Real Lecture 20: More Clustering and intro to graphs

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Clustering Review

Goal:

- Identify groups of related points in a dataset
  - points in each group are more similar to each other than to points in other groups

2 main algorithms:

- hierarchical clustering
- k-means clustering

Hierarchical Clustering

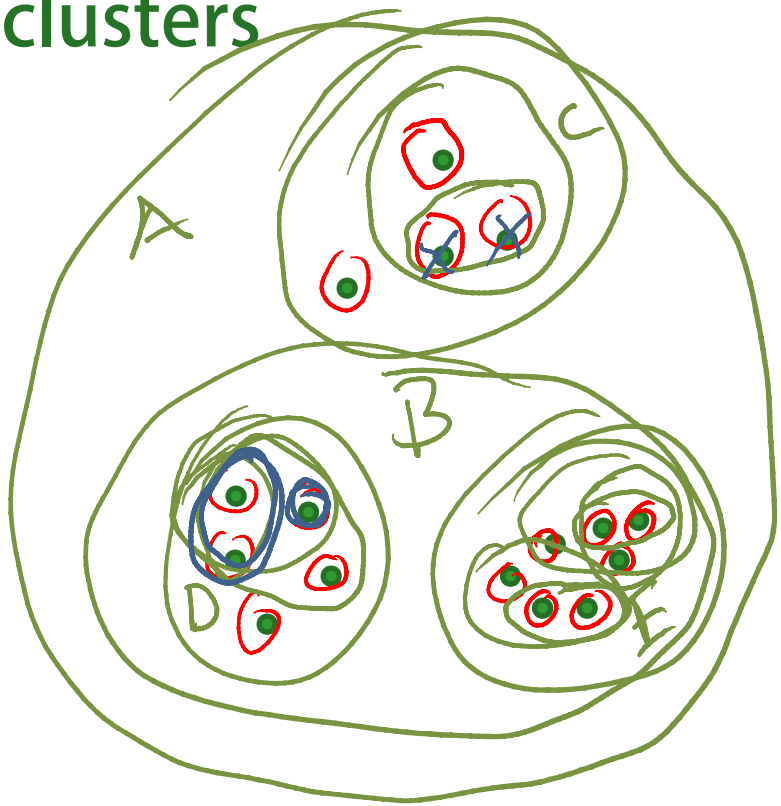
Results in a Tree of Clusters

Make every point a cluster

while(#clusters > 1):

pick the two *closest* clusters

merge them



Data Input Code

Real Data is always messier than you expect

- Spreadsheet had many unused columns
- Annoying quirks in data format

Use simple DataPoint class to store sanitized data

Data: http://bit.ly/121ZpgN

Cleanup Code: http://bit.ly/ZCwSQg

```
class DataPoint:
    def getMovies(self):
        return self.movies

    def getCourses(self):
        return self.courses

    def getHighSchoolLoc(self):
        return hsloc

    def getDreamCity(self):
        return dream
```

Hierarchical Clustering Code

generalizedHierarchical.py

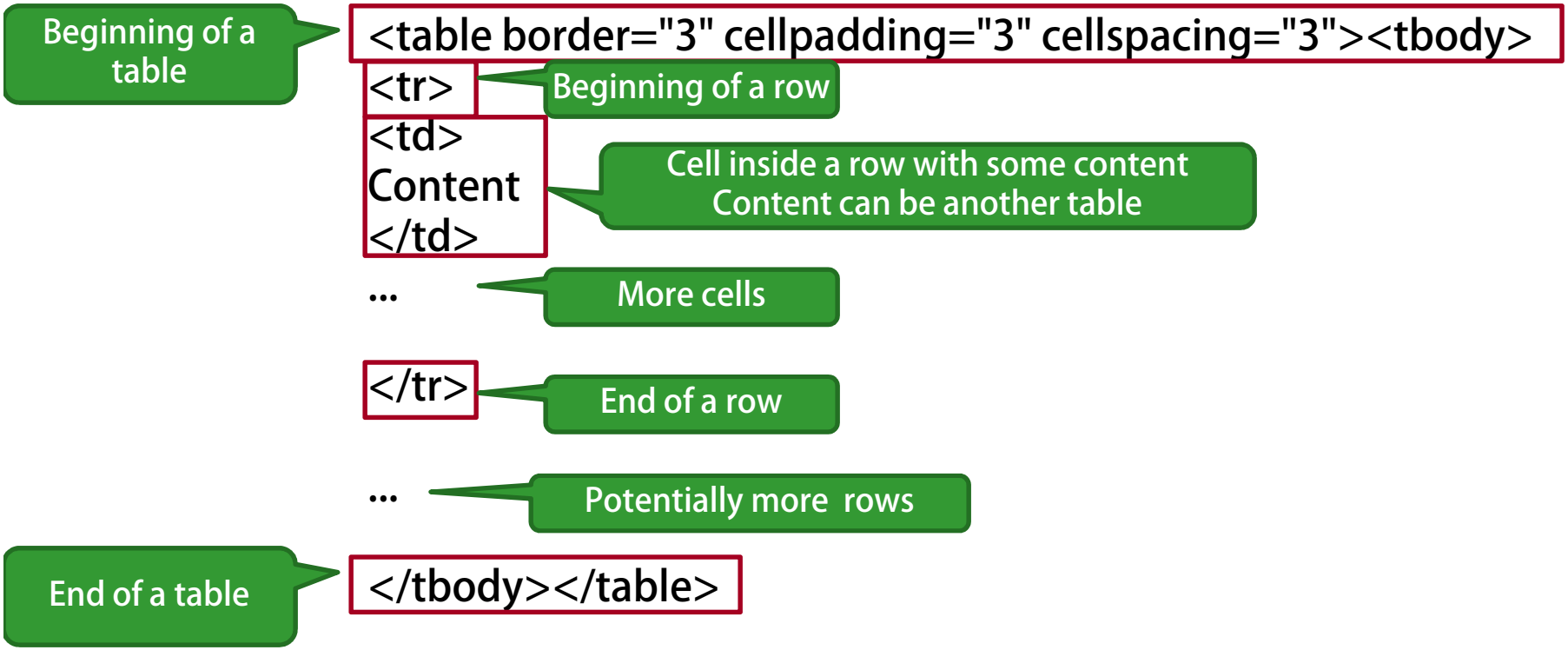
- implements a general version of hierarchical clustering
- uses two classes to support the two different kinds of clusters (Singleton and Composite)
- parameterized by distance function for both individual elements and clusters

http://bit.ly/15cApcE

Printing the Clusters

Cluster classes have a method to print HTML

- Very easy to print tables in HTML



Printing the Clusters

Example:

```
<table border="3" cellpadding="3" cellspacing="3"><tbody>
<tr> <td>Row 1 Col 1</td></tr>
<tr>
<td>
<table border="3" cellpadding="3" cellspacing="3"><tbody>
<tr><td>Row 2.1 Col 1</td>
         <td>Row 2.1 Col 2</td>
</tr>
<tr><td>Row 2.2 Col 1</td>
    <td>Row 2.2 Col 2</td>
</tr>
</tbody></table>
</td>
</tr>
</tbody></table>
```

You can copy and paste this code into a file such as test.html and view that file in your browser

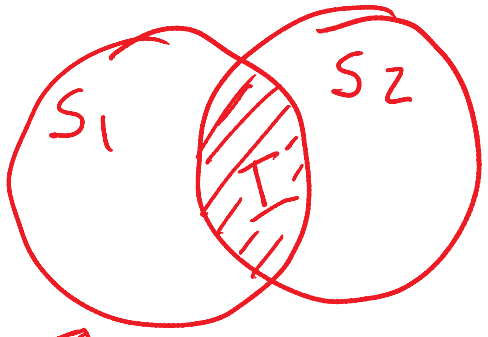
You can also copy it into an html previewer like

http://www.onlinehtmleditor.net/

Exercise 1: Clustering Movies

We want to cluster movies based on your list of movie preferences

- What is the distance between two movies?
  - Idea: If two movies often appear together in people's preference lists, they are "close"

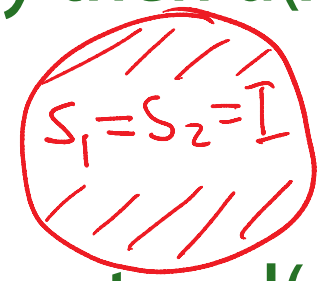


$$d(m_1, m_2) = 2 - \left(\frac{|I|}{|S_1|} + \frac{|I|}{|S_2|}\right)$$

Set of people who like movie m_1

Some good properties of distance

- Non negative: d(x,y) >= 0 ✓
- if x=y then d(x,y) = 0 ✓

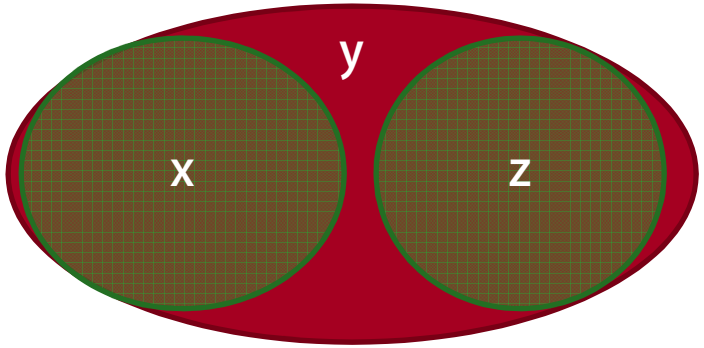


$d(x,x) = 2 - (1+1)$

- Symmetry: d(x,y) = d(y,x) ✓
- Triangle Inequality: d(x,z)<= d(x,y) + d(y,z) ☹

Triangle inequality not satisfied?

Consider what happens in the worst case



$$d(x,z) = 2$$

$$d(x,y) + d(y,z) = 4 - \frac{|I_{xy}|}{|x|} - \frac{|I_{xy}|}{|y|} - \frac{|I_{yz}|}{|y|} - \frac{|I_{xy}|}{|z|} = 4 - \frac{|I_{xy}|}{|x|} - \frac{|I_{xy}|+|I_{yz}|}{|y|} - \frac{|I_{xy}|}{|z|}$$

$\frac{|I_{xy}|}{|x|} = 1$, $\frac{|I_{xy}|+|I_{yz}|}{|y|} > 0$, $\frac{|I_{xy}|}{|z|} = 1$

Using the Clustering Function

The code to create movie clusters is here:

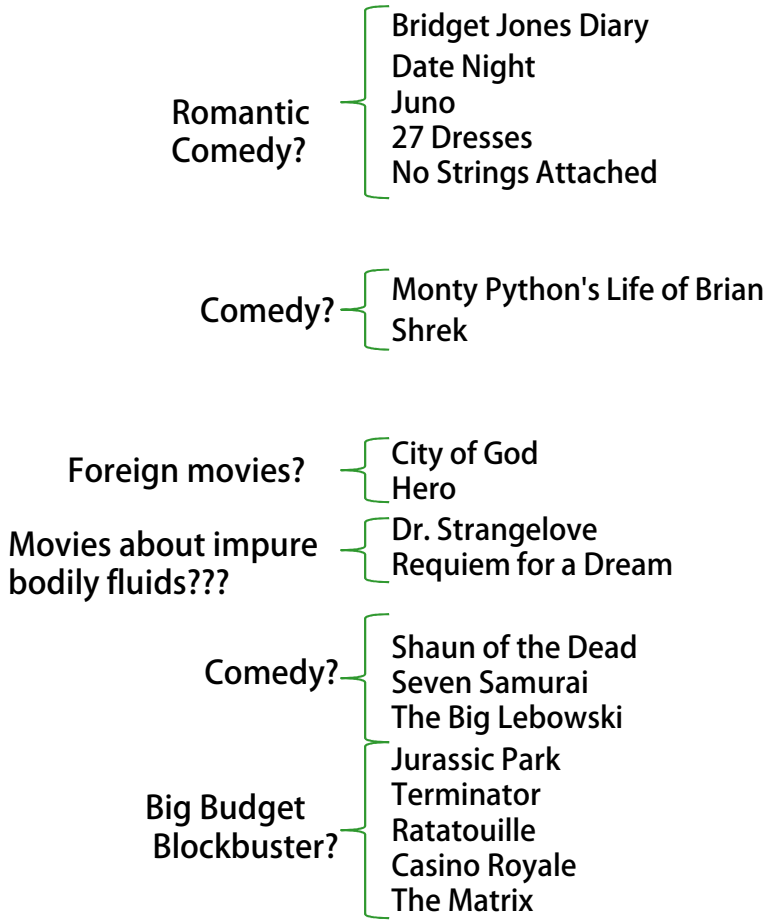
- http://bit.ly/YpPSCQ
- In order to run it, make sure it's in the same directory as cleanup.py and generalizedHierarchical.py as well as the data file
- Remember: to visualize the results you can copy and paste the output to an html file and open it in your browser, or copy and paste to the HTML visualizer here

  http://www.onlinehtmleditor.net/
- Try it with different distance measures and compare the results!

Result of clustering movies

Computed from the dataset as of 1:28 AM
(bigger than the dataset I showed in class)

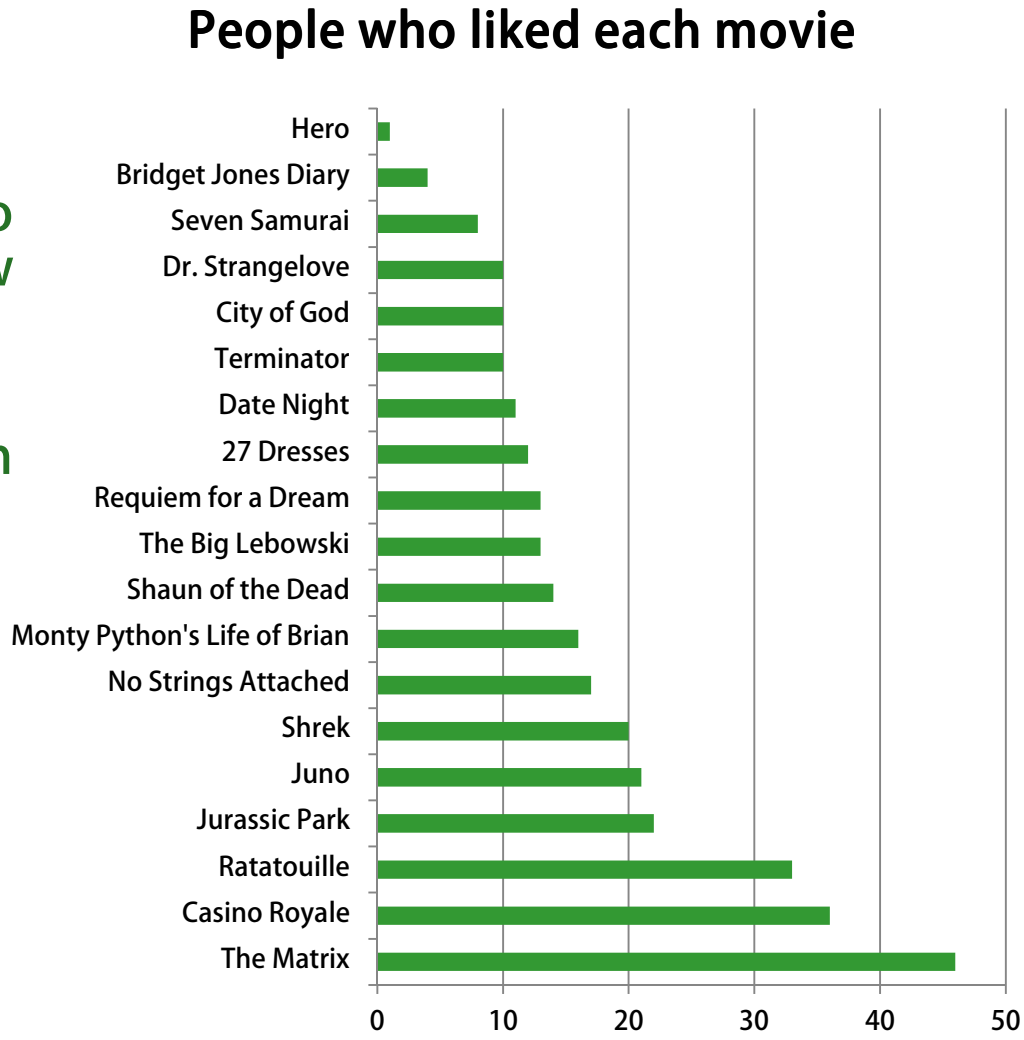
Note that the clusters do seem to correspond to reasonable movie categories
(at least some of them do; hard to say what category includes both
'Requiem for a Dream' and 'Dr. Strangelove')



Some things to keep in mind

Not all movies were liked by the same number of people

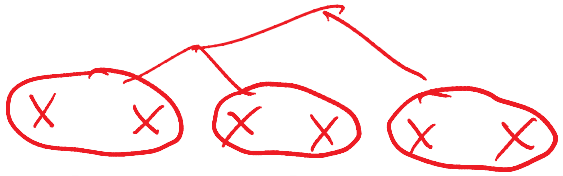
- Only one person liked ‘Hero’, so we can not read much into how it got clustered
- The fact that nobody liked both ‘The Matrix’ and ‘27 Dresses’ is much more significant



What do we mean by closest

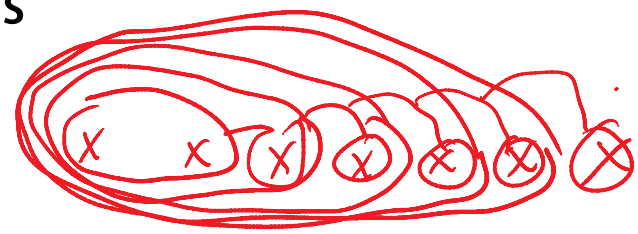
What is the distance between two clusters?

- some choices:
  - distance between the means
    - We can't compute this, but we can compute the mean of the distances between all points



What we used for previous clustering

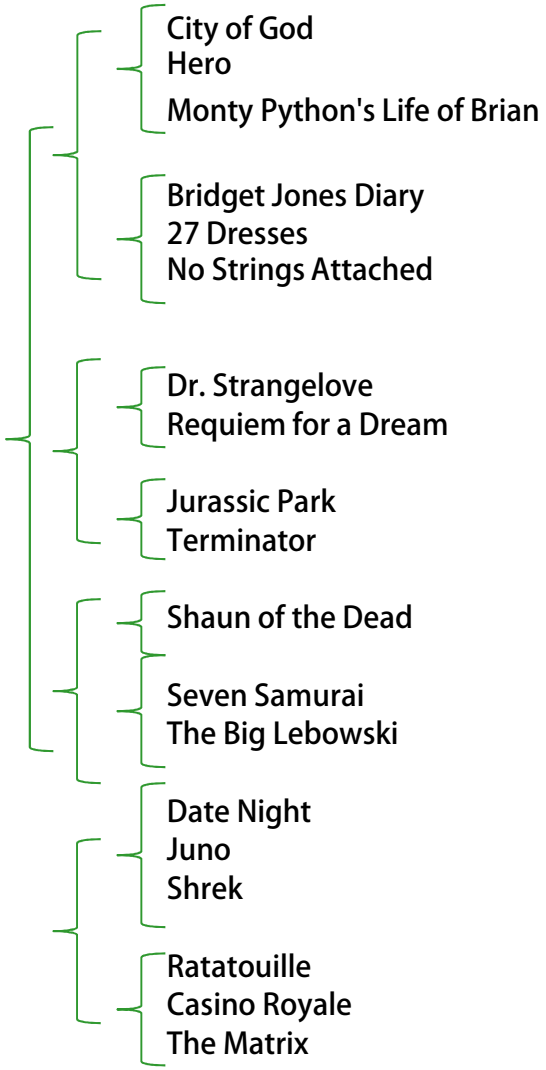
  - distance between the closest points
    - More likely to lead to single point clusters linked with large clusters



  - distance between the farthest points

Clusterings with different measures

Clustering using minimum distance instead of average



Clustering using maximum distance (a little better than min, but average is better)

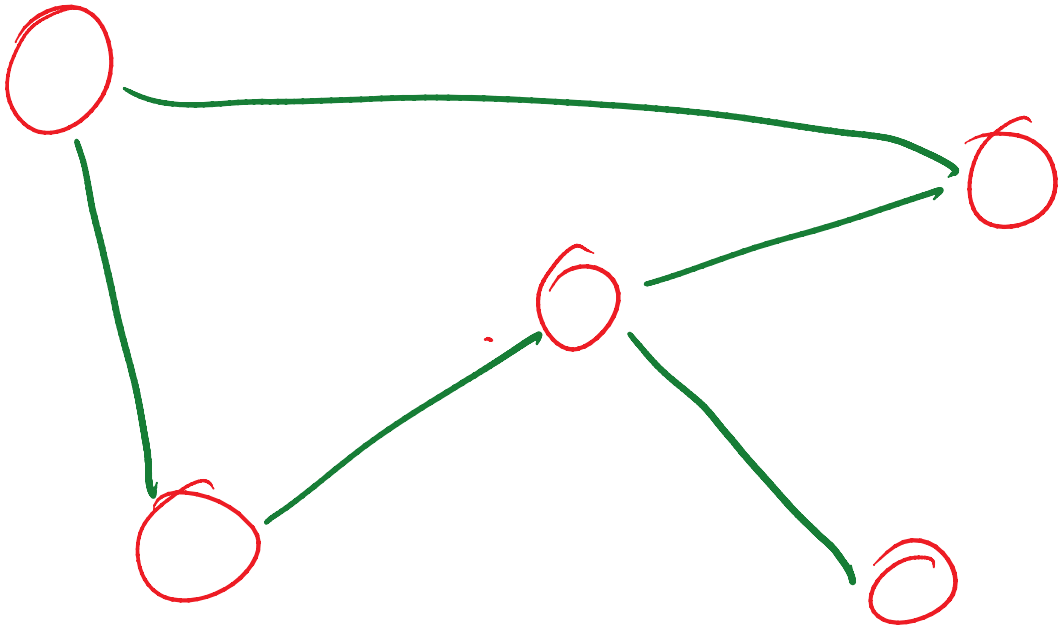
An introduction to graphs

Could we recover the triangle Inequality?

We can, and for that it's helpful to think in terms of graphs

What is a graph

A set of Vertices together with a set of Edges connecting those vertices (V, E)



What is a graph

A set of Vertices together with a set of Edges connecting those vertices (V, E)

- Edges may have direction
- They may also have weights

- Mathematical abstraction that can represent many things
  - People and Friendship relationships
  - Cities and Highways connecting them
  - Movies and their similarity
  - Assignments in a program and which assignments can follow each other
- Expressing something as a graph allows you to exploit their mathematical structure

Graphs

Questions you may want to ask about a graph

- Is there a path between two vertices?
- What is the shortest path between two vertices?
- Can I partition my graph so no two vertices in the same partition share an edge?

If the edges have direction you may also ask

- Does my graph have cycles
- ...

A distance with triangle inequality

Make distance the shortest path between two points

- This will satisfy the triangle inequality