Name:
Number:

CENG464 Quiz 2

1.We are given the following corpus:
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>
Using a bigram language model with add-one smoothing, what is P(Sam | am)? Include <s> and </s> in your counts just like any other token. (35 pts)

$C(\text{Sam}|\text{am}) = 2$
$C(\text{am}) = 3$
V (representing total word types including <s> and </s>) = 11
$P_{\text{laplace}}(\text{Sam}|\text{am}) = \frac{C(w_{n-1}, w_n)+1}{C(w_{n-1})+V} = \frac{2+1}{3+11} = \frac{3}{14}$

2. Assume the following likelihoods for each word being part of a positive or negative movie review, and equal prior probabilities for each class.

	pos	neg
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

What class will Naive bayes assign to the sentence "I always like foreign films."? (35 pts)

$P(\text{I always like foreign films.}|\text{pos}) = 0.09 * 0.07 * 0.29 * 0.04 * 0.08 = 0.000005846$

$P(\text{I always like foreign films}|\text{neg}) = 0.16 * 0.06 * 0.06 * 0.15 * 0.11 = 0.000009504$

Neg class

3. How can we evaluate our classification model ? Give three metrics and explain them. (30)

-> Precision = $\frac{tp}{tp+fp}$
-> Recall = $\frac{tp}{tp+fn}$
-> F1-measure = $\frac{2PR}{P+R}$