University of Canberra

Faculty of Science and Technology

**Programming for Data Science G (11521)**

**Week 6 Tutorial**

**Dictionary and Assignment 1**

**Tasks**

* To practise examples of Sequence, Set and Dictionary
* To apply List and Tuple to Assignment 1 to display data points and cluster centres.

**Create a new Python project**

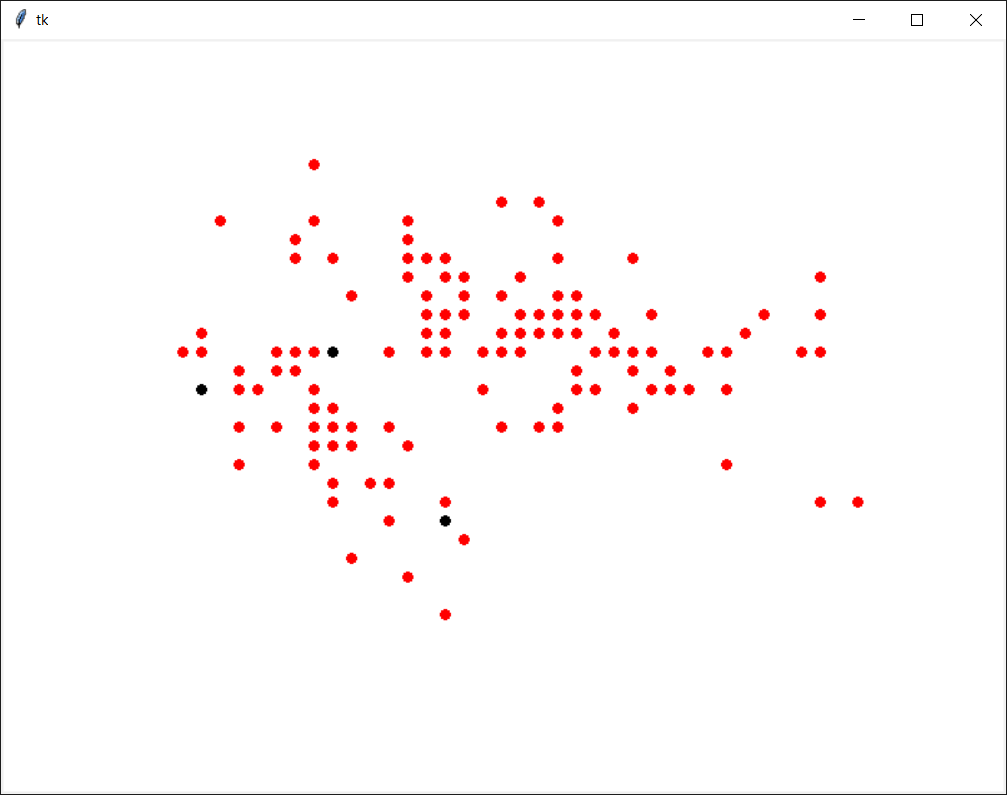
* Create **Week6Tutorial** project (Python)
* Add a module file and name it **io\_data\_module.py**
* There are 4 examples for Sequence, 5 for Set and 7 for Dictionary in Week 5 Lecture
* Review those examples before you answer the questions below.

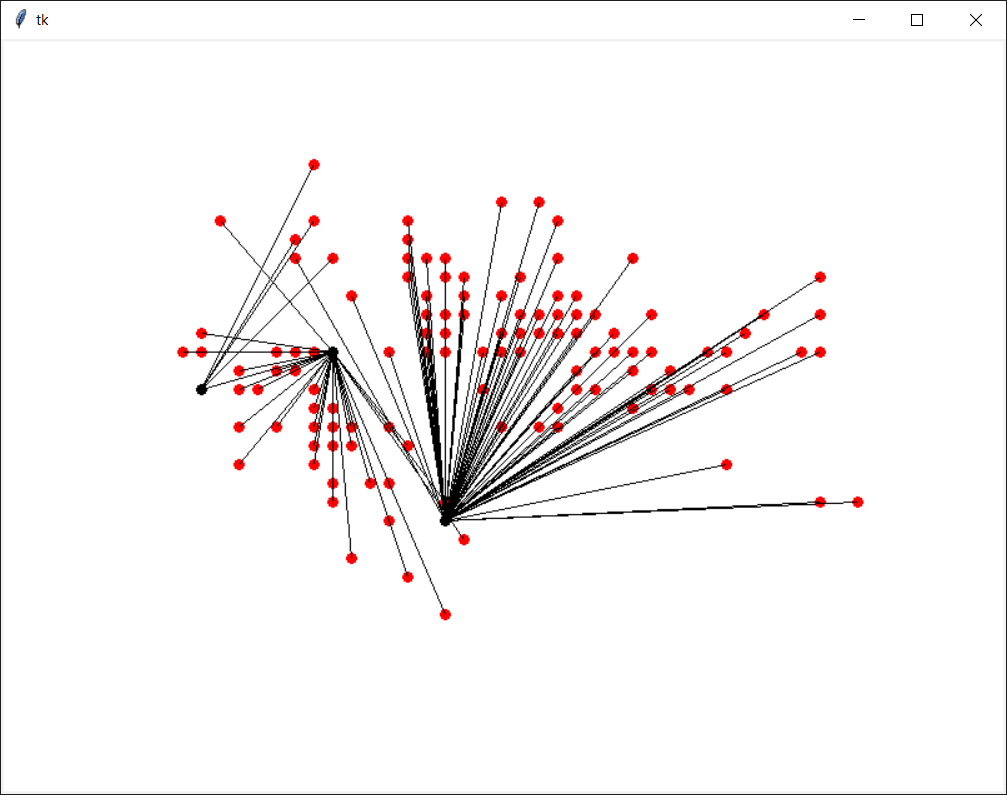
**Question 1:**

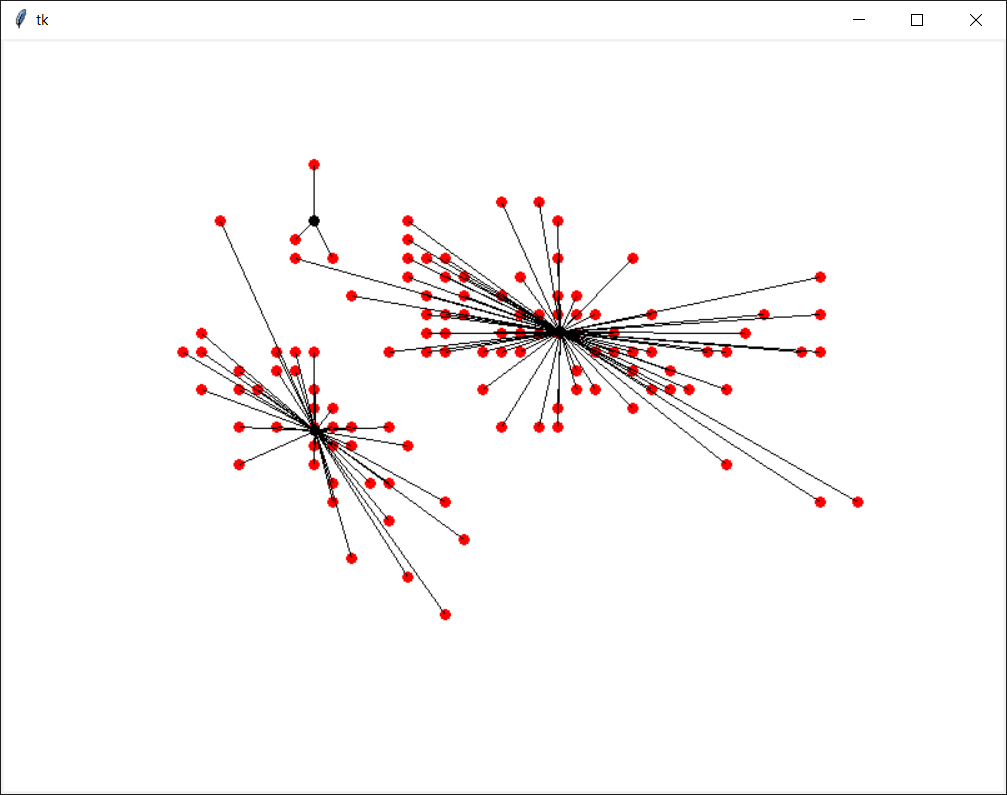
* Implement a Python function (**read\_data\_to\_dict**) that opens the **iris.data** file to read all **150 data samples** and **3 class labels**. This function will output a dictionary that contains 3 key-value pairs as follows
  + **Pair 1**: key: Iris-setosa, value: list of 50 tuples, each tuple is a 4D data sample
  + **Pair 2**: key: Iris-versicolor, value: list of 50 tuples, each tuple is a 4D data sample
  + **Pair 3**: key: Iris-virginica, value: list of 50 tuples, each tuple is a 4D data sample

**Question 2:**

* Implement a Python program in **Week6Tutorial.py** that
  1. (done in **Question 9** in **Week 5 Tutorial**) reads all 150 data samples in **iris.data** and displays them on canvas using the function developed in **Question 8** in **Week 5 Tutorial**, where **x** and **y** are the **first** and **second** values in each data sample, respectively, and
  2. (done in **Question 9** in **Week 5 Tutorial**) displays 3 centre samples centre\_1 = (5.1, 3.0, 1.1, 0.5), centre\_2 = (4.4, 3.2, 2.8, 0.2), and centre\_3 = (5.7, 3.9, 3.9, 0.8) on the same canvas with the **iris.data** data samples as seen below where red dots are iris data samples and black dots are centres.
  3. determines **list\_1** that contains data samples whose nearest centre is **centre\_1**
  4. determines **list\_2** that contains data samples whose nearest centre is **centre\_2**
  5. determines **list\_3** that contains data samples whose nearest centre is **centre\_3**
  6. draws lines from data samples to their nearest centre (Fig. 2 below)
  7. calculates **new\_centre\_1** that is average of data samples in **list\_1**
  8. calculates **new\_centre\_2** that is average of data samples in **list\_2**
  9. calculates **new\_centre\_3** that is average of data samples in **list\_3**
  10. draws lines from data samples to their nearest new centre (Fig. 3 below)

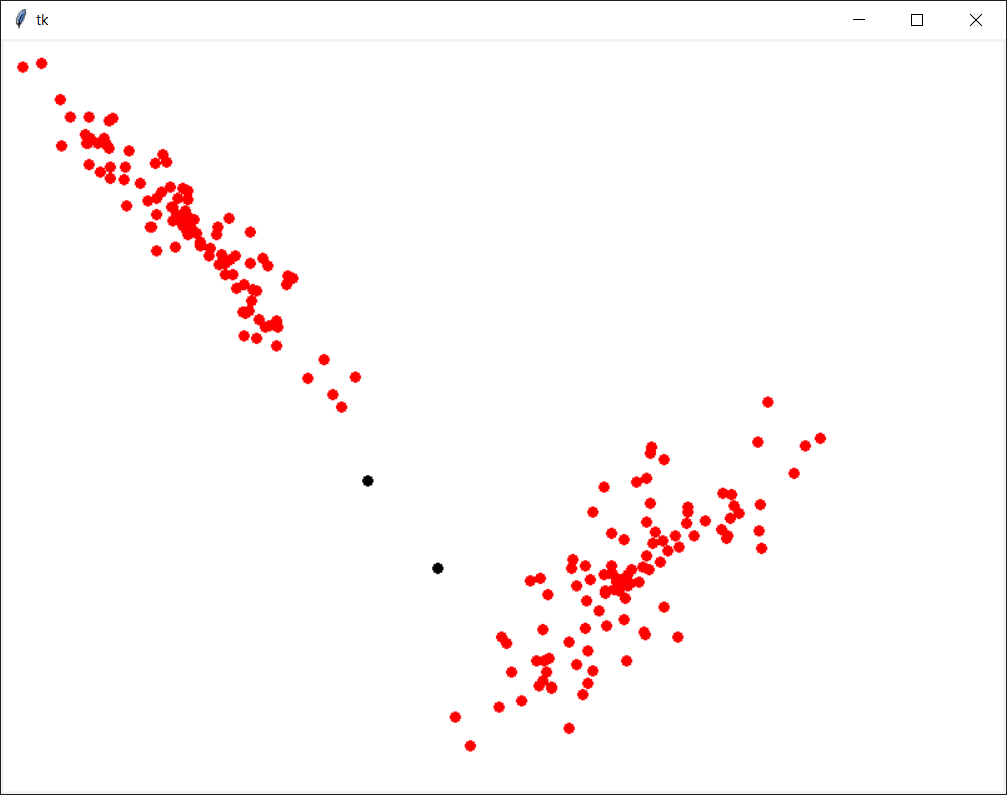


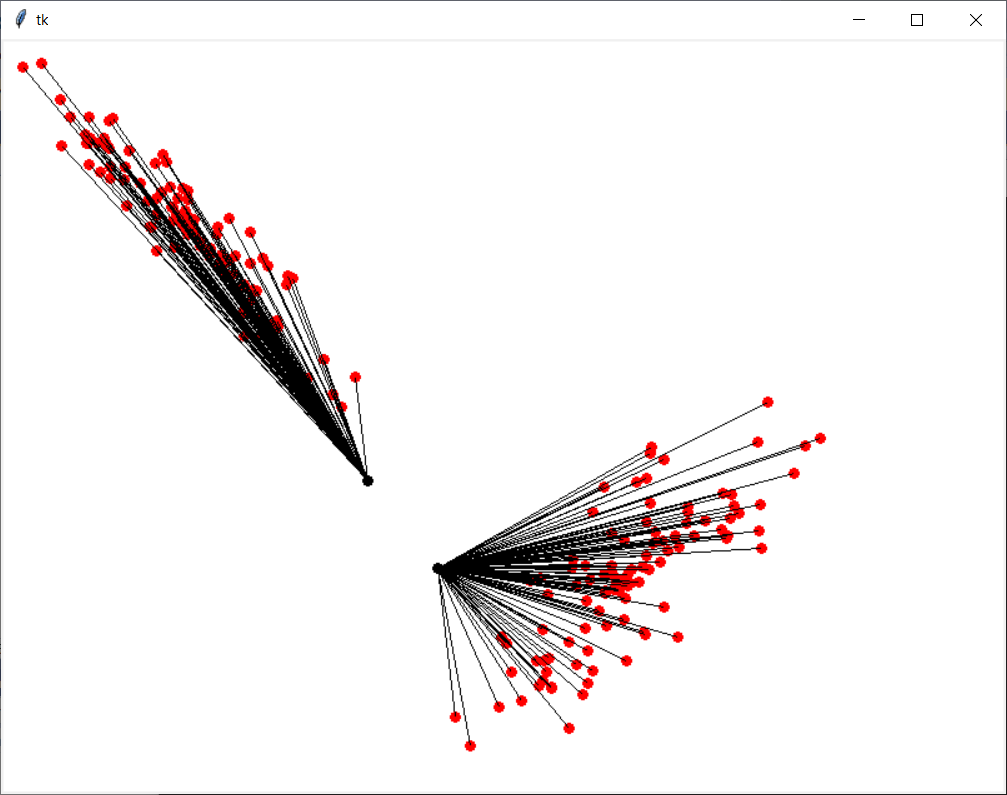


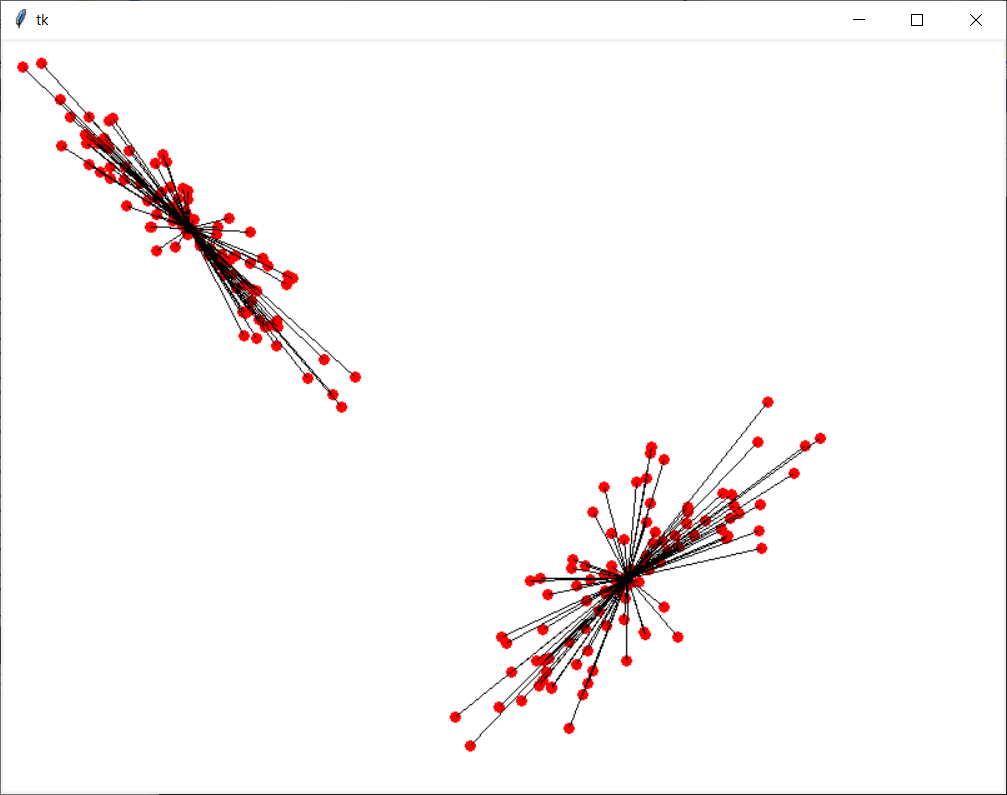


**Question 3:**

* Use the same program in Question 2 and change dataset from **iris.data** to **ellipse1.txt** and change 3 centres to 2 centres **centre\_1 = (2.036779, 2.896883)** and **centre\_2 = (2.836779, 3.896883)** and run the program again (skip step 5 for list\_3 and step 9 for new\_centre\_3). Below are the outputs.







**Total mark for assessment: 3%. Complete and submit all questions: 2%, and lab attendance: 1%. Submit after the due date: -0.5% and -0.5% for each day after. The due date is end of this tutorial session.**