CSE 5522 Artificial Intelligence II

Homework #8: Naive Bayes Classification

Wei Xu, Ohio State University

Your Name: ____________________ OSU ID: ________________

1. **Language Identification**. The Naive Bayes model has been famously used for text classification. In this case, we will use it in the bag-of-words model to determine the language of Twitter posts:

   - Each tweet has binary class label C which takes values in $\{sp, en\}$. The sp stands for Spanish, en stands for English.
   - For a tweet with n words $t_1, \ldots, t_n$, its label is predicted by
   $$\arg\max_c P(C = c | t_1, \ldots, t_n) = \arg\max_c P(C = c) \prod_{i=1}^{n} P(W = t_i | C = c)$$
   - Each word t of a tweet, no matter where in the tweet the word occurs, is assumed to have probability $P(W = t|C)$.

   You are given four tweets as a training set, and one new tweet to classify:

		Tweet	Class
Training	#1	English Wikipedia editor	en
	#2	free English Wikipedia	*en*
	#3	Wikipedia editor	*en*
	#4	español de Wikipedia	*sp*
Test	#5	Wikipedia español el	??

   (a) What values would you estimate for the maximum likelihood parameters for the Naive Bayes model, if not using any smoothing? (Note: Only the parameters that would be involved in the prediction for tweet #5 are listed here.)

$\hat{P}(C)$

en	
sp	

$\hat{P}(W = t|C = en)$

Wikipedia	
español	
el	

$\hat{P}(W = t|C = sp)$

Wikipedia	
español	
el	

What is the probability of tweet #5 being predicted as English or Spanish by this Naive Bayes model?

$P(en|\text{Wikipedia}, \text{español}, \text{el}) =$

$P(sp|\text{Wikipedia}, \text{español}, \text{el}) =$

(b) You are training with the same tweets, but now doing Laplace Smoothing with strength $k = 1$. Re-estimate the parameters. How will this new Naive Bayes model will classify tweet #5?

$\hat{P}(C)$

en	
sp	

$\hat{P}(W = t|C = en)$

Wikipedia	
español	
el	

$\hat{P}(W = t|C = sp)$

Wikipedia	
español	
el	

$P(en|\text{Wikipedia}, \text{español}, \text{el}) =$

$P(sp|\text{Wikipedia}, \text{español}, \text{el}) =$