Midterm for 6.864

Name: ________________________

10	25	35	30	45	Total out of 150

Good luck!

Part #1 10 points

We define a PCFG where the non-terminal symbols are $\{S, A, B\}$, the terminal symbols are $\{a, b\}$, and the start non-terminal (which is always at the root of the tree) is S. The PCFG has the following rules:

Rule	Probability
$S \rightarrow AB$	0.2
$S \rightarrow AA$	0.3
$S \rightarrow BS$	0.5
$A \rightarrow a$	0.7
$A \rightarrow b$	0.3
$B \rightarrow a$	0.2
$B \rightarrow b$	0.8

Question 1 10 points

For the input string *abab*, draw two possible parse trees. Show how to calculate the probabilities of those trees.

Part #2 25 points

Question 2 (5 points)

We train a trigram language model using add-α smoothing on a large corpus composed of Wall Street Journal (WSJ) articles from 2003. Plot the shape of the probability of your **training** corpus under the resulting language model as a function of α for $0 \leq \alpha < \infty$.

Question 3 (5 points)

We now test this language model on WSJ articles from 2004. Plot the probability of your **test** corpus under this language model as a function of α.

Question 4 (5 points)

Plot the *perplexity* on the test corpus under this language model as a function of α.

Question 5 (10 points)

Let V be the size of the vocabulary and W be the size of the test corpus, both in number of words. As $\alpha \to \infty$, what value does the perplexity on the test set approach? Explain your answer.

Part #3 35 points

Consider the task of segmenting a word into prefix, suffix, and stem. For simplicity, assume that every word has *exactly* one prefix, one stem, and one suffix, in that order. Each of the three kinds of word segment is likely to have different distributional characteristics (i.e., use different character sequences) than the other kinds of segments.

Question 6 10 points

Describe the parameters you would use for a probabilistic model for this task. Assume that the choices of prefix, suffix, and stem are independent, e.g., selecting a particular prefix does not affect the distribution over stems.

Question 7 5 points

Show how you would compute the *maximum likelihood* estimate for these parameters, given a corpus of segmented words with prefixes, suffixes, and stems identified.

Question 8 5 points

Compute the likelihood of the **segmented** word "a-bc-de" in terms of the parameters described above (where "a" is a prefix, "bc" a stem, and "de" a suffix).

Question 9 5 points

Compute the likelihood of the **unsegmented** word "abcde" in terms of the parameters described above (i.e., the segmentation is unknown).

Question 10 10 points

Assume now a language in which the choice of suffix depends heavily on the stems. Describe how you could change the parameterization of the model to account for this dependency. Compute the likelihood of the segmented word "a-bc-de" in terms of the parameters described above (where "a" is a prefix, "bc" a stem, and "de" a suffix).

Part #4 30 points

The *Kumaraswamy distribution* is a continuous distribution over the interval $[0, 1]$. It resembles a Beta distribution and is sometimes used instead of the Beta because of its simple analytical form. Its pdf $p(x)$ and cdf $h(x)$ are defined as:

$$p(x) = abx^{a-1}(1 - x^a)^b$$
$$h(x) = 1 - (1 - x^a)^b$$

The variables a and b are real-valued hyperparameters and must be positive. Note that the pdf given here is properly normalized.

The distribution's mode (where it takes its highest value) is

$$x = \left(\frac{a-1}{ab-1}\right)^{1/a}$$

For this problem, assume always that $a = 2$ and $b = 2$.

NB: You do not have to compute out square roots, *e.g.*, no need to simplify expressions like $\frac{3}{\sqrt{5}}$.

Question 11 (10 points)

Given a random number z that is uniformly distributed between 0 and 1, derive the function $f(z)$ that outputs a random draw from the Kumaraswamy distribution.

Question 12 (10 points)

Alternatively, we could sample this distribution via rejection sampling. Using the proposal distribution $q(x) = 1$ for $0 \leq x \leq 1$, compute the lowest possible scaling factor k that makes this proposal distribution amenable for rejection sampling.

Question 13 (10 points)

On average, what is the acceptance rate of your rejection sampler, *i.e.,* what percentage of the time will the rejection sampler accept a sample from the proposal distribution?

Part #5 — 45 points

Consider the probabilistic morphological segmentation model that we discussed in class. Assume that you are provided with parameters Θ induced using the MAP training procedure. We would like to apply this model to perform morphological segmentation on new test data. We assume that the likelihood of a segmentation of the word w into morphemes $m_1, \ldots, m_l$ is given by the following formula:

$$P(m_1, \ldots, m_l) = \prod_{j=1}^{l} \Theta(m_j)$$

You may assume that the set of possible morphemes includes all possible substrings of w, *i.e.*, $\Theta(m_j)$ exists for all substrings.

Your goal is to find the best segmentation using a dynamic programming algorithm.

Question 14 — 5 points

Describe the structure of your dynamic programming table π – specify its dimensions and describe the content of cells in this table.

Question 15 5 points

Specify the base case for your computation.

Question 16 10 points

Specify the recursive case for your computation.

Question 17 5 points

For a word of length n, what is the time complexity of your algorithm?

Question 18 10 points

Suppose that you want to eliminate segmentations that contain more than 5 segments. Describe any changes necessary to the dynamic programming table to implement this constraint. Specify the new base and recursive cases of your algorithm.

Question 19 10 points

Alternatively, suppose that you allow any number of segments, but want to restrict each morpheme to be at most 6 characters. Again, describe any changes necessary to the dynamic programming table to implement this constraint. Specify the new base and recursive cases of your algorithm.