**CS 6320.002: Natural Language Processing**
**Spring 2020**

**Homework 1 Written Component – 50 points**
**Issued 19 Aug. 2020**
**Due 11:59pm CDT 02 Sept. 2020**

**Deliverables:** Answers can be typed directly into Gradescope. LaTeX can be hand typed or generated using Mathpix Snip. 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 Math Review

The problems in this section refresh your memory on concepts from classes you have taken previously that we will use later in this course.

## 1.1 Multivariate Calculus

### 1.1.1 Partial Derivatives (5 points)

$$f(x, y, z) = \frac{x}{y^2} + ze^{x^2}$$

What are $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$? Show your work.

### 1.1.2 The Chain Rule (5 points)

$$\begin{aligned} f(x, y) &= xg(x, y) + 2y \\ g(x, y) &= x^2y - xh(x^2, y) \\ h(x, y) &= xy^2 + 5 \end{aligned}$$

What are $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$? Show your work.

### 1.1.3 Extrema (5 points)

$$f(x) = x \log_2(x) + (1 - x) \log_2(1 - x)$$

What are the values of $x$ corresponding to the minima and maxima of $f(x)$ for $x \in [0, 1]$? Show your work (your math work; graphing it doesn't count!).

## 1.2 Probability and Statistics

### 1.2.1 Conditional Probability (5 points)

Suppose there is a box containing 10 balls; five are black, and five are white. You remove three balls at random, without replacing any. What is the probability that you remove three black balls? Show your work.

Let $B$ and $W$ denote the events that you draw a black or white ball, correspondingly.

### 1.2.2 Bayes's Rule (5 points)

Suppose you have two lab-mates. One (Friend A) talks about computer science 80% of the time, and linguistics 20% of the time; the other (Friend B) talks about linguistics 80% of the time, and computer science 20% of the time. One day, you find a typed note on your desk about computer science. Your lab-mates leave you notes equally often, so you don't know who left this one. What is the probability the note is from Friend A? Show your work.

Let $FA$ and $FB$ denote the events that Friend A or Friend B wrote the note, correspondingly, $CS$ denote the event that the note talks about computer science, and $LI$ denote the event that the note talks about linguistics.

# 2 Language Modeling

The problems in this section are based on the material covered in Week 2.

## 2.1 Smoothing

Suppose we have a training corpus consisting of two sentences:

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

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

If we train a bigram Katz backoff model on this corpus, using $\beta = 0.75$, what is $p_{katz}(\text{sat}|\text{dog})$? What is $p_{katz}(\text{sat}|\text{fish})$? Show your work.

### 2.1.2 Linear Interpolation (5 points)

If we use linear interpolation between a bigram model and a unigram model with $\lambda_1 = \lambda_2 = 0.5$, what is $p_{inter}(\text{dog}|\text{the})$? What is $p_{inter}(\text{dog}|\text{log})$? Show your work.

## 2.2 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.)

## 2.3 Generation (5 points)

Use your code from the programming component of this assignment to train three language models on the provided data file, `shakespeare.txt`: one unigram model, one trigram, and one 5-gram. For each model, generate 5 random sentences with `max_length=10`. Show the sentences you generated with each model.

What are some problems you see with the generated sentences? How do the sentences generated by the different models compare with each other?

## 2.4 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).