**6CS030 Big Data – Nepal**

**2018/9**

**Portfolio – Part 1**

# Worksheet Two – 10%

# Hand-out: Week 7. Due: Week 8 (See Canvas for date)

This worksheet is based on the MongoDB Wiki and Workbook.

1. On Canvas there are three JSON exports from Twitter.

You need to analyse just one of the JSON datasets.

First take your student number and divide it by 3. Use the ***remainder*** value (modulus) to pick one of the following worksheets:

|  |  |  |
| --- | --- | --- |
| **Remainder Value** | **JSON Dataset to use** | **Dataset Generated From** |
| 0 | lex.json | Lex Limbu – Blogger |
| 1 | tia.json | Tribhuvan International Airport |
| 2 | katpost.json | Kathmandu Post |

For example, if your student number is *1712345*, *1712345/3= 2* so you would use the *kermode.json* dataset. See the *Remainder* spreadsheet if you are not sure how to do this.

1. Examine your dataset and carry out the following tasks:

|  |  |  |
| --- | --- | --- |
| **Task no** | **Task** | **Marks** |
| a | Import the data into your own MongoDB database:   * Show the command to do this (you can just show your password as *myPassword*.) * Write a command to show how many documents are in your collection | 20 |
| b | Analyse the data  Write a command to:   * Show one document * Show the unique values in one field * Shows a set of documents based on some criteria. Output just two fields from the document. * Use a regular expression to search for some criteria. The search should be case insensitive | 40 |
| c | Reshape the collection  Write a command to:   * Update a field within the collection * Create a new collection based on a subset of the dataset. Include a query to show a document from the new collection | 20 |
| d | Name one advantage to using this approach for handling Big Data and include brief explanation of why you think this is an advantage. | 10 |
| e | Name one disadvantage to using this approach for handling Big Data and include brief explanation of why you think this is a disadvantage. | 10 |

See Marksheet for further marks breakdown.

Note this is an individual assessment. Any group answers will be classed as plagiarism.

For this exercise you can either use the Mongo Shell or Python Notebook to carry out the commands.

## Python

If using a Python Notebook you will not be able to use the command to import data within the notebook, however, you can document what command you ran in your notebook.

You can use the Print Option to create a PDF version of the file and upload it. Do check that it has printed all the pages and not just the first page (if so, submit the notebook).

## Upload

Upload one of the following: a Word Document, .ipynb file or PDF version of the Python Notebook on Canvas which shows evidence of the above tasks.

You must use the dataset allocated otherwise 0 marks will be allocated.

## Hand-in/Resits

The Hand-in date is the normal teaching day in Teaching Week 8 (w/b 11th March), unless you have been granted an extension through the normal University channels.

If you cannot make this for any reason and have not been granted an extension, or achieved less than 40% in the Worksheet, there will be a final hand-in date for the Worksheets on the 7th May.

Do note unless you have mitigation the mark for each worksheet submitted on the 7th May will be capped at 40%.