Introduction | Problem Set 6 | Contenu du cours 6.00.2x

Introduction

In this problem set, you will use Python and pylab to write a agglomerative hierarchical clustering algorithm. You will use your algorithm to cluster cities across the United States according to some information available about each.

Getting Started

Download: Problem Set 6 skeleton code.

The problem set contains two files:

- cityTemps.txt - a text file containing information about varous cities in the United States. The first few lines starting with the hash (#) represent the column titles of the data available. The lines without the hash represent the actual data, separated by a comma -- in this case, cities and their average temperature in Jan, average temperature in April, average temperature in July, average temperature in October, annual precipitation in inches, and the number of days of precipitation in a year.
- clusterCities.py - a file containing some useful classes and a partial implementation of the `Cluster` and `ClusterSet` classes. In particular, this file contains:
    - A function, `scaleFeatures`, that can be used to scale the dataset features. Scaling essentially normalizes the values to be between 0 and 1 so that certain features do not overwhelm others
    - Functions to read the data in cityTemps.txt and produce a list of objects of type `City` from the data in the file.
    - Classes `Point` and `City`
    - An incomplete implementation of class `Cluster`. You will implement the functions `singleLinkageDist`, `maxLinkageDist`, `averageLinkageDist`.
    - An incomplete implementation of class ClusterSet. You will implement `mergeCluster`, `findClosest`, `mergeOne`.
    - A function that uses classes to implement the clustering and accumulate the results.
    - A function `hCluster` to test the program.