CS224N: Natural Language Processing
Spring 2008
Weekly Quiz

Quiz 1

1: [Conditional Probability] Which of the following is always true? There are two random variables X and Y.

- $SUM_x P(x \mid y) = 1$
- $SUM_y P(x \mid y) = 1$
- $SUM_{x,y} P(x \mid y) = 1$
- all of the above
- none of the above

2: [Perplexity] Suppose a language model assigns the following conditional n-gram probabilities to a 3-word test set: 1/4, 1/2, 1/4. Then P(test-set) = 1/4 * 1/2 * 1/4 = 0.03125. What is the perplexity?

- 2.5
- 1.5
- 2.828
- 0.75
- 3.175

3: Which one of the following is true?

- After smoothing, a probability distribution might not sum up to 1 anymore
- Entropy of a discrete random variable is always non-negative
- The problem of add-one smoothing is that unseen events don't get enough probability mass.
- P(A,B) = P(A)P(B)
- None of the above

4: For the following questions, assume we are using a corpus completely summarized by the unigram counts below (thus V = 20):

Unigram counts:	
brown	29
fox	34
lazy	18
dog	1

plenty	41
tree	1
skim	4
neat	49
syzygy	33
missing	12
napkin	9
cheap	22
fork	10
nickel	1
chocolate	5
syrup	9
short	28
options	13
car	14
concinnity	0
SUM	**333**

What are the following probabilities? (answer as a fraction or a whole number (e.g., "1/2" or "1")):

P_{MLE}(short) =

- 28/333
- 13/333
- 28/353
- 28/334

5: P_{MLE}(concinnity) =

- 1/333
- 1/353
- 1/334
- 0

6: Now assume we are using Laplace smoothing. What are the following probabilities? $P_{Laplace}$(lazy) =

- 18/353
- 19/353
- 19/334
- 18/334

7: $P_{Laplace}$(concinnity) =

- ☐ 0
- ☐ 1/353
- ☐ 1/333
- ☐ 1/334

8: Now assume we are using Good Turing smoothing. What probability mass do we assign to things with zero frequency in our training data?

- ☐ 1/111
- ☐ 0
- ☐ 3/353
- ☐ 3/20

9: Now assume that the above counts were just drawn from a larger corpus, and on that full corpus we collected the following counts of counts:

N_1	112849
N_2	41018
N_3	15608
N_4	5704
N_5	2111
N_6	754
N_7	283
N_8	104
N_9	37
N_{10}	14

With this data, what are the smoothed counts c^* for the following words (assuming the unigram counts given above are still valid)?

c^*(skim) =

- ☐ 844/5704
- ☐ 22816/15608
- ☐ 10555/5704
- ☐ 10555/15608

10: c^*(syrup) =

- ☐ 126/104
- ☐ 126/37
- ☐ 140/37
- ☐ 333/104