due. I acknowledge that I used the following websites or contacts to complete this assignment

* StackOverflow: <https://stackoverflow.com/> for questions on debugging.
* ‘re’ library docs: <https://docs.python.org/3/library/re.html#re.search> for queries of regular expression operations.

# Part 1. Programming

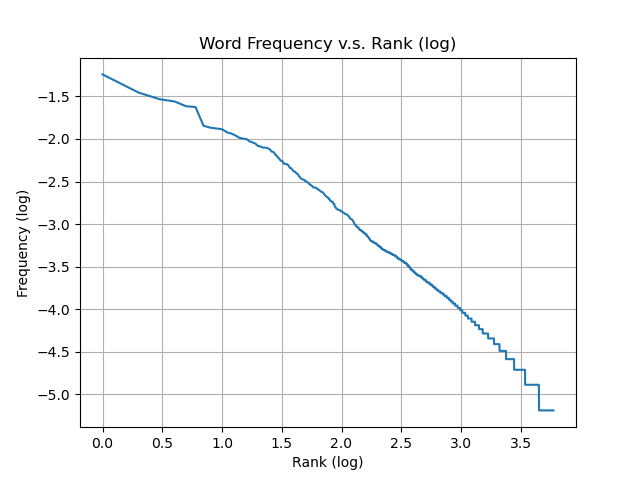
## Objective 1 — Tokenization, Zipf’s law

1. Plot here.

  
  
Justification here.

The plot of default setting is consistent with Zipf’s law because the corpus frequency of a word type is approximately inversely proportional to its rank as demonstrated in the plot above. Zipf’s law aligns to this set of tokens because they are extracted from a large body of text that follows English language and punctuations are well taken cared.

1. Plot here.

  
  
Justification here.

The plot of BPE setting is also consistent with Zipf’s law because the corpus frequency of a word type is approximately inversely proportional to its rank as demonstrated in the plot above. BPE learns a vocabulary list of sub words from a given corpus which is the loaded textual file. Since the training process follows the highest frequency of pairs of characters, it is easy to understand the relationship between the ranks and the frequencies of tokens as per large body of text. Then the stored sub words vocabulary list is used to tokenize the same given text. In addition, BPE model is tuned to generate a list of vocabulary that has a close size to the default one’s.

## Objective 2 — Regular Expressions, Edit Distance

Regex Questions.

1. Answer here.   
   R1 = (^(\S)[\S]\*\2$)
2. Answer here.   
   R2 = (^(?!.\*(\S)(\S)\3\2).\*$)
3. Answer here.   
   R3 = (?=(^(\S)[\S]\*\2)$)(?=(^(?!.\*(\S)(\S)\5\4).\*$))

R4 = (?=(^(\S)[\S]\*\2)$)(?=(^.\*(\S)(\S)\5\4.\*$))

In FSA:

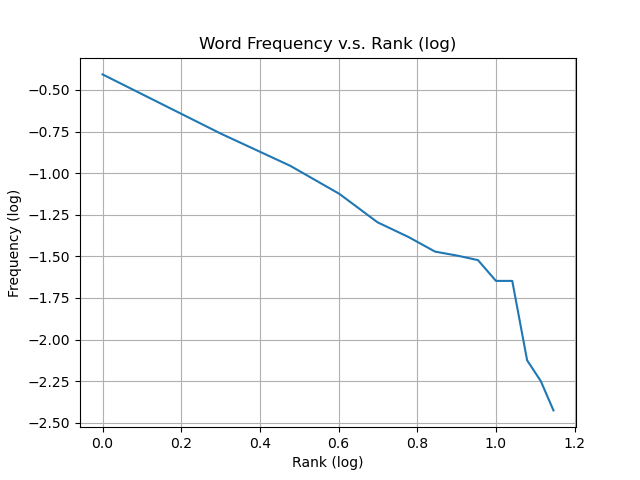
Edit Distance.

Briefly justify how you complete the Task B.

Firstly, check if the word has already aligned with regexp R3. If yes, return True/False by comparing with the value of k. If no, check if the word starts and ends with the same character. Case 1 (starts and ends with the same character): check how many patterns in the form of ‘abba’ in the word and compare this number with k. If k is larger, the target and source will be within an edit distance of k, so return True. If k is smaller, return False. Case 2 (starts and ends with different characters): replace the first character of the word by the last one to align the first and last characters. Then, check how many patterns in the form of ‘abba’ in the word and compare this number with k-1. If k-1 is larger, the target and source will be within and edit distance of k, so return True. If k is smaller, return False.

## Objective 3 — Regular Expressions, Sentiment Analysis

1. Plot here.

  
  
Justification here.

1. Compare with v.s. without emoticons (you may include some case study for illustration).   
     
   Effects of threshold on two methods.

## Objective 4 — Language Modeling

Here are some of the input-output pairs with respect to the generate\_word, generate\_text, and perplexity calls using the text corpus test\_corpus.txt (replace as needed).

...

If you implemented any additional LM variants, you’re welcomed to give high level documentation here.

# Part 2

1. **Language Models**
   1. True or False.   
        
      Justify your answer here.
   2. True or False.   
        
      Justify your answer here.
   3. True or False.   
        
      Justify your answer here.