Problem Set 5 (80 points)

1. A Matlab class called ngramcounts has been written for you. See the documentation for details on how to use this class. Write the following functions that use class ngramcounts:

   a. (20 points) Write function get_counts that processes a file to build a cell array of ngramcounts objects with the following signature:

   ```matlab
   counts = get_counts(N, filename)
   % Build counts for all N-grams up to order N
   % from the given file
   % counts is an array of ngramcounts objects.
   % counts{1} is the unigram
   % counts(Species 2000 & ITIS) is the bigram counts
   % ... counts{N}
   ```

   b. (20 points) Write function Pfrequentist that creates ngramcounts objects that use the frequentist (relative frequency) model of probability without any type of smoothing.

   ```matlab
   % ngrams = Pfrequentist(counts)
   % Given an ngramcounts object with N-gram counts,
   % return a new object that has frequentist probabilities
   % of each N-gram.
   ```

   For both problems, generate a bar chart showing the distribution of counts/probabilities for a bigram model with a context consisting of the word “the.” (These should look the same, except the vertical axis will be counts for one and probability for the other.) The function `bar` and ngramcounts method `get_values` will be useful for this.

   Blackboard files for this problem:
   - ngramcounts.m – ngramcounts class
   - train.txt – Transcripts of people talking from the Switchboard corpus (Godfrey et al., 1992). Format is one sentence per line. Add sentence markers `<s>` and `</s>` before processing.
   - example.mat – Contains counts, a cell array of ngramcounts created on train.txt.
     This can be used for checking 1a and as input for 1b if you are unable to complete 1a.

2. (40 points) Suppose that we have a four word codebook where codewords are given the names of colors (easier on humans than cw1, cw2, cw3, cw4), and the following discrete HMM has been trained:

   ```latex
   \pi = \begin{bmatrix} .7 \\ .2 \\ .1 \end{bmatrix}
   ```
Assume that the following observation sequence has been seen: (yellow, purple, blue). Answer the following questions. You may either do the work by hand (showing your work), or write a Matlab program\(^1\) to determine the likelihood of this sequence considering all paths through the model?

### Reading Assignment (100 points)

Read “Good-Turing Smoothing Without Tears” (Gale, 1994) and write a double spaced summary of the article that should not exceed two pages. Describe the problem that Gale is addressing and how he solved it. The goal of this work is to demonstrate that you understand and can present at a conceptual level the concepts of this article. As always, be careful to avoid plagiarism as it can result in you failing the class. Turn in a hard copy as well as a soft copy to TurnItIn via Blackboard.

### REFERENCES


---

\(^1\) You could be asked to solve a small problem similar to this on your final exam. If you choose to write a program to solve this (it might take you longer than doing it by hand), make sure that you are prepared to do it by hand for a very short sequence should you be asked to do so on the final.