Name: _____
CS159: Natural Language Processing
PReview #12
Due with the final exam

Write a regular expression that matches the course codes for all of the CS courses offered at the 5Cs. You can see all of the current offerings at `https://jics1.hmc.edu/ICS/Portal_Homepage.jnz?portlet=Course_Schedules`, and should assume that future course codes will match the same format, though new course numbers might be introduced and courses might be offered at different campuses in the future.

■

If you enter the sentence "I know he knows the head of the bank. " into Google Translate, it generates the Spanish translation "Yo sé que él sabe que el jefe del banco ." If you then translate that sentence *back* to English, it generates the sentence "I know he knows that the head of the bank."
Using what you've learned about the parts of a machine translation system, explain what went wrong in the translation. What sort of ambiguity was relevant here? Why did the MT system not translate it correctly? What sorts of modifications to the translation system might be helpful in correcting it?

■

(From Jurafsky and Martin's NLP textbook): Suppose someone took all the words in a sentence and reordered them randomly. Explain how you could write a program that takes as input such a bag of words and produces as output a guess at the original order. You can assume you have access to an n-gram language model. How would you use the Viterbi algorithm to design a computationally-tractable solution?

■

Give a valid POS tagging of the following sentence, by adding tags in the form "word/POS":

```
The     better     better     better     best     the     rest.
```

Explain how the assumptions of an HMM would or would not allow a viterbi decoder to correctly tag this sentence, given sufficient training data.

■

Consider the confusion matrix below, which gives the results of a (completely fabricated) system for predicting sentiment in movie reviews. Each row indicates an actual label, and each column indicates a system-hypothesised label. For example, this table says that there were 5 reviews that were really neutral but that the system hypothesized were positive.

	Positive	Neutral	Negative
Positive	86	15	3
Neutral	5	30	2
Negative	2	20	57

Under this scheme, comment on the types of errors that the system seems most prone to. What is it good at? What might the system's engineers focus on improving? How might your answer vary based on the intended use of the system?

■

If a movie company is interested in knowing more about when and why people don't like their movies, they might be willing to group "positive" and "neutral" reviews into the same category. This would reduce the task to a binary classification task with the goal of flagging "negative" reviews. For this characterization of the task, give the precision and recall of the system described above.

■

Write a context-free grammar in chomsky normal form that can generate (at least) the sentences below:

```
The cat sleeps.
The mouse eats the cheese.
Tom steals the cheese from the mouse.
```

■

Draw one resulting parse tree for each of the sentences above.

■

Could any sentences have more than one parse under the grammar you generated? If so, give an example. If not, justify how you know that each sentence has at most one parse.

■