ENSF 612: Assignment 1

Marks: 25

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. You have until Oct 18 to submit this.
3. Each student will use Databricks community edition notebook to write the assignment. Once completed the student will share the notebook with the TAs and with Dr. Uddin
4. 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)

Assume the size of the file.txt below is 100 GB.

Is there anything wrong with the following Spark code? If so, how can you fix it?

```
from collections import defaultdict

text_file = sc.textFile("file.txt")
counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).collect()
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 (20 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-Java contains 50,000 questions from Stack Overflow that are tagged as `java’.
- SO-Python contains 50,000 questions that are tagged as `python’.
- SO-Javascript contains 50,000 questions that are tagged as `javascript’.
- 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 (each question has 5 marks)

1. How can we preprocess the textual contents in the files?
    a. Write a short description of how you can answer this and then write a short program to answer this question.
    b. Write a function to tokenize and remove stop words from each of the files.
    c. Write a function to remove any other noise in the text files (first define what is a noise in the texts and then write code to remove the noise)
2. What are the most frequent keywords in the textual contents of each programming language?
    a. Write a short description of how you can answer this and then write a short program to answer this question.
3. What percentage of questions in each programming language has accepted answers?
    a. Write a short description of how you can answer this and then write a short program to answer this question.
4. What types of questions are asked for each programming languages?
    a. Write a short description of how you can answer this and then write a short program to answer this question.
    b. E.g., we can say a question can be of four types: How (e.g., how to solve this?), What (e.g., what is a recommended way of solving this?), Why (e.g., why is my program crashing?), or Other (everything else)
    c. To check for ‘why’ questions, you can whether the question has started with \`why’ word. You can apply similar rules for find “what” and “how” type of questions.

Hint:

1. Use Python BeautifulSoup to parse HTML files. https://www.crummy.com/software/BeautifulSoup/bs4/doc/