PHYS10003 Project and Professional Skills

‘Big Data’ programming task

Deadline for submission via TurnitIn: Friday 18th January 2019 23:59

Task brief

To write a program in the C programming language to read and process data related to a ‘Big Data’ project. You will be assigned to a team of three or four in which you will work to write and test the code together. Each team will submit a single piece of code but the individual contributions should be highlighted with the use of comment lines.

Big data is a term used to describe data sets which are too large or complex to use traditional data processing tools. Big data is currently a real buzzword with the concepts permeating science and society at large. Increasingly complex physics experiments (think Atlas at the LHC or Advanced LIGO) generate incredible quantities of data every second from a large number of detector channels. In addition, with the advent of social media and devices which are constantly connected to the internet, society is generating vast amounts of multi-parameter data which is being stored and analysed by the corporations running these networks.

In addition to computer scientists, statisticians and mathematicians, physicists are in high demand when it comes to the analysis of these data. The skills, both the hard computative and the soft problem solving, you gain as part of a physics degree lend themselves to applications in which large data sets are involved.

Each of the three data sets can be found in the relvant folder on the Moodle page.

Team 1 : Nuclear masses data set

James Deary
Shaun Donnelly
Connor Douglas
Kim Paterson

The data file set for Team 1 to process is a list of atomic masses arranged in terms of atomic number, *Z*, and mass number, *A*. In total there are over 3000 entries each with three columns of data. These entries represent the known atomic nuclei.

The tasks set for Team 1 are to:
(a) first design and write code which will read this large file and assign each of the quantities for each entry to a variable. The code should be well documented through the use of comments.
(b) process the data such that you can calculate the neutron number for each of the nuclei
(c) write the new data to a separate file in which the neutron number is also now part of the data file
(d) process the data such that the cluster separation energy, S_N, of any nucleus can be calculated by the user where this can be calculated using

$$S_N\left({}^{A}_{Z}Y\right)=\left(M_A\left({}^{A-a}_{Z-z}X\right)-M_A\left({}^{A}_{Z}Y\right)+M_A\left({}^{a}_{z}C\right)\right)c^2$$

where M_A are the atomic masses of the appropriate nuclei.
(e) write the processed data to a file which should inculde proton number, *Z*, neutron number, *N*, and cluster separation energy.

Team 2 : IceCube neutrino observatory data set

Jose Alves
Dion Blackburn
Graeme Clark
Caitlin McGallagly

The data file assigned to Team 2 is a data file containing measurements of muon fluxes taken with the IceCube neutrino observatory based at the South Pole. There are a number of columns of data including declination, right ascension, muon energy loss, muon energy and the date and time of the measurements.

The tasks set for Team 2 are to:
(a) first design and write code which will read this large file and assign each of the quantities for each entry to a variable. The code should be well documented through the use of comments.
(b) process the data such that a user can obtain a histogram of muon energy (remember that a histogram is simply a representation of the number of events which fit within a particular range of values). Ideally the user should be able to specify the range of energy values.
(c) write the histogram data to a new data file which could be used to plot the histogram.
(d) process the data such that an average muon detection rate, and associated uncertainty, can be determined for each day of the experiment.
(e) write the detection rate and associated uncertainty in addition to a representation of the date to a new file which can be used for plotting.

Team 3: Baseball statistics 1980s-2010s data set

Ross McDonald
Alejandro Maza
Greig Oliver
Paolo Sassarini

The data file assigned to this team is slightly different to the others in that the data represented have nothing to do with physics but contain salaries of Major League baseball players in the US from mid 1980s to early 2010s. There are five colums of data pertaining to over 23,000 player entries. The data represented are as follows: the year, a shorthand for the team, the league in which the team play, a player id and their salary.

The tasks set for Team 3 are to:
(a) first design and write code which will read this large file and assign each of the quantities for each entry to a variable. The code should be well documented through the use of comments.
(b) process the data such that a user can determine the average salary for a US Major League baseball player for each of the 27 years of data.
(c) write these data to a new file which could be used for plotting.
(d) process the data such that the player with the highest salary and separately the player with the highest salary to year average ratio can be obtained.