Name: ____________________________
CS159: Natural Language Processing
PReview #4
Due in class Feb 16, 2015

Review

If you randomly pick a word from the Brown corpus “mystery” category, what is the probability that the word is at least six letters long? What if you pick a word from the “hobbies” category? The “government” category?
Use the fact that there are 57169 words in the mystery category, 82345 words in the hobbies category, and 82345 words in the government category to calculate the marginal probability of a word chosen at random being at least six letters long if we combined all three of the above categories.

∎

Again using the Brown corpus “government” category, what is the probability that a word w_i is at least six letters long given that the previous word w_{i-1} is at least six letters long? Given that result, what can you conclude about the (in)dependence of the lengths of consecutive words in that data set?

∎

Preview

Suppose that we have the following data points that we want to cluster with the k-means algorithm:

x	**y**	assignment 1	assignment 2
1	4		
2	5		
2	0		
1	8		
0	2		
3	2		
3	5		
4	6		
5	4		

For clustering into two groups, we will have two means. Suppose that we initialize them at (1,0) and (5,6):

iteration	center 1	center 2
0	(1,0)	(5,6)
1		
2		
3		

Step through three iterations of the k-means algorithm. In the top table, indicate whether each point is assigned to cluster 1 or cluster 2 in each iteration. In the bottom table, indicate what the newly-computed means are for each cluster. Use the rest of this page to show enough of your work for us to understand how you got your results.

■