Principles of Computer Science II

Working with Data Sets

Ioannis Chatzigiannakis

Sapienza University of Rome

Lecture 19

Introduction to CSV

- CSV : Comma separated values
- Other separators are supported
- Module CSV provides useful functions to handle structured text files

```
5.1,3.5,1.4,0.2,setosa
4.9,3.0,1.4,0.2,setosa
4.7,3.2,1.3,0.2,setosa
4.6,3.1,1.5,0.2,setosa
5.0,3.6,1.4,0.2,setosa
5.4,3.9,1.7,0.4,setosa
```

Import File

Download file:

http://aima.cs.berkeley.edu/data/iris.csv

```
import csv
f = open("iris.csv")
dataset = []
for row in csv.reader(f):
    newLine = []
    for col in range(0, 4):
        newLine.append(float(row[col]))

    newLine.append(row[4])
    dataset.append(newLine)
```

Iris Flower Dataset

- Study of 150 Iris Plants
- 3 different types: Setosa, Versicolour, Virginica
- Each plant sampled for
  1. Sepal Length
  2. Sepal Width
  3. Petal Length
  4. Petal Width

Exercise – Scatter plots

- Create Scatter plots for all iris species:
  1. Sepal Length vs Sepal Width
  2. Sepal Length vs Petal Length
  3. Sepal Length vs Petal Width
  4. Sepal Width vs Petal Length
  5. Sepal Width vs Petal Width
  6. Petal Length vs Petal Width

```
plt.plot([row[0] for row in dataset if row[4]=='setosa'],
         [row[1] for row in dataset if row[4]=='setosa'])
plt.xlabel("Sepal Length")
plt.ylabel("Sepal Width")
plt.axis([0, 9, 0, 6])
plt.show()
```

Exercise – Histograms

- For each iris species create Histograms:
  1. Sepal Length
  2. Sepal Width
  3. Petal Length
  4. Petal Width

```
plt.figure()
plt.hist([row[0] for row in dataset if row[4]=='setosa'])
plt.show()
```

Exercise – 3D Plots

- For each iris species create 3D plots:
  1. Sepal Length vs Sepal Width vs Petal Length
  2. Sepal Length vs Petal Length vs Petal Width
  3. Sepal Width vs Petal Length vs Petal Width

```
from mpl_toolkits.mplot3d import Axes3D
plt.clf()
fig = plt.figure(1, figsize=(8, 6))
ax = Axes3D(fig, elev=-150, azim=110)
ax.scatter3D([row[0] for row in mergedata if row[5]==0],
             [row[1] for row in mergedata if row[5]==0],
             [row[2] for row in mergedata if row[5]==0],
            c='blue')
plt.show()
```

3^{rd} Assignment

- Work in groups of 3
- Generate a random list of integers (where n = {10, 100, 1000, 10000, ...}) and do a benchmark analysis for each one:
  - Python Lists with Quicksort, MergeSort
  - Binary Tree insertion, get random element, get max, delete random element
  - Python Heap insertion, get max, delete max
- Use timeit to measure execution time
- Produce plots for each of the above results visualizing the performance of the different algorithms
- Email `ichatz@dis.uniroma1.it`
  Subject: [PCS2] Homework 3
  A link to a github repository with your python code.
- **Deadline: 11/December/2017, 23:59**