**CS 6320.002: Natural Language Processing**
**Fall 2023**

**Homework 1 Written Component — 45 points**
**Issued 28 Aug. 2023**
**Due 11:59pm CDT 18 Sept. 2023**

**Deliverables:** Answers can be typed directly into Gradescope. See the assignment guide for more details.

**What does it mean to "show your work?"** Write out the math step-by-step; we should be able to clearly follow your reasoning from one step to another. (You can combine "obvious" steps like simplifying fractions or doing basic arithmetic.) The point of showing your work is twofold: to get partial credit if your answer is incorrect, and to show us that you worked the problem yourself and understand it. We will deduct points if steps are missing.

# 1 Language Modeling

## 1.1 Smoothing

Suppose we have a training corpus consisting of two sentences:

- The cat sat in the hat on the mat
- The dog sat on the log

Our fixed vocabulary is $V = \{\text{cat, dog, fish hat, in, log, mat, on, sat, the}\}$.

### 1.1.1 Discounting and Katz Backoff (5 points)

If we train a bigram Katz backoff model on this corpus, using $\beta = 0.75$ and no end token, what is $p_{katz}(\text{sat}|\text{dog})$?

What is $p_{katz}(\text{sat}|\text{fish})$? Note that "fish," despite not appearing in the training set, is part of the vocabulary $V$. Show your work.

### 1.1.2 Linear Interpolation (5 points)

If we use linear interpolation between a bigram model and a unigram model, using $\lambda_1 = \lambda_2 = 0.5$ and no end token, what is $p_{inter}(\text{dog}|\text{the})$? What is $p_{inter}(\text{dog}|\text{log})$? Show your work.

## 1.2 Perplexity (5 points)

What is the maximum possible value that the perplexity score can take? What is the minimum possible value it can take? Explain your reasoning and give an example of a training corpus and two test corpora, one that achieves the maximum possible perplexity

score and one that achieves the minimum possible perplexity score. (You can do this with a single short sentence for each corpus.)

### 1.3 Applications (5 points)

Authorship identification is an important task in NLP. Can you think of a way to use language models to determine who wrote an unknown piece of text? Explain your idea and how it would work (you don't need to implement it). You must use language modeling the receive credit! Other approaches do not count.

## 2 Sentiment Analysis & Classification

### 2.1 Naive Bayes — 10 points

We have a training corpus consisting of three sentences and their labels:

- The cat sat in the hat, 1
- The dog sat on the log, 1
- The fish sat in the dish, 0

**A.** Suppose we train a Naive Bayes classifier on this corpus, using maximum likelihood estimation and unigram count features without any smoothing. What are the values of the parameters $p(c)$ and $p(f|c)$ for all classes $c$ and features $f$? You can simply list the parameters and their values; no need to show the arithmetic. You can skip parameters with value 0, and you can leave your answers as fractions.

**B.** What class would our Naive Bayes classifier predict for the test sentence "The cat sat"? Show your work, ie. show the calculations for the predicted probabilities of both classes.

### 2.2 Logistic Regression — 5 points

The last step of the programming component asks you to get the top $k$ most important features for your sentiment classifier. When doing this, why do we sort by absolute value? Explain why we do this rather than sorting by the raw weight values (1-2 sentences).

## 3 Part-of-Speech Tagging — 10 points

Suppose we have a training corpus consisting of two tagged sentences:

- | The | can | is | in | the | drawer |
  |---|---|---|---|---|---|
  | DT | NN | VB | PP | DT | NN |
- | The | cat | can | see | the | fish |
  |---|---|---|---|---|---|
  | DT | NN | VB | VB | DT | NN |

**A.** Suppose we train a simple HMM part-of-speech tagger on this corpus, using maximum likelihood estimation, bigram tag transition probabilities, and a single meta-tag `<s>` (the

start tag). What are the values of the parameters $p(t_i|t_{i-1})$ and $p(w_i|t_i)$ for all tags $t$ and words $w$? You can simply list the parameters and their values; no need to show the arithmetic. You can skip parameters with value 0, and you can leave your answers as fractions.

**B.** What parts of speech would the trained HMM tagger in the previous problem predict for the test sentence "The fish can see the can," using Viterbi decoding? Show your work, ie. the dynamic programming table $V$. You can leave your answers as fractions.