1. (19 points) Assume a bigram language model is trained on the following corpus of sentences using MLE with linear interpolation for smoothing (with the bigram \( \lambda \) weight set to 0.9 and the unigram \( \lambda \) weight set to 0.1). Since the unigram model does not need to estimate \( P(<s>) \), just completely ignore the start token when estimating the unigram model.

- \(<s>\) man marries woman \(</s>\)
- \(<s>\) woman marries man \(</s>\)
- \(<s>\) woman marries woman \(</s>\)
- \(<s>\) man divorces woman \(</s>\)
- \(<s>\) woman divorces man \(</s>\)

What is the estimated probability of the following test string? Show your work. You only need to calculate the parameters of the model sufficient to solve this particular problem.

- \(<s>\) man marries man \(</s>\)
2. (19 points) Consider the HMM below where the transition probabilities are shown in the graph and the observation probabilities (where $V=\{A,B\}$) are in the tables below each state.

Use the Forward algorithm to compute the probability of generating the short output string: “A B”.
Show the values computed for each of $\alpha_t(j)$ parameters as they are computed, showing your work.
3. (19 points) Given the following context-free grammar, draw trees for all parses produced for the sentence: “Guard tests like gold.” Under each tree, briefly paraphrase in English the meaning of each parse as best you can and circle the most likely interpretation.

\[
\begin{align*}
S & \rightarrow NP \ VP, \ S \rightarrow VP, \\
NP & \rightarrow Adj \ NP, \ NP \rightarrow N, \ NP \rightarrow NP \ PP, \\
VP & \rightarrow V, \ VP \rightarrow V \ NP, \ VP \rightarrow VP \ PP \\
PP & \rightarrow Prep \ NP, \\
Prep & \rightarrow like, \\
N & \rightarrow guard, \ N \rightarrow gold, \ N \rightarrow tests, \\
V & \rightarrow guard, \ V \rightarrow tests, \ V \rightarrow like, \\
Adj & \rightarrow gold, \ Adj \rightarrow guard
\end{align*}
\]
4. (19 points) Consider the following simple PCFG:

\[
\begin{align*}
S & \rightarrow NP \ VP \quad 0.6 \quad Pronoun \rightarrow I \quad 0.5 \\
S & \rightarrow VP \quad 0.4 \quad Pronoun \rightarrow YOU \quad 0.5 \\
NP & \rightarrow Pronoun \quad 0.4 \quad PlurNoun \rightarrow FLIES \quad 0.5 \\
NP & \rightarrow PlurNoun \quad 0.6 \quad PlurNoun \rightarrow ARROWS \quad 0.5 \\
VP & \rightarrow Verb \quad 0.2 \quad Verb \rightarrow LIKE \quad 0.6 \\
VP & \rightarrow Verb \ NP \quad 0.2 \quad Verb \rightarrow FLIES \quad 0.4 \\
VP & \rightarrow VP \ PP \quad 0.6 \quad Prep \rightarrow LIKE \quad 0.4 \\
PP & \rightarrow Prep \ NP \quad 1.0 \quad Prep \rightarrow WITH \quad 0.6
\end{align*}
\]

Use the probabilistic CKY algorithm (the Viterbi analog for PCFGs) to find the most probable parse tree for the sentence:

FLIES LIKE ARROWS
First, below show the changes to productions that require conversion to CNF:

Next, show the triangular CKY table with each cell filled with all its constituents together with their probabilities, showing your work.

Finally, show all final parse trees for this sentence (in the original grammar) with their probabilities.
5. (24 points) Provide short answers (1-3 sentences) for each of the following questions:

Consider the following joke:

There are two fish in a tank. One says to the other, “How do you drive this thing?”

Explain what specific type of ambiguity in language understanding makes this humorous.

Do the same for the following famous dialog from the disaster-movie spoof “Airplane!”:

Rumack: You’d better tell the Captain we’ve got to land as soon as we can. This woman has to be gotten to a hospital.
Elaine Dickinson: A hospital? What is it?
Rumack: It’s a big building with patients, but that’s not important right now.

What two colloquial “catch phrases” best characterize the difference between 1) a generative model and 2) a discriminative model?

Given feature vector $X$ and a class variable $Y$, what is difference between the probability distribution that is effectively estimated for a discriminative versus a generative model?
What is “smoothing” and why is it critical to many areas of statistical language processing?

What is the basic idea behind a “back-off” approach to smoothing?

Why does CKY parsing take $O(n^3)$ time, where $n$ is the length of the sentence?

Semantic similarity between the distributional representations of two words can be computed in different ways, for example using Euclidean distance or cosine. Which of the two is preferable, and why?

(Extra credit, 2pts) After becoming a billionaire working for a hedge fund (Renaissance Technologies), Robert Mercer was awarded the ACL “Lifetime Achievement Award” in 2014. Where did he work as a statistical NLP researcher before he left for Wall Street?
(Extra credit, 2pts) Mercer has recently donated millions of dollars to which current presidential candidate?

(Extra credit, 2pts) What famous sentence did Noam Chomsky use to illustrate that people can make grammaticality judgements even for sentences whose semantics is very unclear?