Homework 5

GOAL	Demonstrate understanding of clustering and classification concepts
DEADLINE	Thursday, November 14th– by the beginning of class
TOTAL POINTS	85 points
HINT	Including intermediate calculations will help me give you better feedback if something goes wrong 😉

PROBLEM 1 (10 POINTS)

Provide an example of how you would use clustering to support IR-related tasks. What are the features (beyond simple word tokens) that you would use to represent objects you want to cluster? What measure you would use to estimate the similarity between any two objects? How would you evaluate the outcome?

PROBLEM 2 (10 POINTS)

Assume that you want to do a classification using a very fine-grained hierarchy, such as one describing all the families of human languages. Suppose that, before training (or probability computation, depending on the classification model), you decide to collapse all of the labels corresponding to Asian languages into a single "Asian Language" label. Discuss the negative consequences/drawbacks of this decision.

PROBLEM 3 (10 POINTS)

Nearest neighbor clusters are not symmetric, in the sense that if instance A is one of instance B's nearest neighbors, the reverse is not necessarily true. Explain how this can happen. Hint: use a diagram to illustrate your point.

PROBLEM 4 (25 POINTS)

Given the following table, use Laplacian smoothing to:

- ***a)*** Estimate a Multinomial Naive Bayes classifier and apply the classifier to the test document, i.e., Doc ID = 5.
- ***b)*** Estimate a Multiple-Bernoulli Naive Bayes classifier and apply the classifier to the test document, i.e., Doc ID = 5.

Show ***all the intermediate steps*** (e.g., word probability values as needed, intermediate computations, etc.) for full credit.

	Doc ID	*Words in Document*	*Class*
Training Set	1	Barcelona Vienna Lisbon	World
	2	Lisbon Paris Madrid	World
	3	Barcelona Girona Bilbao	Spain
	4	Barcelona Bilbao Madrid	Spain
Test Doc	5	Barcelona Bilbao Bilbao	?

PROBLEM 5 (15 POINTS)

Using documents 1, 4, and 7 as the initial centroids, perform a K-means clustering for the documents in the following table; K = 3 and the distance is determined by the cosine similarity measure.

Doc ID	*Document Text*
1	Hot chocolate cocoa beans
2	Cocoa Ghana Africa
3	Beans harvest Ghana
4	Cocoa butter
5	Butter truffles
6	Sweet chocolate
7	Sweet sugar
8	Sugar cane Brazil
9	Sweet cake icing
10	Sweet sugar boot
11	Cake black forest

PROBLEM 6 (15 POINTS)

Consider the following two-dimensional instances. Find the K-nearest neighbor of instances (1,-1) and (1,1), for K = 3 and distance computed using the Euclidian metric. What is the cost of joining the 2 clusters if you use single linkage? And if you use average linkage?

(-4,-2), (-3,-2,), (-2,-2), (-1,-2), (1,-1), (1,1), (2,3), (3,2), (3,4), (4,3)