Name: Drashti Mehta

Lab 10

We will continue to work on our Metabolomics Database. As I said earlier, the mass spectra are typically used to annotate unknown compounds by comparing their spectra to the spectra of known compounds.

To compare the similarity of two spectra, we'll use the following formula:

$$Similarity(X,Y) = \frac{\sum_{i} \sqrt{x_{i} \cdot y_{i}}}{\sqrt{\sum x_{j}} \cdot \sqrt{\sum y_{k}}}$$

Where x_i and y_i are the intensities of the common peaks in spectra X and Y. Those peaks should have the same m/z value. In the denominator, $\sum x_j$ is the sum of all intensities in the spectrum X, while $\sum y_k$ is the sum of all intensities in the spectrum Y.

Example:

Assume that spectra X and Y have the following peaks:

Spectrum X		Spe	Spectrum Y	
m/z	intensity	m/z	intensity	
51	100	43	800	
63	70	51	400	
75	200	75	50	

The two spectra have two m/z values in common: 51 and 75. Therefore the nominator is equal to:

$$\sum_{i} \sqrt{x_i \cdot y_i} = \sqrt{100 \cdot 400} + \sqrt{200 \cdot 50} = 200 + 100 = 300$$

The first term above is for m/z=51, and the second term is for m/z=75. In the denominator, we have the sum of all intensities in X and the sum of all intensities in X. Therefore

$$Similarity(X,Y) = \frac{300}{\sqrt{100 + 70 + 200} \cdot \sqrt{800 + 400 + 50}} = \frac{300}{\sqrt{370} \cdot \sqrt{1250}} = 0.4411$$

Now, when we know how the similarity is calculated, let's write an SQL query to do it automatically!

The similarity formula will look much simpler if we normalize the intensities so that the sum of all intensities in a spectrum is equal to one. In this case, the denominator in the similarity formula is always 1 so we don't need to calculate it. I've already normalized the intensity values for you, so you can download the database file with the normalized intensities from Canvas.

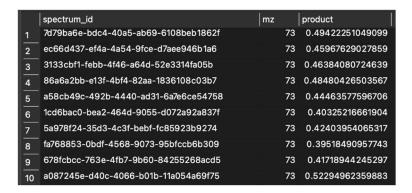
Name: Drashti Mehta

The following table contains m/z and intensity values of our query spectrum:

m/z	intensity
73	0.76
75	0.09
116	0.04
169	0.06
217	0.04

Task 1 (25pts)

Find all peaks with m/z=73 and calculate $\sqrt{I \cdot 0.76}$, where *I* is the intensity of each peak. The result should look something like this:



In SQL, use the function SQRT() to calculate the square root.

Report the SQL statement.

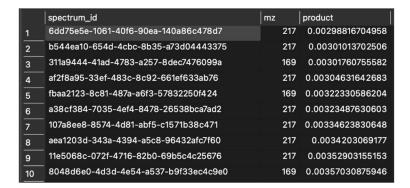
ANSWER:

select spectrum_id, mz, sqrt(intensity * 0.76) as product from peaks where mz = "73"

Task 2 (25pts)

Write similar SQL queries for the other four peaks and combine their outputs. You need to use UNION ALL to combined all of the outputs.

The result should look something like this (after ordering by the sqrt-values):



Name: Drashti Mehta

Report the SQL statement.

```
select spectrum_id, mz, sqrt(intensity * 0.09) as product from peaks where mz = "75" select spectrum_id, mz, sqrt(intensity * 0.04) as product from peaks where mz = "116" select spectrum_id, mz, sqrt(intensity * 0.06) as product from peaks where mz = "169" select spectrum_id, mz, sqrt(intensity * 0.04) as product from peaks where mz = "217"
```

UNION:

select spectrum_id, mz, sqrt(intensity * 0.76) as product from peaks where mz = "73" UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.09) as product from peaks where mz = "75" UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.04) as product from peaks where mz = "116" UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.06) as product from peaks where mz = "169" UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.04) as product from peaks where mz = "217" order by product ASC

Task 3 (25pts)

Finally, calculate the sum of sqrt-values for each spectrum in the database. Use ORDER BY for aggregating the values.

	spectrum_id	similarity
1	cfc6d4c1-0e20-4b09-9ccf-0d7e9d34324b	0.64120356332594
2	dbc80f2a-7d73-44ae-a1a7-5735c99551c8	0.62005044964943
3	ae155599-9f0a-4752-98a7-d2525e977f77	0.61970822750075
4	0a3048e5-a7ef-4f4b-9436-403856f8aa37	0.61518256347962
5	a087245e-d40c-4066-b01b-11a054a69f75	0.59920468567096
6	1e072b0d-9460-40d1-ad6d-3b69f80a81ed	0.59315571754465
7	8ed825a3-ddec-4dd7-a5b0-06f581ec90ca	0.59158810290938
8	ec66d437-ef4a-4a54-9fce-d7aee946b1a6	0.59102951432407
9	96abedee-ce79-4e8c-91d3-04b8b78fbfe6	0.58674623099658
10	c2eeed30-7112-4a90-b9a7-a86be89b8e62	0.58568111153211

If everything is correct, the column similarity will give the similarity calculated between the query spectrum and every spectrum in the database. If you sort the output by the similarity in the descending order, the spectrum #49 73.0000@7.6741 MS1+ (cfc6d4c1-0e20-4b09-9ccf-0d7e9d34324b) should be on the top of your list.

Report the SQL statement.

ANSWER:

select spectrum_id, **sum(product) as similarity** from (select spectrum_id, mz, sqrt(intensity * 0.76) as product from peaks where mz = "73" UNION ALL

Name: Drashti Mehta

select spectrum_id, mz, sqrt(intensity * 0.09) as product from peaks where mz = "75"

UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.04) as product from peaks where mz = "116"

UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.06) as product from peaks where mz = "169" UNION ALL

select spectrum_id, mz, sqrt(intensity * 0.04) as product from peaks where mz = "217")

group by spectrum_id order by similarity desc

Task 4 (25pts)

In SQLiteStudio, you can see how long does it take to execute SQL query. Check how long it takes to execute the query from Task 3.

You can speed it up by adding an index on one of the attributes in the table **peaks**. Try to add such an index and see whether it's improved the execution time for the query from Task 3.

On my machine, I had 0.005 sec without index and 0.001 with index.

Report: (1) SQL code for adding the index, (2) describe whether it has improved the running time.

- 1) CREATE INDEX [M/Z] ON peaks (mz);
- 2) Indexing improved the execution of query by speeding up and decreasing time of execution. Earlier taking 0.007s took 0.002s after indexing

