Honor Code & Collaborating with classmates

For the final exam (not including the special instructions for the bonus portion), you may not work with your fellow classmates outside of discussions on Piazza. On Piazza, you'll be able to ask and answer questions, similarly to what we can ask and answer for homework.

Carefully read the following piazza question guidelines:
1. Questions like "what is this question asking about" or "can you clarify X question" are welcome and encouraged.
2. Questions about previous quiz/assignment questions examples are welcome and encouraged.
3. In general, Piazza questions/answers should not include final exams answers or partial answers.
We won't be monitoring Piazza 100% constantly for the next week, so if you don't know if it's allowed, make a private post and we will let you know.

Formatting Guidelines

You will turn in ***one pdf per section.*** Each pdf should be formatted as follows:

Your name here
Final exam section number here

Question 1: An example essay question
(your answer here, which may include images, example code, and math as helpful, make sure to cite your sources)

Question 2: An example numerical question (work required)
(your work here, which must be included for full credit and may include images, example code, etc, make sure to cite your sources)

Your answer here, highlighted in yellow and bold

If you are working in a Jupyter Notebook, you can use the following in a Markdown cell to produce this effect: <mark style="background-color: yellow; font-weight:bold">Your answer here</mark>

Question 3: An example multiple choice or true/false question (no work required)

Your answer here, highlighted in yellow and bold

If you are working in a Jupyter Notebook, you can use the following in a Markdown cell to produce this effect: <mark style="background-color: yellow; font-weight:bold">Your answer here</mark>

Citations:
(The citations relevant to your answers for this section here. You may cite the textbook, slides provided in class, youtube videos, other notes provided in class, or any other sources you used to answer this question.)

Including code in your answers

You may answer any question that you like using example code. When you use example code, you must write this code yourself. You are allowed to use any code that you have written for this class in previous assignments, quizzes, etc where relevant.

No questions require that you write code, though you may find it useful for many questions.

Section 1: Data & Pre-processing (10 points)

1. Describe in 2 - 3 sentences what a vocabulary is for an NLP system and what an unknown word is for NLP systems. Next, make an argument (2 - 3 sentences) about whether dealing with unknown words is relevant for [**no**, **some**, or **all**] NLP tasks.

2. Name one effect the following two differences in tokenization might have on a downstream task and explain why it would happen (2 - 3 sentences).

sentence 1: When I was a child, I wasn't allowed to watch Sesame Street!
sentence 2: When I was a child , I was n't allowed to watch Sesame Street !

Multiple Choice

3. **True** or **False:** Given infinite training data, a language model will not have to deal with unknown words.

Section 2: Statistical Models (70 points)

1. For the following data, answer the given questions.

   data:
   <s> Summer is close </s>
   <s> I am excited for summer </s>
   <s> I am sad ski season ended </s>
   <s> I am sad winter ended </s>
   <s> Swimming season is close </s>
   <s> I am excited for swimming soon </s>

   a. What is the frequency of unknown words in the following training data if we choose a vocabulary before training? You may assume that your vocabulary includes <s> and </s> tokens.

   vocabulary:
   am
   for
   I
   is
   sad
   season
   spring
   summer
   end

   b. Give an example generated sentence where:
   - the sentence is not in the training data but the underlying words are
   - it could be generated using Shannon's method for generation

2. Using a Naive Bayes classifier, with bag-of-words features, what is P(y = spam | x) and what label will your classifier assign the sentence "remember to give Big Bird your password today !" given equal priors for each class and the following table of learned probabilities?

	spam	not spam
remember	.27	.10
to	.01	.01
give	.16	.27
your	.20	.29
password	.11	.21

3. You are a geologist studying historical earthquakes in Hawaii. You don't have access to seismic data from a particular year but you do know how many earthquake insurance packages were purchased in each of the months. Your job is to use the insurance data to predict whether an earthquake happened in a given month or not

What is the probability of "earthquake, no earthquake, earthquake" for the first three months of the year given the observation 10, 30, 10 insurance packages bought, and the following data?

pi = [.8, .2]
P(10 | earthquake) = .2
P(20 | earthquake) = .4
P(30 | earthquake) = .4
P(10 | no earthquake) = .5
P(20 | no earthquake) = .4
P(30 | no earthquake) = .1
P(earthquake | no earthquake) = .4
P(no earthquake | no earthquake) = .6
P(earthquake | earthquake) = .7
P(no earthquake | earthquake) = .3

4. Is using the Viterbi algorithm equivalent to taking the argmax of each column in a probability table? why/why not? Give an example sequence of observations that demonstrates your argument.

5. Why is a logistic regression classifier considered a discriminative model? (2 – 3 sentences)

6. Can we use the same algorithm to train an MEMM as we used for Logistic Regression? Why/why not? (2 – 3 sentences and/or example code)

7. In an HMM, we directly model transition probabilities between states at time *i* and *i - 1*. Is this transition probability modeled in an MEMM? If yes, how? If no, why not?

8. You are designing a logistic regression classifier to assign movie reviews to the classes **positive** or **negative**.

Training data (each review is from a different movie):

positive: A film bursting at the seams with sheer, unadulterated joy: watch it, and the world seems just that little bit brighter...

negative: A film that seeks to inspire our loathing for nearly every frame it sits on the screen, and that is before we even attempt to decipher the mediocrity that is the music or the nonexistence of the story.

positive: feminist, foul-mouthed and funny, turning the formulaic tropes of bawdy comedies inside out and giving us a couple of teen heroines who feel real and very 2019.

a. Give 4 different features that you would use in your sentiment analysis system.
b. Convert the given training data into the features they would be represented as.
c. Report the assigned label from your mini classification system using the following already trained weights and the given testing data.

Testing data (each review is also from a different movie):

It doesn't do anything new or even terribly distinctive, but maybe it didn't have to. It just had to be good enough to stick the landing, and it does that.

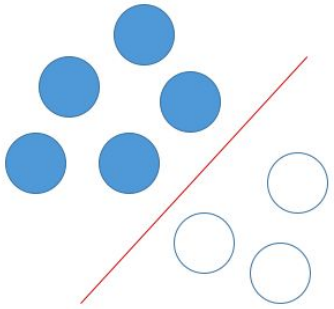
I think the film is a reflection of a most unpleasant mind, a mean, sly, sadistic little mind.

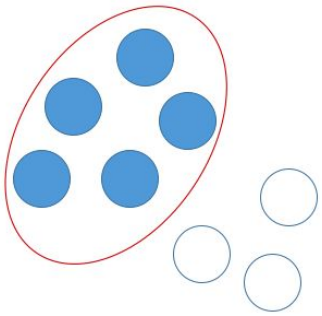
weights: [0.5, 1, -0.5, -1], bias term: 0.25

Multiple Choice

9. In a training set with examples **X** and corresponding labels **Y**, a generative model would seek to learn _____ from the training data: (Choose all that apply)
    a. P(X)
    b. P(X|Y)
    c. P(Y|X)
    d. P(Y)

10. Which of the following shows a distribution that a generative model might learn? (Choose one)

A. 

B. 

11. Select all of the ways in which we can deal with the unknown word problem during the training step of a language model: (Choose all that apply)
    a. Convert all low count words to <UNK> during training
    b. Choose a vocabulary, convert all words not in vocabulary to <UNK>
    c. Convert all words with high character counts to <UNK> during training
    d. Randomly convert .05 of the words in the training set to <UNK>
    e. Nothing special needs to be done - language models trained on enough data are robust enough to not encounter problems with unknown words

12. In a language model, we want to _____ (maximize/minimize) perplexity and _____ (maximize/minimize) dev/test set probability.

13. What is it called when the weights for logistic regression features fit the training set perfectly? (Choose one)
    a. Generalization
    b. Regularization
    c. Overfitting
    d. Entropy

14. Does an MEMM for POS tagging generate multinomial or binary classifications? (Choose one)
    a. Multinomial
    b. Binary

15. Must an MEMM use a sigmoid or softmax classification function? (Choose one)
    a. MEMMs must use the Sigmoid function
    b. MEMMs must use the Softmax function
    c. The classification function that an MEMM uses is task-dependent

Section 3: Neural Models (15 points)

1. How many weights must be trained for a Feed-forward Neural Network that does the sentiment classification task from question 8 in the Statistical Models section?
    - Assume that your FFNN has two hidden layers, the first layer with 2 hidden units, the second layer with 3 hidden units.
    - Include bias terms for the input and each hidden layer.
    - Use the same input features that you used for question 8.

2. Explain how words are represented when used in neural language models (2 – 3 sentences) **and** how these representations are produced using the skip-gram algorithm (2 - 3 sentences).

3. Explain how an RNN with 1 hidden layer that uses “regular” hidden units (hidden units with a single non-linearity) can perform POS tagging. Make sure to address all of the following:
    a. How many weight matrices must be trained for the network
    b. What are the inputs to each layer
    c. What operation(s) is/are happening in the hidden layer
    d. What operation(s) is/are happening in the output layer
    e. How POS tags are assigned to a given sentence, assuming greedy decoding

Multiple Choice

4. Which of the following is a Feed-forward Neural Network most similar to? (Choose one)
    a. N-gram Language Model
    b. Naïve Bayes Classifier
    c. HMM
    d. Logistic Regression Classifier
    e. MEMM

5. **True** or **False**: an RNN performing POS tagging of the last word in a sentence is influenced by the first word in the sentence.

6. For skip-gram word embeddings, the dimensionality is _____ (dependent on/independent from) the size of the vocabulary of the training data.

Section 4: Machine Translation & Transfer learning (20 points)

1. Explain how the E-M algorithm works for IBM Model 2.
    a. How is the model initialized?
    b. What does the E-step involve (specific to IBM Model 2)?
    c. What does the M-step involve (specific to IBM Model 2)?
    d. Why is using the E-M algorithm necessary?

2. Explain how the encoder-decoder architecture works for neural machine translation with no attention used by answering the following questions.
    a. What is the output of the encoder portion of the network? (both at each timestep and from the encoder portion as a whole)
    b. What is the output of the decoder portion of the network? (at each timestep)
    c. Which words from the input sentence is the decoder portion "aware of"?
    d. What is one way to improve a neural machine translation model that does not use attention.

3. Explain how transfer learning can be used (2 – 3 sentences) **and** what one important limitation that it faces is (2 – 5 sentences).

Multiple Choice

4. How is a contextualized word embedding different than a "regular" word embedding? (choose the best single answer)
    a. contextualized word embeddings are less sparse
    b. contextualized word embeddings capture polysemy better
    c. contextualized word embeddings are higher dimensionality
    d. contextualized word embeddings are trained using the CBOW algorithm

5. Which of the following are common examples of a source task for transfer learning? (Choose all that apply)
    a. copy task
    b. edit task
    c. language modeling
    d. machine translation
    e. NER recognition
    f. POS tagging
    g. sentiment analysis
    h. sentence end detection
    i. summarization task

Section 5: Bias (10 points)

1. What is one way that NLP systems are biased? Give an example grounded in a specific task, using a specific model that we covered in this class, then explain why the bias that you've identified might occur (use model specifics) and explain why it is a difficult problem to solve?

Section 6: Bonus (up to +20 points)

The criteria/rubric bonus section will be released on 4/27/2020 and will be an application of an NLP model to art. You will have complete creative freedom and will be allowed to use any python libraries and data sources that you choose.