CSE 20 Spring 2022
Homework 1 solution
Due date: Wednesday, Apr 6, 2022 at 11:59pm

1. (12 points) For each set, list out the elements in roster notation: (you do not need to show your work.)

   (a) $\{\texttt{AA}, \texttt{A}, \texttt{C}\} \circ \{\texttt{AA}, \texttt{A}\}$

      Solution:
      $\{\texttt{AAAA}, \texttt{AAA}, \texttt{AA}, \texttt{CAA}, \texttt{CA}\}$

   (b) $\{\texttt{AA}, \texttt{A}, \texttt{C}\} \times \{\texttt{AA}, \texttt{A}\}$

      Solution:
      $\{(\texttt{AA}, \texttt{AA}), (\texttt{AA}, \texttt{A}), (\texttt{A}, \texttt{AA}), (\texttt{A}, \texttt{A}), (\texttt{C}, \texttt{AA}), (\texttt{C}, \texttt{A})\}$

   (c) $\{(\texttt{A}, \texttt{C}), \texttt{A}\} \times \{\texttt{C}, (\texttt{A}, \texttt{U})\}$

      Solution:
      $\{((\texttt{A}, \texttt{C}), \texttt{C}), ((\texttt{A}, \texttt{C}), (\texttt{A}, \texttt{U})), (\texttt{A}, \texttt{C}), (\texttt{A}, (\texttt{A}, \texttt{U}))\}$

   (d) $\{x \in \mathbb{N} \mid 9 \leq x^2 < 50\}$

      Solution:
      $\{3, 4, 5, 6, 7\}$

   (e) $\{x^2 \mid x \in \mathbb{N} \textit{ and } 9 \leq x^2 < 50\}$

      Solution:
      $\{9, 16, 25, 36, 49\}$

   (f) $\{\lambda, 00\} \circ \{1, 11, 111\}$

      Solution:
      $\{1, 11, 111, 001, 0011, 00111\}$

2. (10 points) Recall the recursively defined set S of all RNA strands over the bases in the set $B = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{U}\}$:

   Basis Step: $\texttt{A} \in S, \texttt{C} \in S, \texttt{U} \in S, \texttt{G} \in S$
   Recursive Step: If $s \in S$ and $b \in B$ then $sb \in S$

   Exercise: Recursively define a function *AminusC* from the set S to the set $\mathbb{N}$ that computes the number of As minus the number of Cs. For example, $AminusC(AGGUCUC) = -1$

   Your function needs to follow the format of a recursive function. Namely,

   Basis Step: $<$ *your basis step* $>$
   Recursive Step: $<$ *your recursive step* $>$

   Solution:

   Basis Step:

   - $AminusC(\texttt{A}) = 1$
   - $AminusC(\texttt{C}) = -1$
   - $AminusC(\texttt{G}) = 0$
   - $AminusC(\texttt{U}) = 0$

   Recursive Step: If $s \in S$ then

   - $AminusC(s\texttt{A}) = AminusC(s) + 1$
   - $AminusC(s\texttt{C}) = AminusC(s) - 1$
   - $AminusC(s\texttt{G}) = AminusC(s)$

- $AminusC(s\mathtt{U}) = AminusC(s)$

3. (15 points) Recall the way we encoded Netflix ratings in class. Let's say that Netflix only has 4 titles:

Bridgerton, Squid Game, Ozark, Tiger King.

Then for example, a user with the 4-tuple: $(-1, 1, -1, 0)$ is a user who does not like Bridgerton or Ozark, likes Squid Game and they are indifferent to Tiger King. (maybe they haven't seen it.)

For example:

Suppose Netflix has 6 Users:

User 1:	(1,	0,	-1,	0)
User 2:	(-1,	-1,	1,	-1)
User 3:	(1,	1,	0,	1)
User 4:	(0,	-1,	0,	0)
User 5:	(-1,	-1,	1,	1)
User 6:	(-1,	-1,	-1,	-1)

In class we learned about a way to compare two users by taking the *Euclidean distance* of their 4-tuples.

We can apply the same logic to compare two titles.

Exercise 1: (4 points) For each title, create a 6-tuple where the first entry is the rating of User 1, the second entry is the rating of User 2 and so on.

Solution:

- Bridgerton: $(1, -1, 1, 0, -1, -1)$
- Squid Game: $(0, -1, 1, -1, -1, -1)$
- Ozark: $(-1, 1, 0, 0, 1, -1)$
- Tiger King: $(0, -1, 1, 0, 1, -1)$

Exercise 2: (4 points) Which pair of Titles are "closest" and which pair of Titles are "furthest" from each other using the *dist* metric? (show all 6 calculations.)

Solution:

- Bridgerton and Squid Game:

$$dist((1, -1, 1, 0, -1, -1), (0, -1, 1, -1, -1, -1)) =$$

$$\sqrt{(1-0)^2 + (-1-(-1))^2 + (1-1)^2 + (0-(-1))^2 + (-1-(-1))^2 + (-1-(-1))^2} =$$

$$\sqrt{1+0+0+1+0+0} = \sqrt{2} \approx 1.41$$

- Bridgerton and Ozark:

$$dist((1, -1, 1, 0, -1, -1), (-1, 1, 0, 0, 1, -1)) =$$

$$\sqrt{(1-(-1))^2 + (-1-1)^2 + (1-0)^2 + (0-0)^2 + (-1-1)^2 + (-1-(-1))^2} =$$

$$\sqrt{4+4+1+0+4+0} = \sqrt{13} \approx 3.61$$

- Bridgerton and Tiger King:

$$dist((1, -1, 1, 0, -1, -1), (0, -1, 1, 0, 1, -1)) =$$

$$\sqrt{(1-0)^2 + (-1-(-1))^2 + (1-1)^2 + (0-0)^2 + (-1-1)^2 + (-1-(-1))^2} =$$

$$\sqrt{1+0+0+0+4+0} = \sqrt{5} \approx 2.24$$

- Squid Game and Ozark:

$$dist((0,-1,1,-1,-1,-1),(-1,1,0,0,1,-1)) =$$
$$\sqrt{(0-(-1))^2+(-1-1)^2+(1-0)^2+(-1-0)^2+(-1-1)^2+(-1-(-1))^2} =$$
$$\sqrt{1+4+1+1+4+0} = \sqrt{11} \approx 3.31$$

- Squid Game and Tiger King:

$$dist((0,-1,1,-1,-1,-1),(0,-1,1,0,1,-1)) =$$
$$\sqrt{(0-0)^2+(-1-(-1))^2+(1-1)^2+(-1-0)^2+(-1-1)^2+(-1-(-1))^2} =$$
$$\sqrt{0+0+0+1+4+0} = \sqrt{5} \approx 2.23$$

- Ozark and Tiger King:

$$dist((-1,1,0,0,1,-1),(0,-1,1,0,1,-1)) =$$
$$\sqrt{(-1-0)^2+(1-(-1))^2+(0-1)^2+(0-0)^2+(1-1)^2+(-1-(-1))^2} =$$
$$\sqrt{1+4+1+0+0+0} = \sqrt{6} \approx 2.45$$

Bridgerton and Squid Game are the “closest” titles and Bridgerton and Ozark are the “furthest” titles.

Another way to compare two titles is by using the *dot product* of their 6-tuples. The dot product is computed by multiplying corresponding entries and adding up the products. For example:

$$dot((-1,1,-1,0,1,0),(1,0,0,1,-1,1)) =$$
$$(-1)(1)+(1)(0)+(-1)(0)+(0)(1)+(1)(-1)+(0)(1) = -2$$

Exercise 3: (3 points)

Compute the dot product of Squid Game and Tiger King.

Solution:

$$dot((0,-1,1,-1,-1,-1),(0,-1,1,0,1,-1)) =$$
$$(0)(0)+(-1)(-1)+(1)(1)+(-1)(0)+(-1)(1)+(-1)(-1) = 0+1+1+0-1+1 = 2$$

We can use the dot product to define another measure of how similar two titles are. It is called the cosine similarity and for two n-tuples v, w it is defined as:

$$cossim(v,w) = \frac{dot(v,w)}{\sqrt{dot(v,v)}\sqrt{dot(w,w)}}$$

For example, the cossim of $(-1,1,-1,0,1,0),(1,0,0,1,-1,1)$ is:

$$cossim((-1,1,-1,0,1,0),(1,0,0,1,-1,1)) =$$
$$\frac{dot((-1,1,-1,0,1,0),(1,0,0,1,-1,1))}{\sqrt{dot((-1,1,-1,0,1,0),(-1,1,-1,0,1,0))}\sqrt{dot((1,0,0,1,-1,1),(1,0,0,1,-1,1))}} =$$
$$\frac{-2}{\sqrt{4}\sqrt{4}} = -1/2$$

Exercise 4: (3 points)

Compute the cosine similarity of Squid Game and Tiger King (you can use your calculation from the previous exercise.)

Solution:
For this calculation, we have to find the dot product of Squid Game and Tiger King by themselves.

$$dot(SG, SG) = dot((0, -1, 1, -1, -1, -1), (0, -1, 1, -1, -1, -1)) =$$
$$= (0)(0)+(-1)(-1)+(1)(1)+(-1)(-1)+(-1)(-1)+(-1)(-1) = 0+1+1+1+1+1 = 5$$

$$dot(TK, TK) = dot((0, -1, 1, 0, 1, -1), (0, -1, 1, 0, 1, -1)) =$$
$$= (0)(0) + (-1)(-1) + (1)(1) + (0)(0) + (1)(1) + (-1)(-1) = 0 + 1 + 1 + 0 + 1 + 1 = 4$$

$$cossim(SG, TK) = \frac{dot(SG, TK)}{\sqrt{dot(SG, SG)}\sqrt{dot(TK, TK)}} = \frac{2}{\sqrt{5}\sqrt{4}} \approx 0.45$$

The distance function and the cosine similarity function are two different ways to measure how similar two user ratings are.

Exercise 5: (for fair effort completeness:) (1 point)

In your own words, what are the differences and similarities of these two functions? Which one do you like better for the task of comparing two titles and why?

Solution:
The distance function will always give you non-negative number and the closer you are to 0, the closer the two tuples are. This is a good way to compare a single title with a bunch of other titles to see which is the "closest" match.
The cossim function gives you a value in the range $[-1, 1]$ where values near 1 are very similar. The cossim function is a little more telling than the dist function when it comes to titles that are "different". If the cossim is close to -1, then that means that the titles are close to "opposite". If title A and title B have a cossim close to -1 then you can say with confidence that if a person likes title A, they are likely to not like title B and vice versa.
For this reason, the cossim may give you more useful information if you are trying to recommend titles to different users.

4. (10 points) Color in computer is often represented as a 3-tuple (r, g, b) where r represents the red component, g represents the green component and b represents the blue component. Each component r, g, b must be an integer between 0 and 255 (inclusive.)

   For example, a nice lavender color is $(199, 176, 252)$. White is $(255, 255, 255)$ and black is $(0, 0, 0)$.

   Let $A = \{x \in \mathbb{N} \mid 0 \leq x \leq 255\}$ be the range of each of the rgb components.

   Let's say that you have 8 different paint colors:

A	(103,	221,	21)
B	(28	,102,	71)
C	(164,	110,	70)
D	(67,	12,	119)
E	(187,	70,	124)
F	(148,	79,	150)
G	(226,	230,	42)
H	(27,	188,	202)

Think about how you would describe a measure of "closeness" of two colors. Out of the 8 colors, using your description would you say that A and B should be considered the "closest" pair? What is a function you can use to quantify this idea?

Let's consider the cossim formula for two colors v, w:

$$cossim(v,w) = \frac{dot(v,w)}{\sqrt{dot(v,v)}\sqrt{dot(w,w)}}$$

Exercise 1: (4 points)

It looks like the two green colors (A and B) should be closest. Compute the cossim of these two colors.

$$cossim(A,B) = \frac{dot(A,B)}{\sqrt{dot(A,A)}\sqrt{dot(B,B)}} =$$

$$\frac{dot((103,221,21),(28,102,71))}{\sqrt{dot((103,221,21),(103,221,21))}\sqrt{dot((28,102,71),(28,102,71))}} =$$

$$\frac{(103)(28)+(221)(102)+(21)(71)}{\sqrt{103^2+221^2+21^2}\sqrt{28^2+102^2+71^2}} = \frac{26917}{\sqrt{59891}\sqrt{16229}} \approx 0.86$$

Exercise 2: (4 points)

Consider the two purply colors D and F. Compute the cossim of these two colors.

$$cossim(D,F) = \frac{dot(D,F)}{\sqrt{dot(D,D)}\sqrt{dot(F,F)}} =$$

$$\frac{dot((67,12,119),(148,79,150))}{\sqrt{dot((67,12,119),(67,12,119))}\sqrt{dot((148,79,150),(148,79,150))}} =$$

$$\frac{(67)(148)+(12)(79)+(119)(150)}{\sqrt{67^2+12^2+119^2}\sqrt{148^2+79^2+150^2}} = \frac{28714}{\sqrt{18794}\sqrt{50645}} \approx 0.93$$

Exercise 3: (for fair effort completeness) (1 point)

How do you interpret the results of Exercise 1 and Exercise 2? according to the results, would you say the two greens are closer or the two purples are closer and why?

Solution:
The cossim of the two purples is closer to 1 than the cossim of the two greens to according to this metric, the two purples are closer.
To me, the two greens are both "green" but I would say that D is purple and F is mauve. So since they have different names, I would probably consider them to be farther away. But mathematically according to the cossim, they are closer.

Exercise 4: (for fair effort completeness) (1 point)

Is the Euclidean distance function a better or worse way to compare colors? Why or why not?

Solution:
The distance function will give you a different meaning of "close".
For example, if you compare the dist of a light green and a dark green, then they will be far apart because of the light and the dark. On the other hand, the cossim of a light green and a dark green will be close to 1 because they have the same "hue" meaning that if you plot the two points in three dimensional space, they will be "pointing" in roughly the same direction.
I think that if you are concerned with "hue" (the actual color not necessarily if it is dark or light,) then the cossim is better.
But actually matching exact colors (considering if it is dark or light) then dist is better.