**ENSF 612: Assignment 2**

Marks: 20

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 with a link and in a PDF and upload the PDF in D2L dropbox folder "Assignment 2"
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.

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 preprocess each file data as follows (by processing the column Body). (5 marks)
    a. Remove stop words
    b. Remove all punctuation marks
    c. Remove all non-alphabetic characters
2. Write a function that could inform of the potential topics in the files as follows (by processing the column Body). (8 marks)
    a. For each file, find the top 10 most frequent keywords (after preprocessing above) and visualize it using a bar chart
    b. Across the three files, show the overall top 30 keywords (i.e., across the three files) and visualize it using a bar chart
3. Write a function to show for each file the percentage of questions that does not have an accepted answer. (3 marks) (hint: use the column "AcceptedAnswerId)
4. Write a function to show for each file the total amount of posts created by year and visualize it with a bar chart (4 marks) (hint: use CreationDate to infer the year)

Hint:

1. Use Python BeautifulSoup to parse HTML content of the column Body (i.e., the column . https://www.crummy.com/software/BeautifulSoup/bs4/doc/