ASSN 8: SIMILARITY COEFFICIENTS

A. A pizza restaurant (John's Pizza, Bleeker St, NYC) offers only a fixed set of possible combinations of ingredients.
1) For the given pizza data ('pizza_bin.txt'), discuss which similarity coefficient you think would be most appropriate to measure the similarity of the pizzas. The main considerations are 1) respecting the measurement level of the variables and 2) defining similarity in the most "meaningful" way. The eight binary variables represent presence/absence of eight ingredients (cheese, tomato sauce, anchovies, onions, sausage, mushroom, peppers, meatballs).
2) Using the measure you select, calculate the proximities among pizzas 1, 2 and 3 (these three observations are shown below).

p#	ch	ts	an	on	sa	mu	pe	me
01	1	1	0	0	0	0	0	0
02	0	1	1	0	0	0	0	0
03	1	1	1	0	0	0	0	0
..								

B. For Fisher's iris data ('iris_mlt.txt'),
1) discuss which similarity measure you think would be most appropriate for distinguishing between TYPES of irises (that means the measure should tend to make irises of the same type look similar and irises of different types look different). The four variables are: sepal length, sepal
width, petal length, petal width, (The fifth variable is a category code variable, = 1 for *setosa*, 2 for *versicolor*, and 3 for *virginica*. This variable is not to be used in the clustering – it represents *a priori* info that we usually would not have access to in doing a clustering. We might use it here to try to validate our clustering).
2) Using the measure you selected, calculate the proximities among the first specimen of each type (shown below):

spec	sl	sw	pl	pw	c	name
01	51	35	14	02	1	set
51	70	32	47	14	2	ver
101	63	33	60	25	3	vig

C. For the demographic data on African countries in the file 'africa_mlt.txt' (sample lines below), propose an appropriate measure of the similarity between countries. Discuss why this is an appropriate or the most appropriate measure. [OPTIONAL: if the data were in the form shown in the additional provided file 'Africa demographics.pdf', how would your answer change? In this document, we have additional variables containing numeric codes for dominant ethnic group, dominant religion, languages spoken, etc.]

Country	area	popM	grow	life	lit %ed	labr	%ag	%ot	GDP	GDPgr	perc	p%ag	p%in	p%ot	imprt	imptUS	exprt	expUS	USaid	date
Angola	481351	8	2.7	38	20-99	1.9	60	40	4.2	0	550	29	27	44	1500	103	1600	1010	1.9	1975
Benin	43483	4	3.1	41	20-43	1.5	70	30	1.1	-4.2	310	35	16	49	590	13	304	0.3	0.8	1980
Botswana	220000	1.1	3.3	50	30-93	0.4	75	25	0.7	0	750	11	1	88	740	19	640	57	11.4	1966
Burkina	106000	6.9	2.5	42	5-8	2.7	83	17	0.9	-1.3	157	35	20	45	230	21	110	0.1	15.6	1960
Burundi	10747	4.8	2.6	42	25-29	1.9	93	7	1.2	3	255	51	15	34	198	9	79	2	6	1962
Cameroon	183568	9.8	2.7	47	65-70	3	83	17	6.7	5	734	30	9	61	1100	66	1904	721	20.5	1960