Advanced Data Science Lab 01

Hands on Python

Total Marks 10

Activity A (5 marks)

For this lab session you are expected to practice python code that was demonstrated during the lecture session – these python notebooks are

- Lecture 2 - Python Hands-on (Python 3.x)
- Introduction to Python – Practice
- iPython Shell Commands

Online iPython notebook can be found on

- https://colab.research.google.com/
- https://try.jupyter.org/

Google Co Laboratory provides a more reliable python instance then jupyter. For Co Laboratory you will need a google account (i.e., gmail). You can use your personal account.

You can choose to practice using IDE locally installed or online; however, you need to extract the code first above mention notebooks. Also, iPython Shell Commands notebook will only work for iPython.

Please make sure you cover basic syntax, lists, conditional, iterative statements, function, and develop a comprehensive understanding of list, dictionaries, indexing and slicing.

Note: Please note, for this task you are only required to practice the aforementioned notebooks - no deliverable or report is expected from this task.

Activity B (5 marks)

You are expected to write a python code for counting words in a string i.e., count how many times a word appears in a string.

Your final code must fulfil the following requirements:

- You code should accept user input as a string
- Use **dictionary data structure** to hold words and their respective counts
- Special characters must be removed from words before adding them to the dictionary data structure
- Numeric values must be counted as single entry in the dictionary regardless of their values. For example in “I have £35 to spend on 5 different items” your code should count numeric values as “2” this is because two numeric values has appeared in the string.
- You code must not differentiate between uppercase, lowercase and init-cap words i.e., “UNIVERSITY”, “university” and “University” must be considered same.

Some hints for your exercise:

You can declare a dictionary structure using:

```
wordCount = dict()
```

You can split a string into a list of words which you can be processed individually.

```
words = text.split()
```

Convert a word into lower case and then process it

```
word.lower()
```

Write a function which takes a word as in put a get rid of any special character in it i.e., full stop, comma etc. For that you may need to convert a word into character list and process individual character to identify whether it in special character list.

```
specialChar = [',', '.', '@', '&']
    clearWord = []
    for char in word:
        if char in specialChar:
            print 'skip character!'
        else:
            clearWord.append(char)
```

There are many ways to complete this task. If you divide that problems into sub problems and work on them one by one you will complete it in no time – don't try to solve the complete problem in one GO (*subject to proficiency in python!*)

Activity C (optional)

Building on Activity B, rather than taking "user input", you can read data from Wikipedia i.e. Web Scraping. Once you have dumped the data, perform the following:

- Count how many times a word appears by ignoring stop words such as "the", "a", "an", "in"
- Identify the total number of hyperlinks in the dumped data.
- List all references in the dumped data

NOTE: this task is intended only for these how are relatively proficient in python.

Lab submission

You are expected to submit the code of Activity B, along with the screenshot of its output.

Submit a single (MS Word or PDF) file containing the code, and the screenshot.

The submission is through Trunitin.

Your submitted code must be commented - explaining the code and logic line by line.

Submission due date: *please refer to moodle site.*