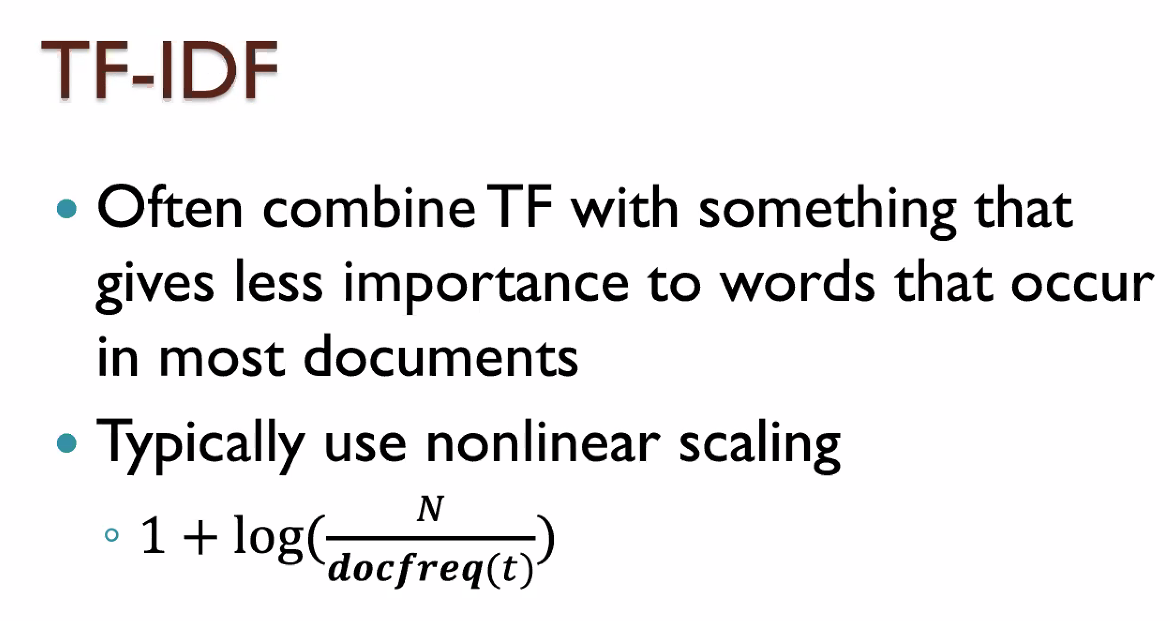
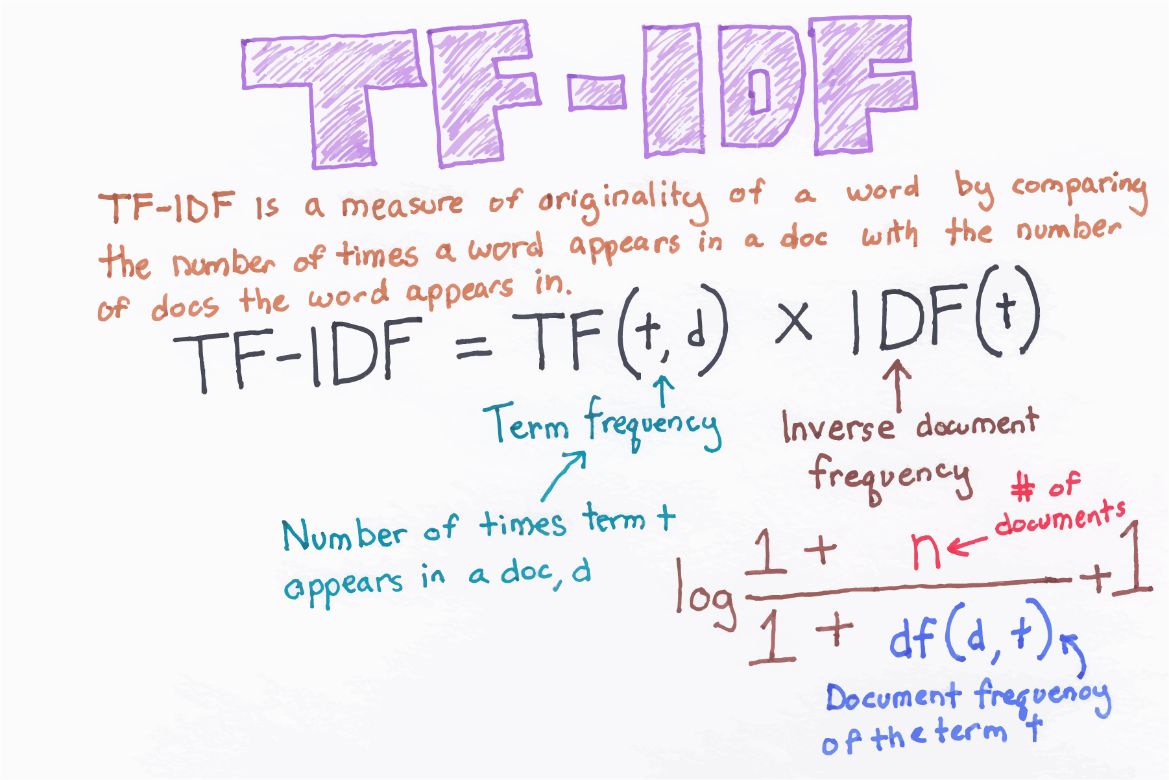
**Document classification:**

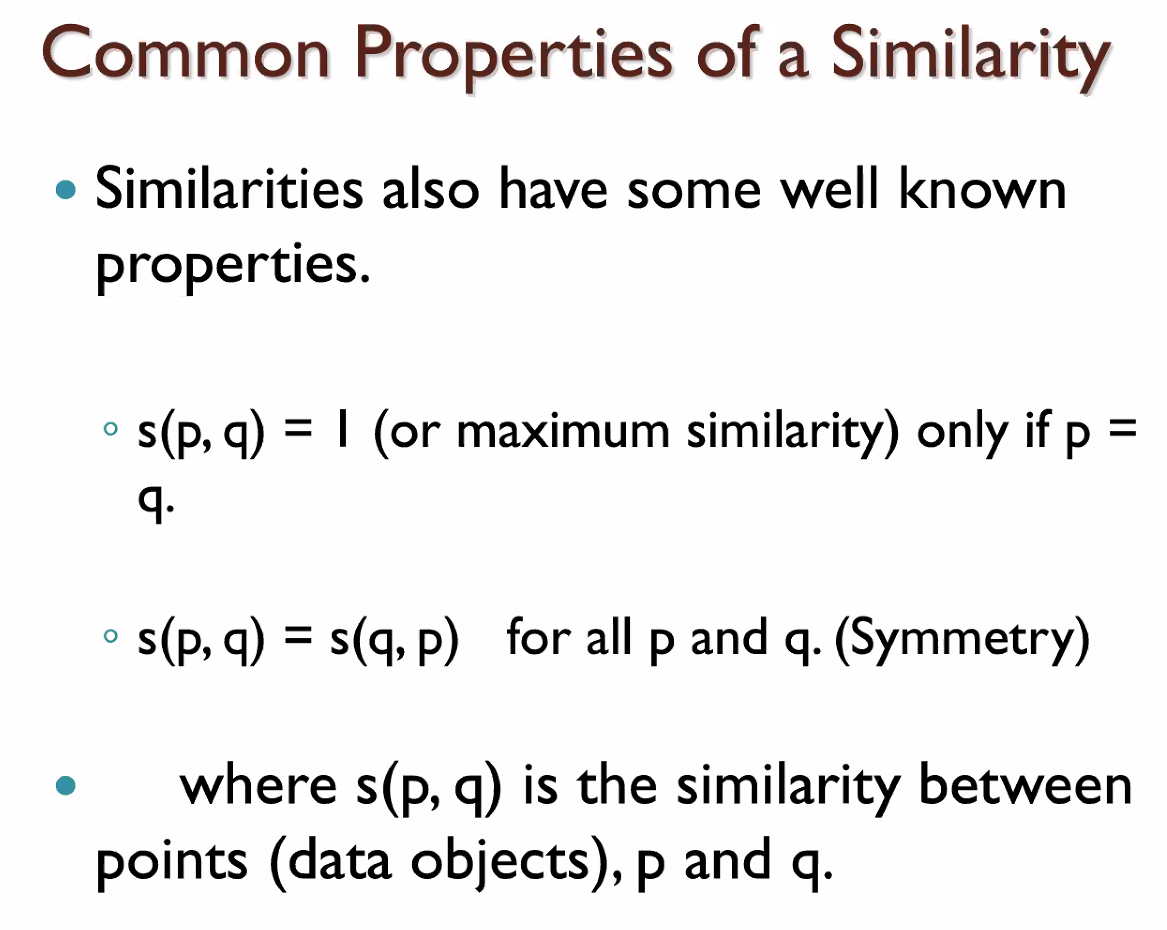
feature vector representation (Ex. Bag of words) row and columns

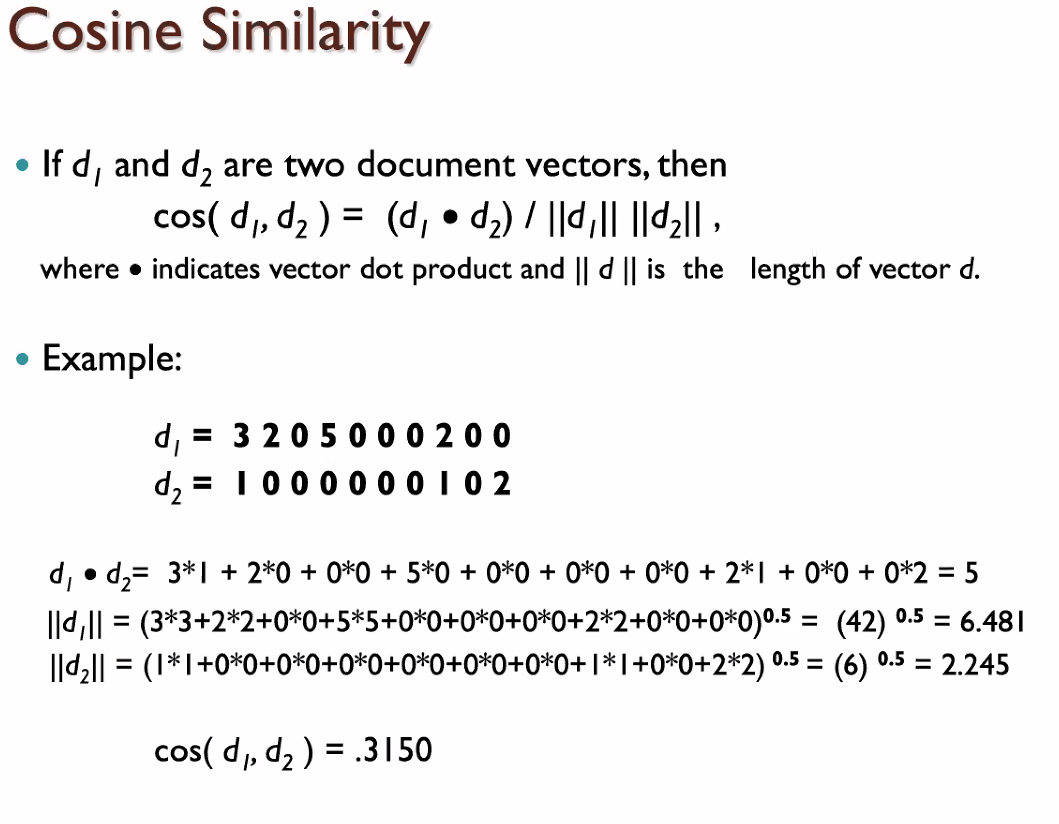
TF ( W2, d ) // term frequency

**TF-IDF**

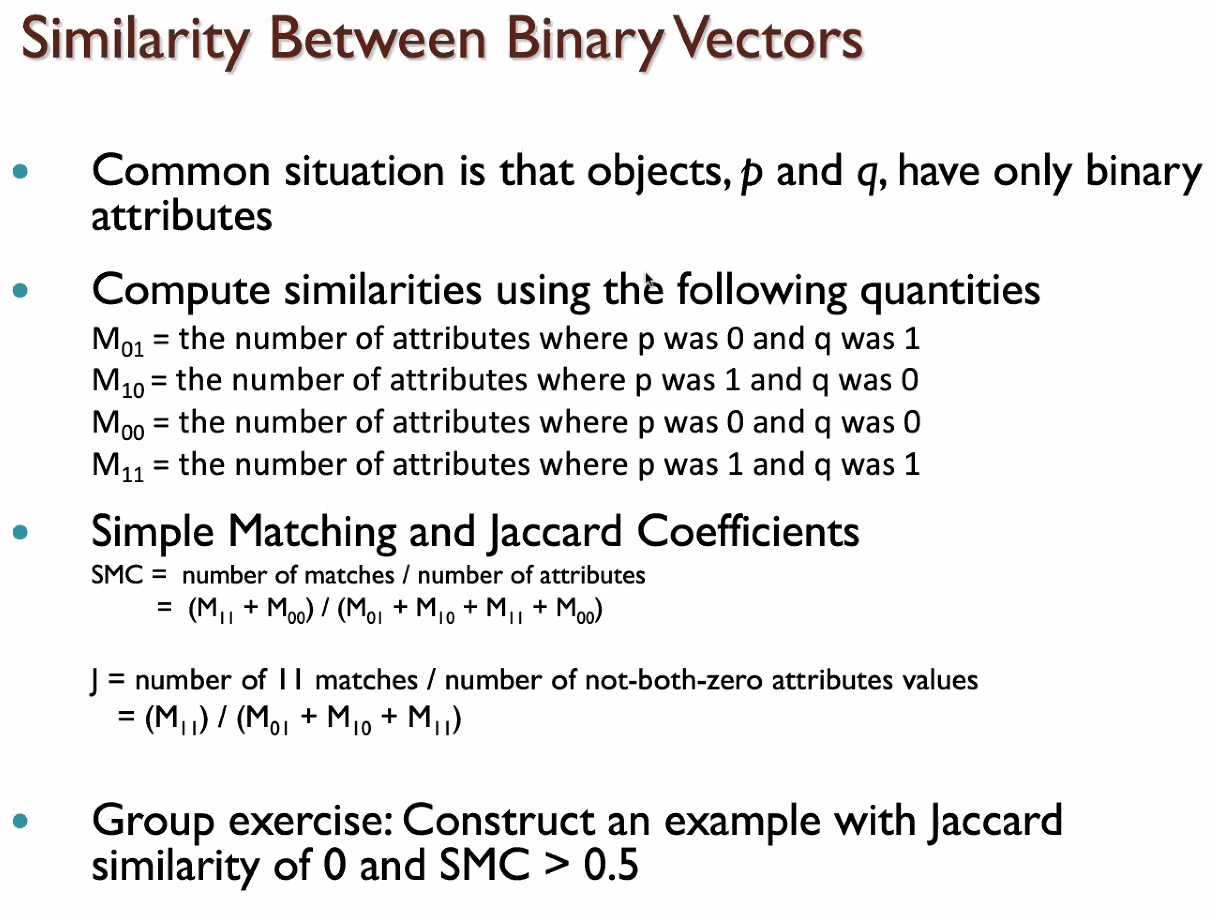




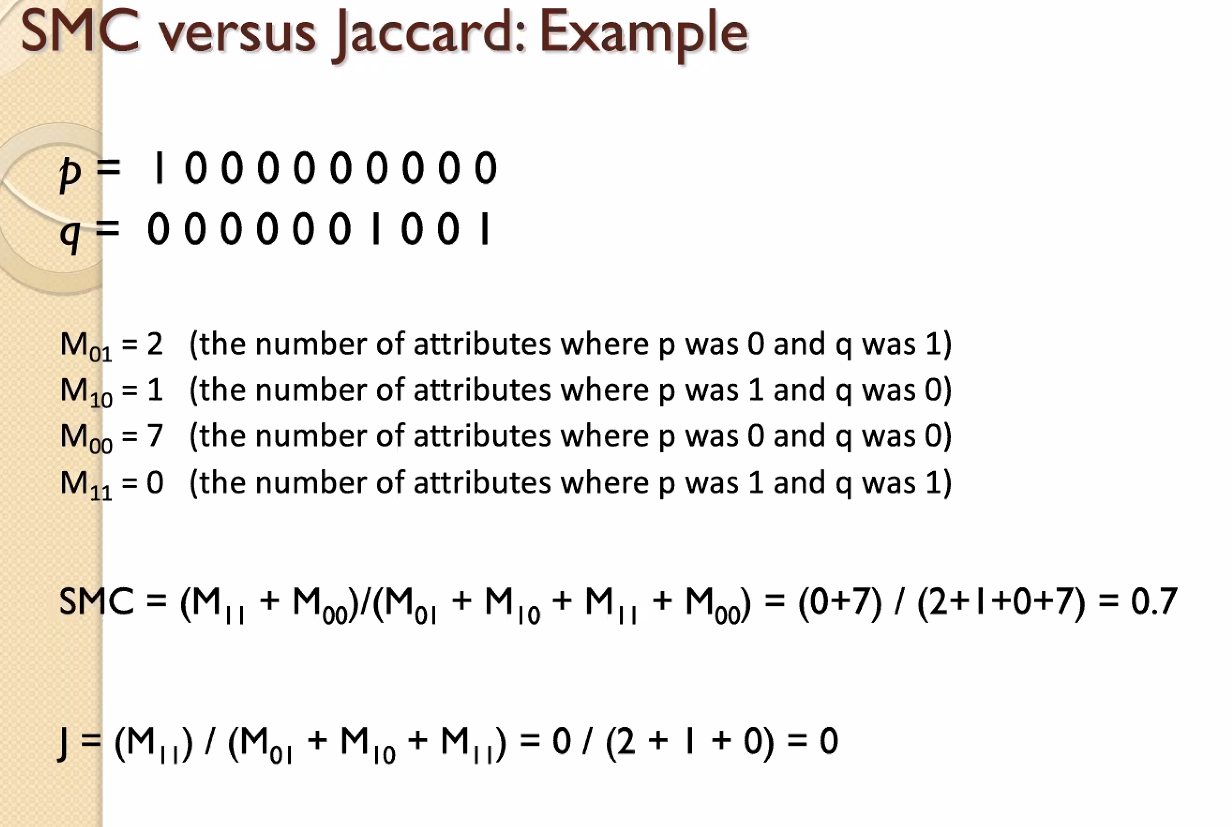


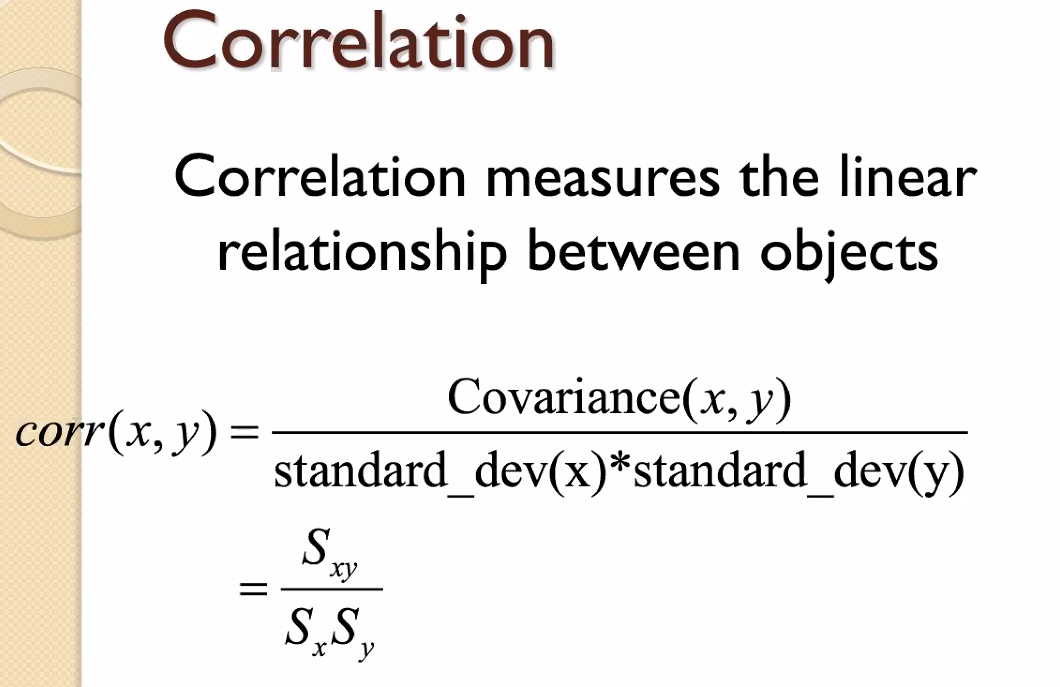


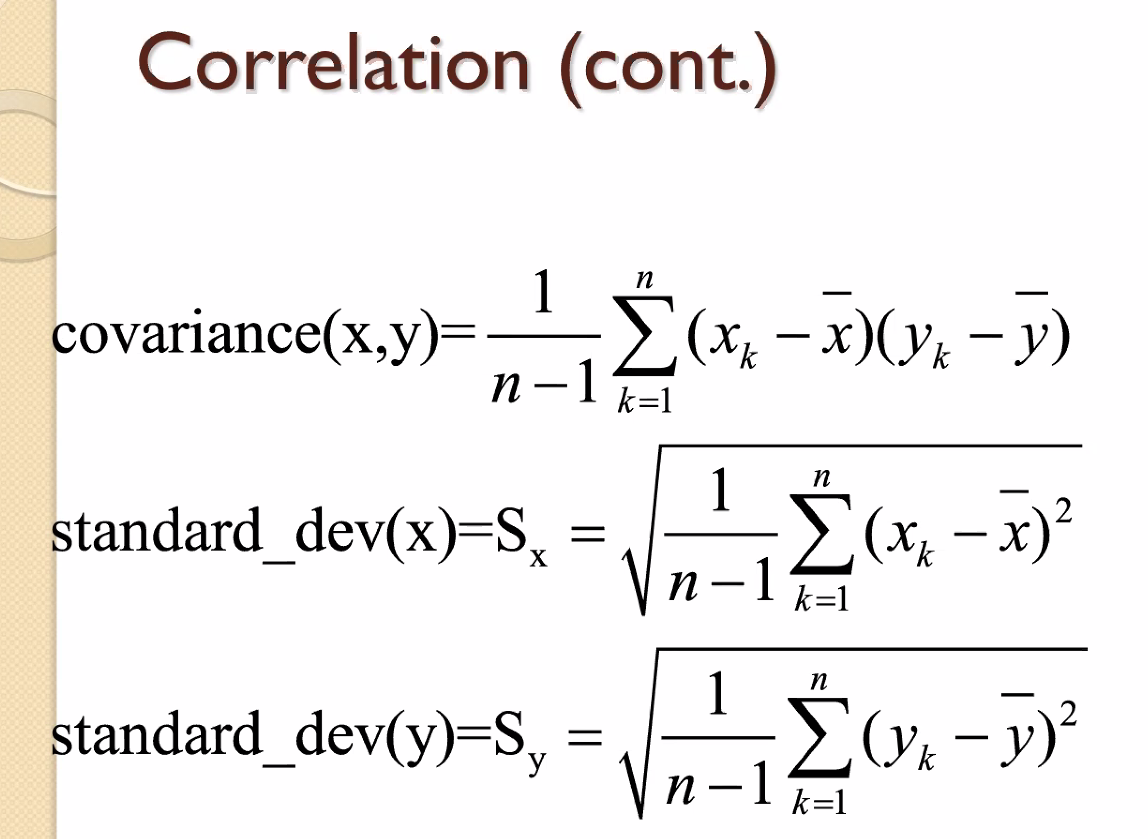
**To assess similarity between binary vectors: (CATEGORIES is good to use)**

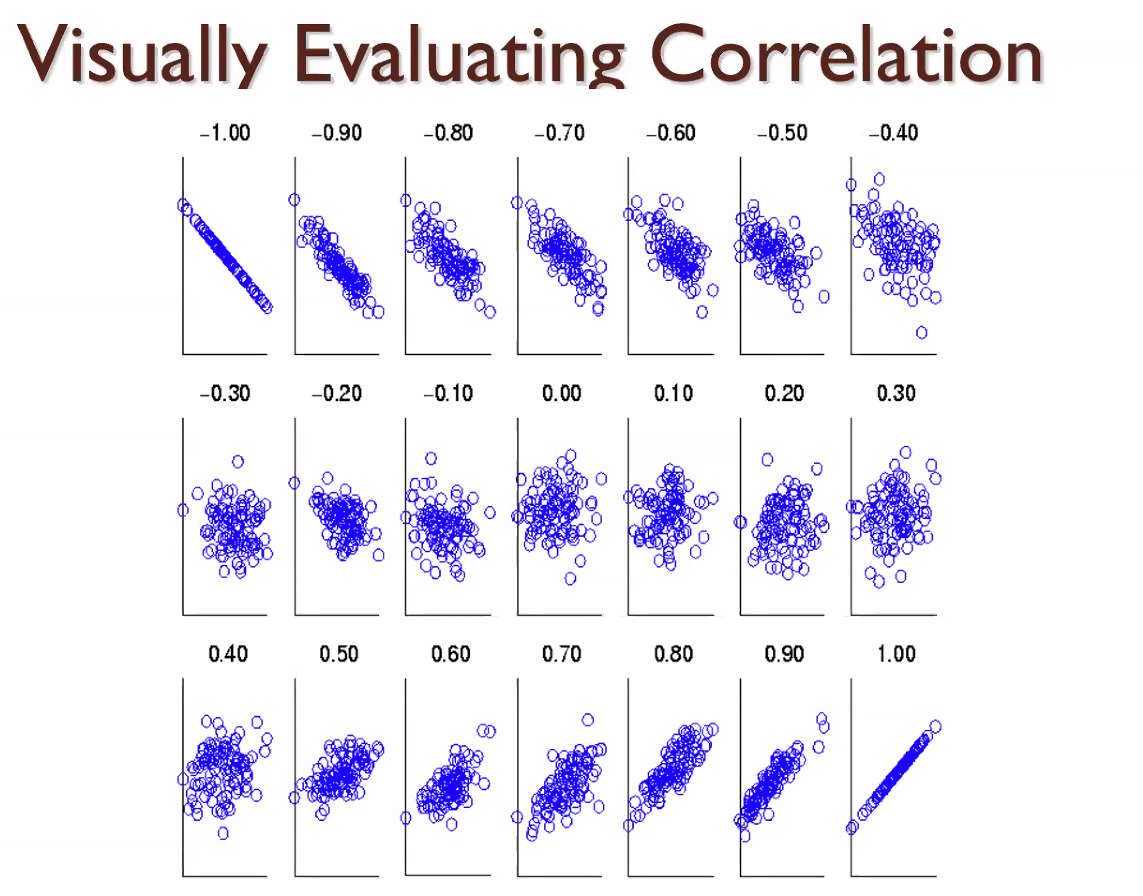


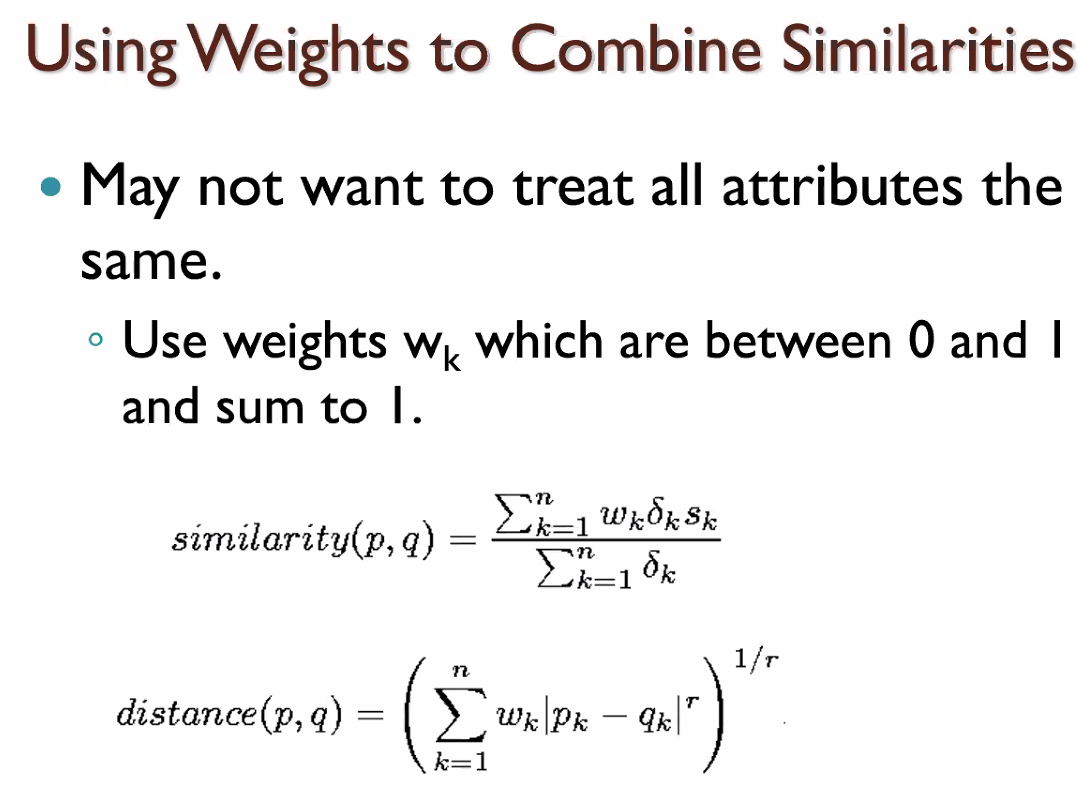
**EXAMPLE OF A SOLUTION TO THE GROUP EXERCISE:**



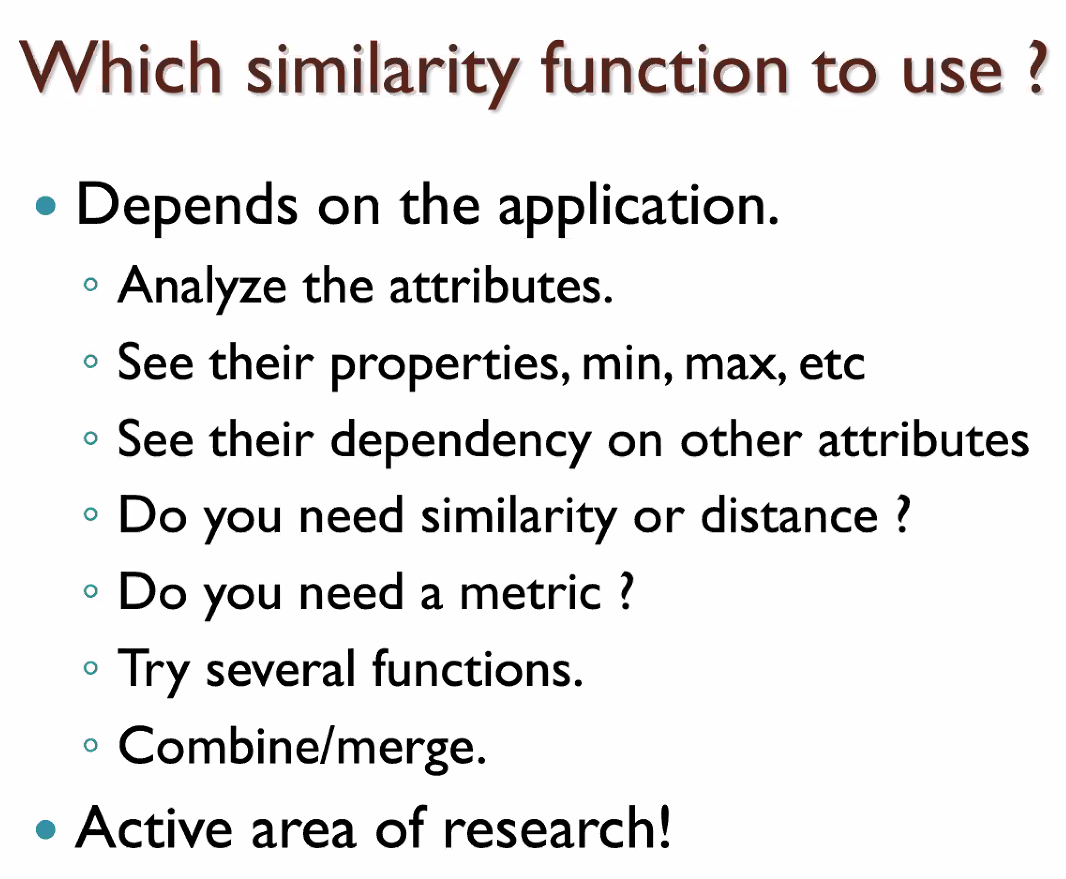






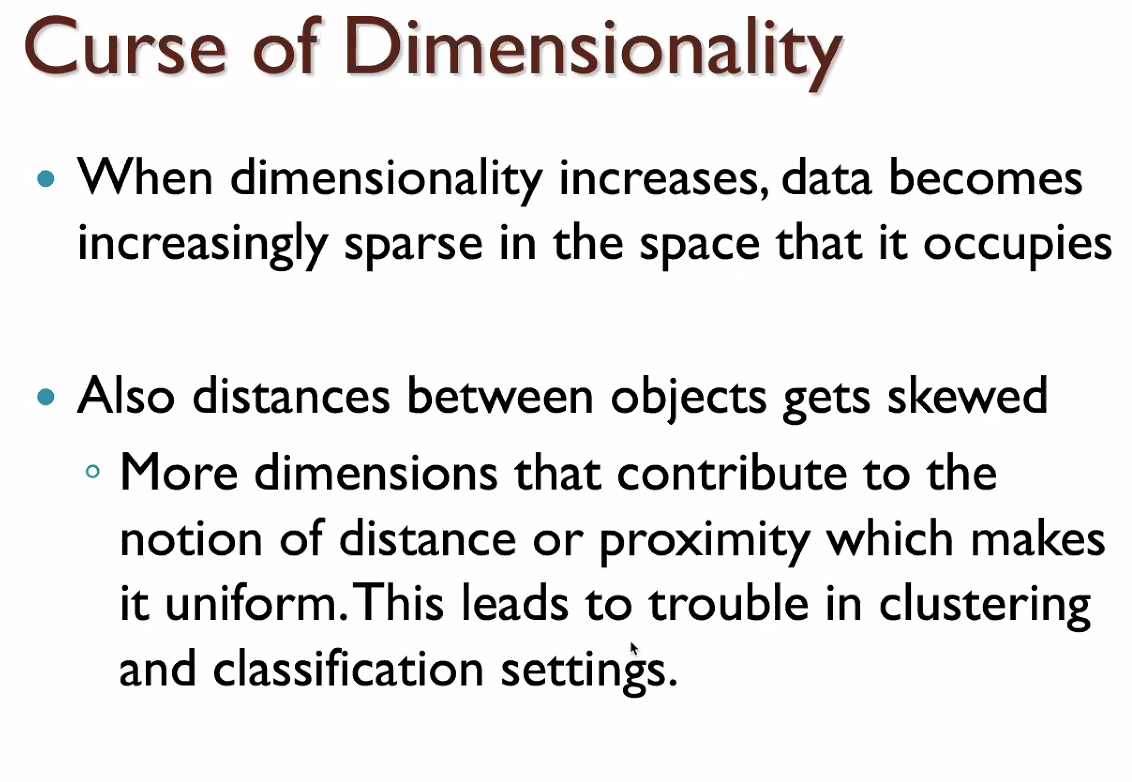


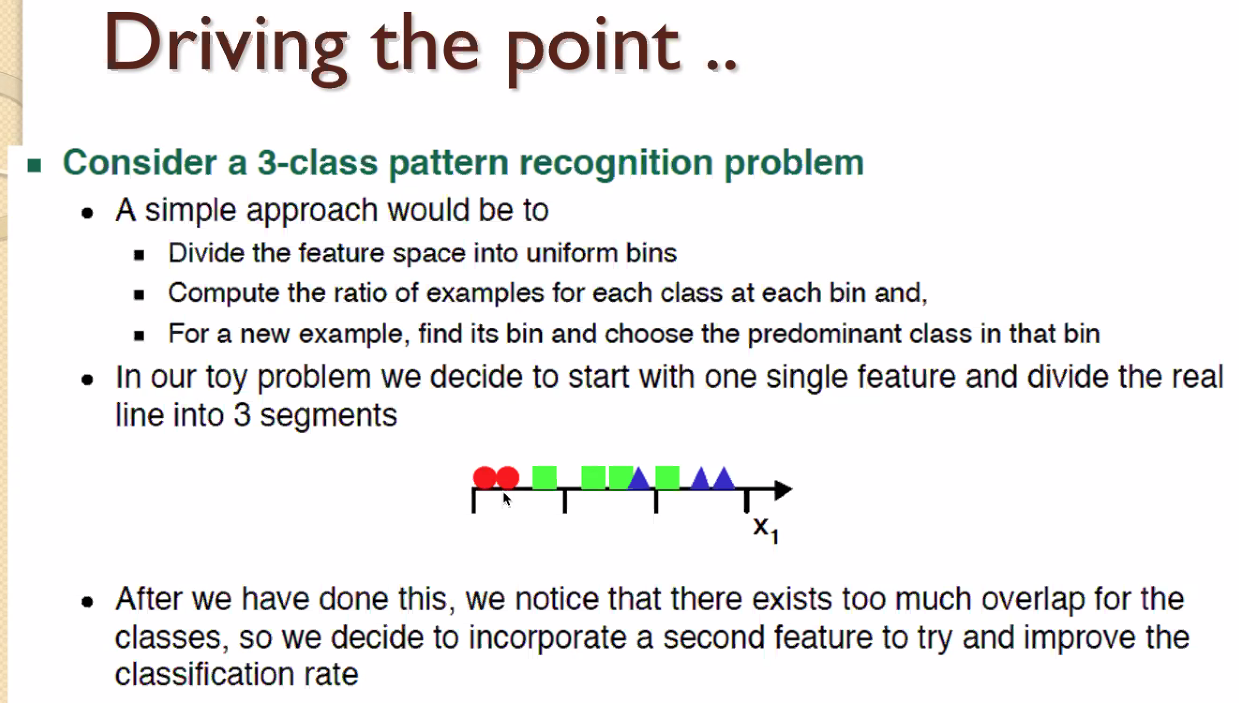
All the weights sum to one and define importance or rank weight value.

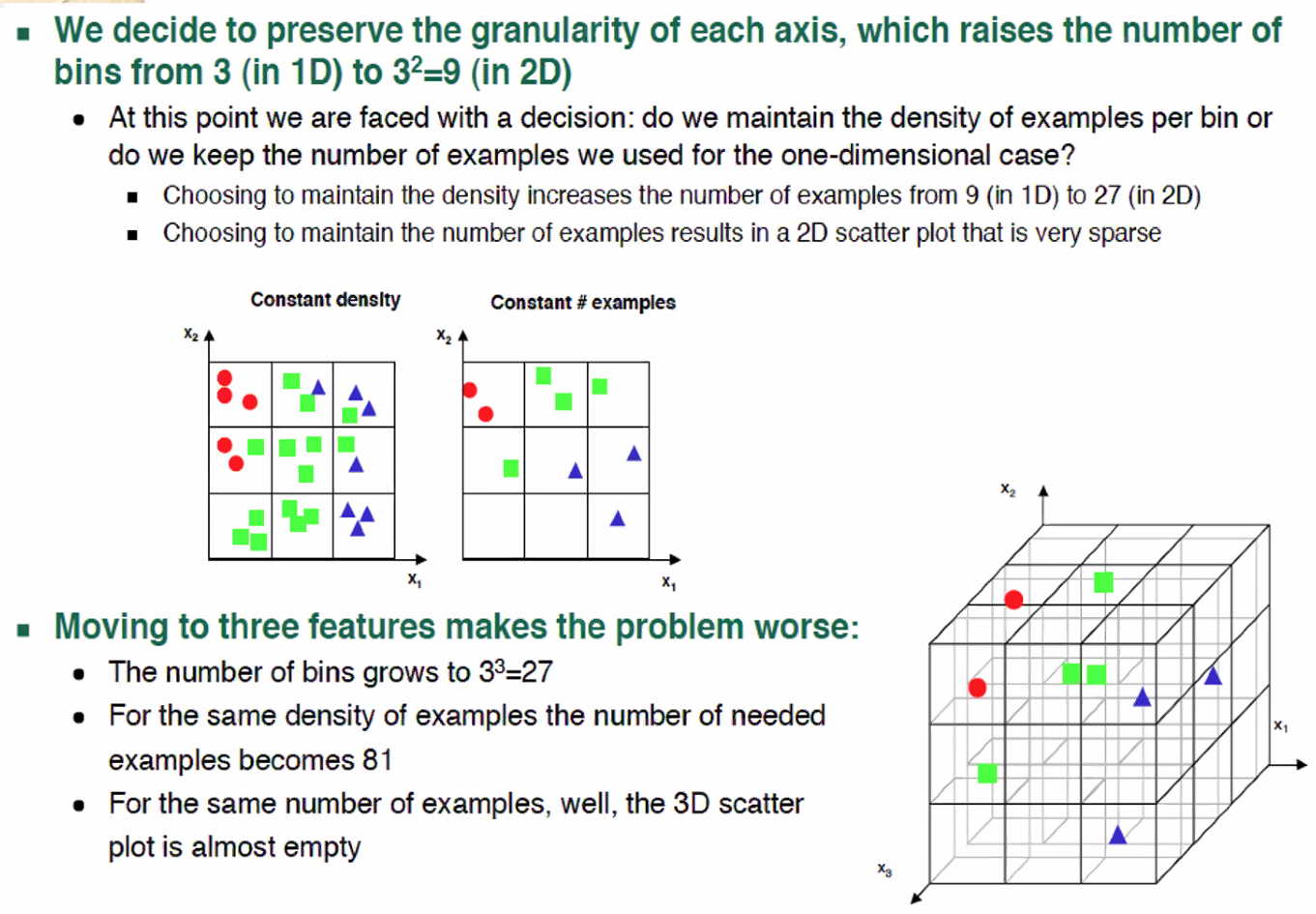


CURSE OF DIMENSIONALITY



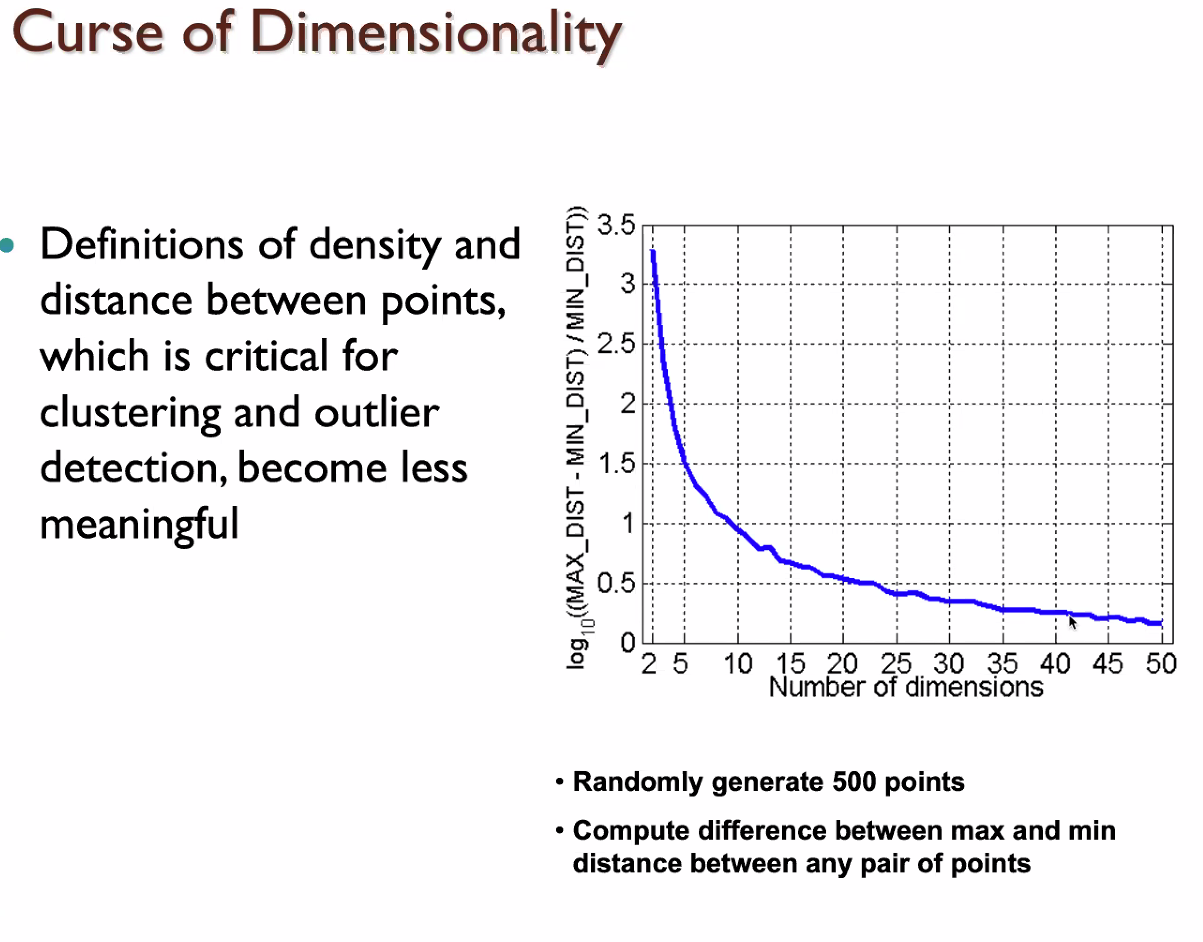




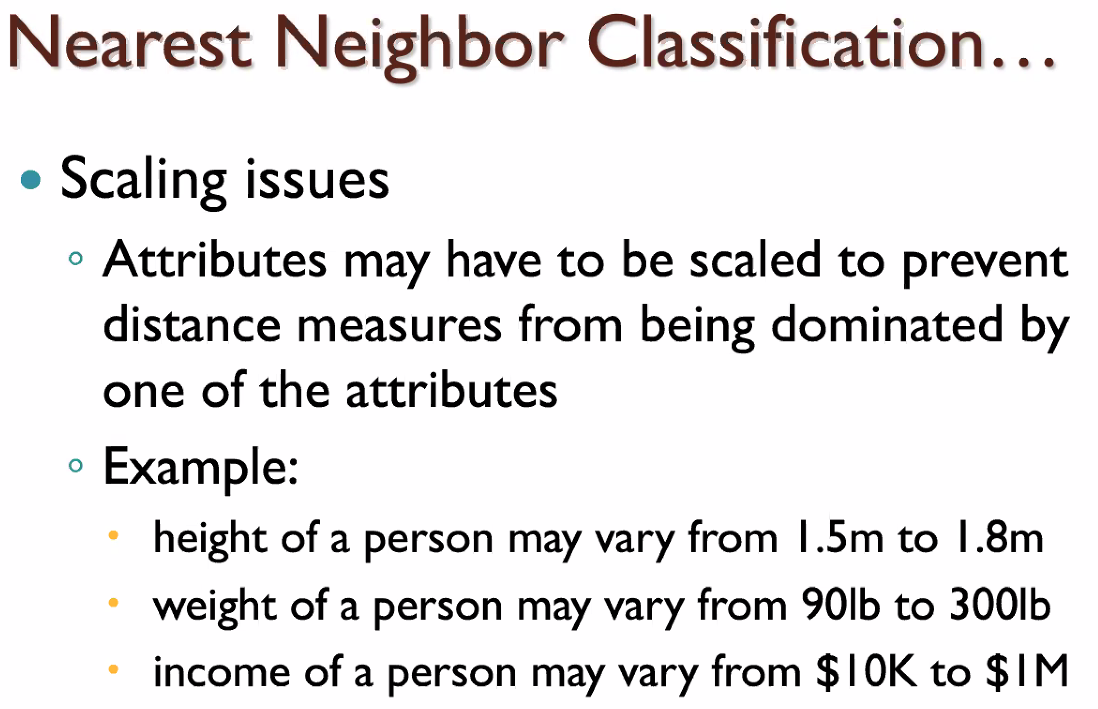


All the near neighbor are far away, but the nearest is actually far.

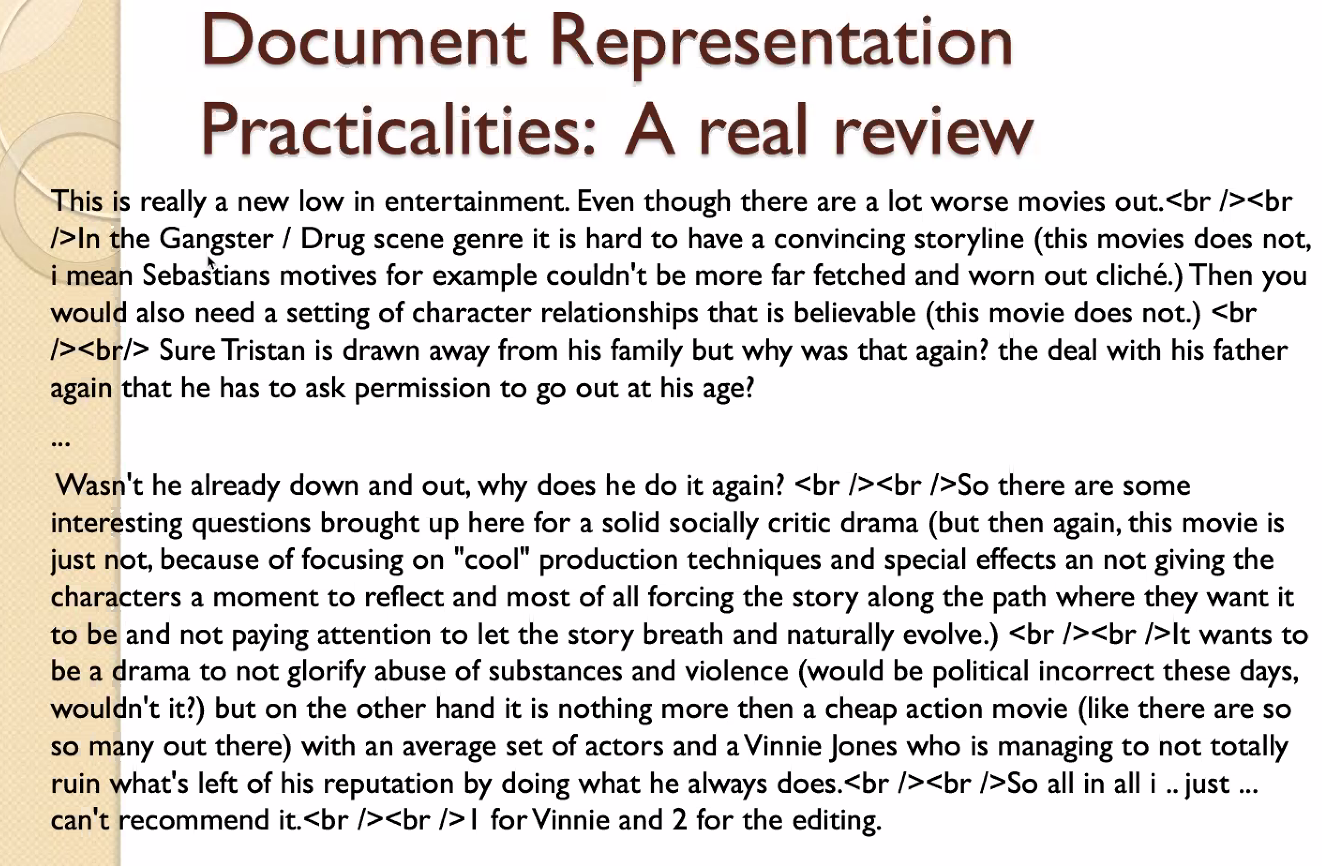
Does my data have enough density to classify the distance



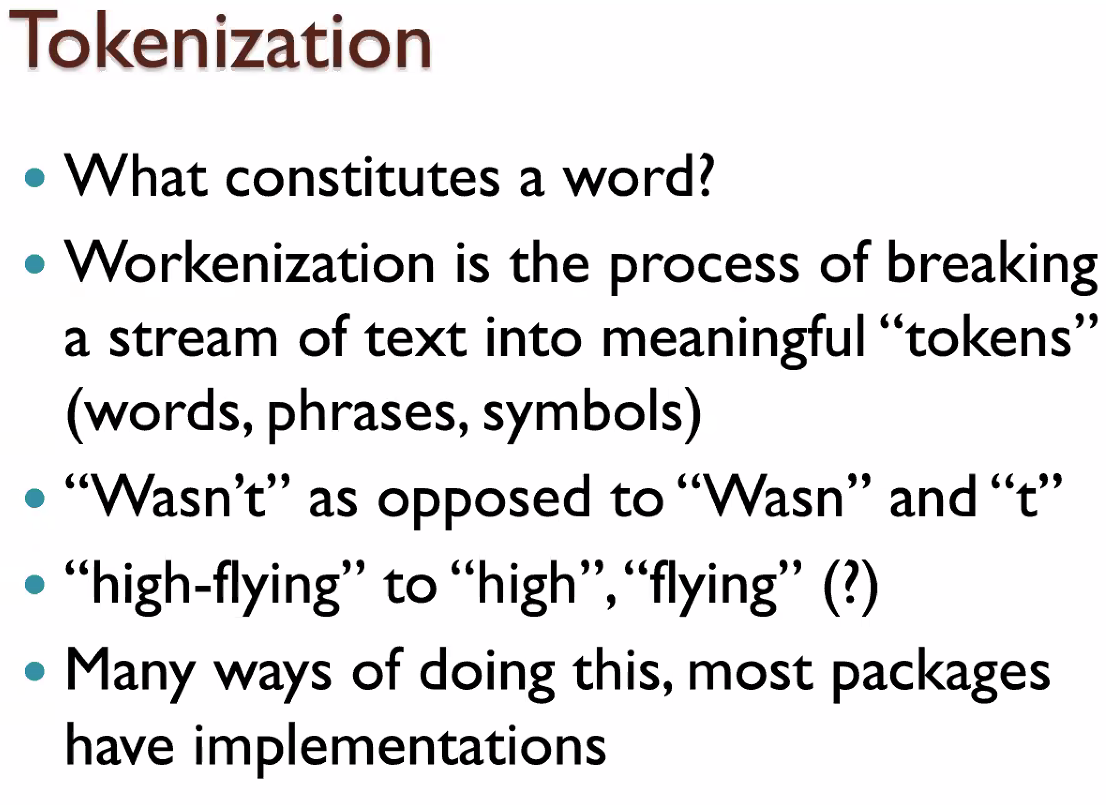
ISSUES:

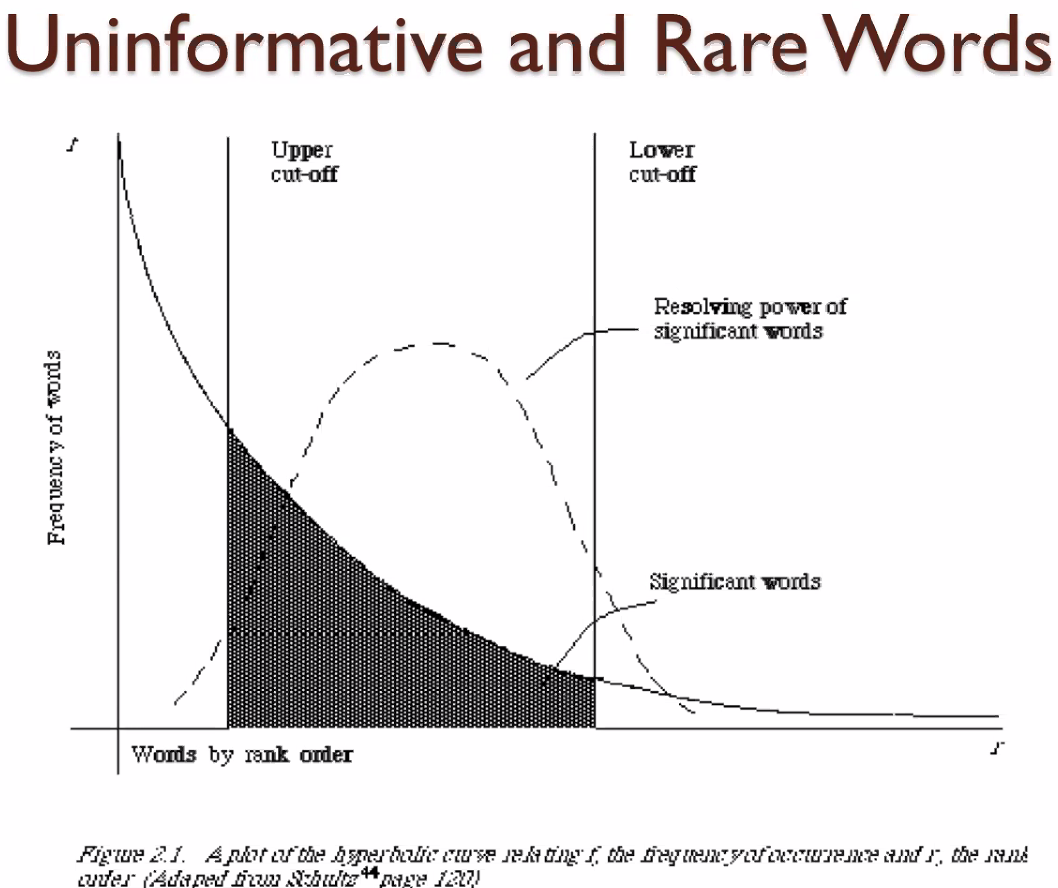


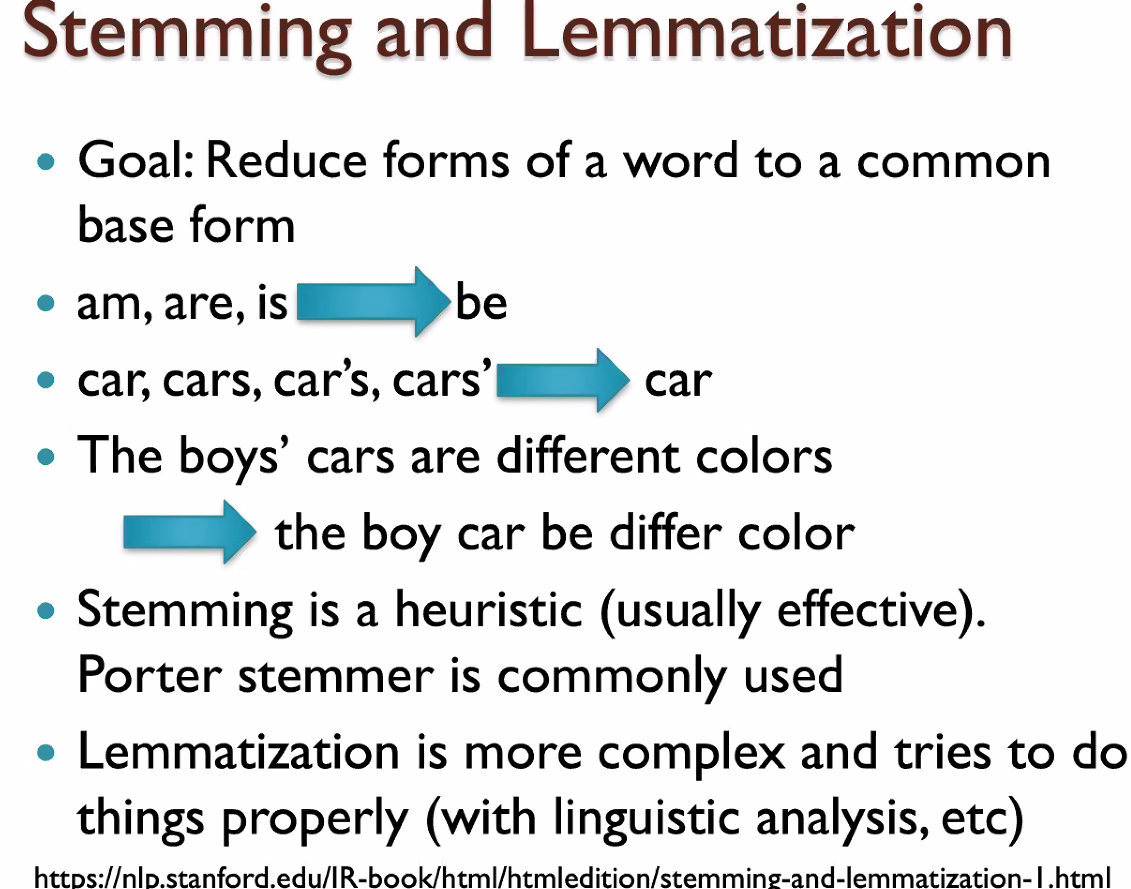
**RELEVANT FOR THE HOMEWORK:**



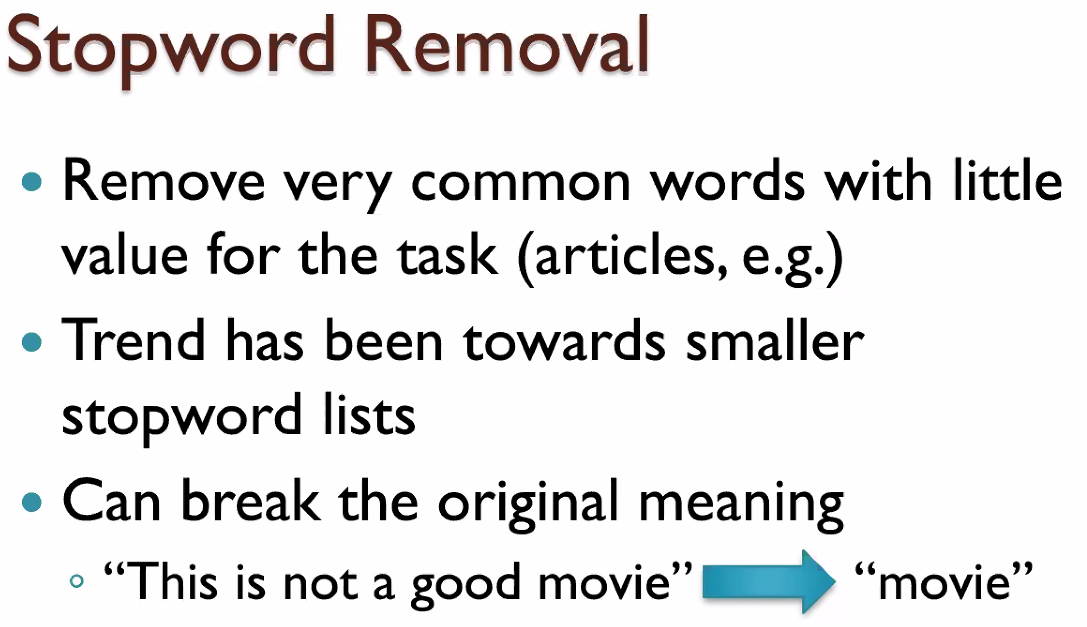
**Lots of noise ^, what you do:**







**For grammatical variances ^**



**NLTK can be useful for some homeworks**