Steps in Cosine Similarity

We had a Dataset of different people of whom we had to calculate similarities among them.

But we can't apply that directly on the dataset. So,

Step 1:

We have to choose the column at which we want to match different people. For this, we have the 'Super Type' column.

Step 2:

We can't work with the letters. So, applying pandas.get_dummies() to the column, we can have numeric values for the column.

Step 3:

In this step, we have to convert the dataset into an array using to_numpy(). If needed we may have to reshape the array.

Suppose that ***x*** and ***y*** are the first two arrays That
is, x=(5,0,3,0,2,0,0,2,0,0) and y=(3,0,2,0,1,1,0,1,0,1). How similar are ***x*** and ***y***?

We can do this using the formula –

$$similarity(x, y) = (x * y) / (|x| * |y|)$$

where x and y are two different persons.

Step 4:

we get:

$x^t * y = 5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1$

$= 25$

$||x|| = \text{root of } (5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2)$

$= 6.48$

$||y|| = \text{root of } (3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2)$

$= 4.12$

Applying these values in the formula , we have -

similarity(x, y) = 25/ (6.48 * 4.12)

= 0.94

So, these two persons can be said quite similar for their score being close to 1.