**ENSF 612: Assignment 1**

Marks: 15

The assignment is worth 5% of course marks

**Instructions:**

1. This assignment must be completed individually. You cannot copy from others or consult with others. Any identification of such unauthorized actions may result in 0 grade for this assignment.
2. This is a take home assignment.
3. Each student will use Databricks community edition notebook to write the assignment.
4. Once completed the student will share the notebook as a link in PDF and upload the PDF in the D2L dropbox folder named "Assignment 1".
5. To answer each question, use the notebook feature markdown to write your texts and the code blocks to write your code. All code needs to be executable. The grading will be done by reading the texts and the code and by running the code.

**Question 1 (5 Marks)**

Suppose you have a job to do word counting on a large amount of textual data. You should only return words with at least 100 counts. The data could be found in two files (filename1, filename2). Each file is 2 TBs. You have found that your own computer has only 8 GB of RAM. Is there anything wrong with the following Spark code that could prevent it from running in your own laptop? If so, how can you fix it?

```
from collections import defaultdict

filteredKV1 = readFile(filename1)
filteredKV2 = readFile(filename2)
```

*// The following pyspark method is called to read two files, do word counts on each file, and to return the word counts*

```
def readFile(filename):
    infile = sc.textFile(filename) // here assume that sc is SparkContext
    counts = infile.flatMap(lambda line: line.split(" ")).map(lambda word:
    (word, 1)).collect()
    return doFilter(counts1)

def doFilter(count):
    key_val = defaultdict(int)
    for item in counts:
        key = item[0]
        val = item[1]
        key_val[key] += int(val)
    filtered_key_val = dict()
    for k, v in key_val.items():
        if v >= 100:
          filtered_key_val[k] = v
    return filtered_key_val
```

**Question 2 (10 Marks)**

Write a program in **pyspark** that will use the following three files:

There are three files with 150,000 questions that are asked about three programming languages in Stack Overflow, java, python, and javascript. The files are shared in D2L (Assignment 1 files).

- SO-Spark contains 50,000 questions from Stack Overflow that are tagged as `apache-spark'.
- SO-ML contains 50,000 questions that are tagged as `machine-learning'.
- SO-Security contains 50,000 questions that are tagged as `security'.
- The posts are collected from Stack Overflow posts table. Details about Stack Overflow posts table can be found here: https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

You will code in pyspark to answer the following questions

1. Write a function to load each of the CSV files. (5 marks)
    a. Input to the function: filename
    b. Output from the function: a pyspark dataframe with the file contents.
2. Write a function that will show the total ViewCount per file (5 marks)