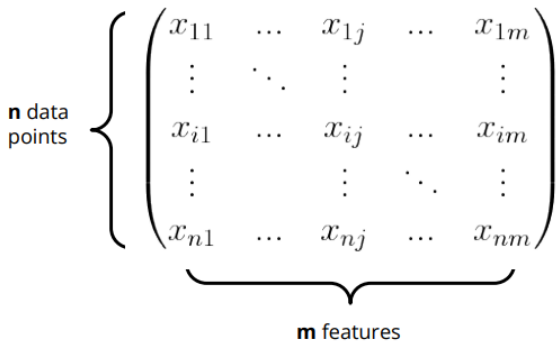
CS506 Lecture

☑ Override filters	☐
≡ Room	KCB 104

Notes Distance and Dissimilarity:

Data in the matrix:



Over here n represents the data points but m represents as a sort of features almost like X and Y coordinates.

Dissimilarity function:

- to uncover interesting structures from our data we need a way to compare data points

- A dissimilarity function is a function that takes two objects (data points) and returns a large value if these objects are dissimilar (or I guess not similar)
- in a sense, it is a way to compare data points
- distance is one we will use

Distance

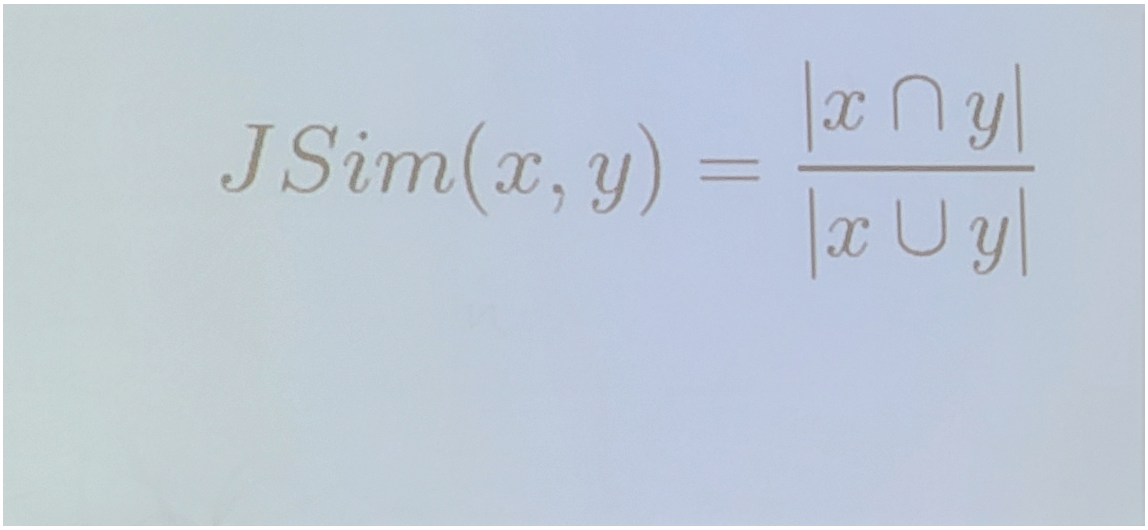
- properties
  - d(i,j) = 0 if and if i =j
  - d(i,j) = d(j,i)
  - d(i,j) $\leq$ d(i,k) + d(k,j) : Triangle inequality
- Minkowski Distance
  - $x = (x_1, ...x_d) \wedge y =$ $(y_1 ... y_d)$ for some d dimensional real space $p \geq 1$
  - P is a Parameter
  - d is the dimension

$$L_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

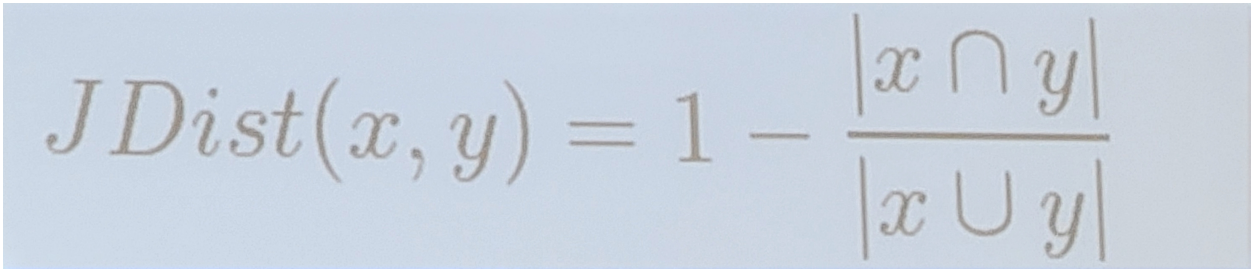
- Euclidean Distance:
  - The distance between two points is the length of the line between them
- Manhattan Distance:

- The distance between two points is the length of going up and across from two points (**NOT DIAGONAL**)

Jaccard Similarity:

- Example: union of all the words used in two documents in an 2d matrix or array or vectors
  - How would you would u compare the distance, you can use the Minkowski distance
  - If you have two documents with only 2 words differing the Manhattan distance would be 2 which means u lose a lot of info. Thus this is jaccard Similarity
- 
$$JSim(x,y) = \frac{|x \cap y|}{|x \cup y|}$$

- Similary


$$JDist(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

Cosine Similarity:

- The cosine of the angle of the two vectors
  - Two proportional vectors: 1
  - Two orthogonal vectors: 0
  - Two opposite vectors : -1
- To get a corresponding dissimilarity function we can do
  - 1- cos_similarity(x,y)

- Why would you want angle instead
  - Say you have two things talking about the same topic, but one uses more words and is longer over all
  - Thus distance would show than farther apart and imply they are very different
  - Thus angle will show that the are still quite similar even though one is longer and the other is shorter