KEY -- (2020) -- ASSN 8: SIMILARITY COEFFICIENTS

A. For the pizza data ('pizza_bin.txt'), discuss which similarity coefficient you think would be most appropriate (the main considerations are 1) respecting the measurement level of the variables and 2) defining similarity in the most "meaningful" way).
Using the measure you select, calculate the proximities among pizzas 1, 2 and 3 (these three observations are shown below). The eight variables represent presence/absence of eight ingredients (cheese, tomato sauce, anchovies, onions, sausage, mushroom, peppers, meatballs).

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
..								

These are binary variables, hence an "association coefficient" might be appropriate. But which one? In my judgment, two pizzas are similar to the extent that they both have an ingredient (but not so much because they both DON'T have an ingredient) – that is, the ABSENCE of an ingredient does not seem so informative. Therefore, Jaccard's coefficient might be best – it does not count the (0,0) cell in the 2 x 2 association matrix: S = a/(a+b+c).

To compute the Jaccard coefficient between pizza #1 and #2, go across the first two rows above, counting the number of joint 1's, joint 0's and mismatches of each type:

Freq: P2:

P1:	1	0
1	1	1
0	1	5

so S_{12} = a/(a+b+c) = 1/(1+1+1) = 0.333

For pizzas #2 and #3,

Freq: P3:

P2:	1	0
1	2	0
0	1	5

so S_{23} = a/(a+b+c) = 2/(2+0+1) = 0.667

Also, S_{13} = a/(a+b+c) = 2/(2+0+1) = 0.667

B. For Fisher's iris data ('iris_mlt.txt'), 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). 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
```

One can argue that profile similarity is the most relevant type of measure here (because size of a flower might merely indicate maturity), so the correlation between two rows might be appropriate. On the other hand, my botany might be in error, so Euclidean distance seems like a reasonable answer, too.

We could perhaps investigate this question empirically, by investigating which type of proximity measure is more highly correlated with the TRUE partition of these specimens into the three varieties.

Using Euclidean distance, the distance between the first two specimens above is: $d(1,51) = SQRT[(51-70)^2+(35-32)^2+(14-47)^2+(2-14)^2] = 40.04$
Etc.

C. For the demographic data on African countries ('africa_mlt.txt'), 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 attached 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

If we include the additional (nominal-level) variables containing the category codes for ethnic group, etc., then I see two reasonable approaches: 1) pick a similarity measure that explicitly allows for variables of different measurement level (i.e., Gower's coefficient), or 2) do some fancy dummy coding of the categorical variables, then use a numeric similarity / distance measure (e.g., Euclidean).

If we include only the variables shown in the file (see table above), then we might get by with a numeric-based measure of similarity (e.g., Euclidean distance or correlation), with appropriate variable standardization, to deal with the problem that the variables abovce are on wildly different scales.