Project 2 FAQ

1. **About vocabulary, how should I optimize my vocabulary?**

   First, you need to use tokenize and lemmatization (not stemming).
   Second, optimize your vocabulary. Here are some parts you must consider in your program.

   - If the word is Ask_HN, you vocabulary should contain ***Ask_HN*** NOT ***Ask, HN***.
   - If the titles are "Data Science is a subject" and "I chose Computer Science as my major", you vocabulary should contain ***Data Science, Computer Science*** NOT ***Data, Science, Computer***.
   - If the title includes some words, which you think is not useful for classification, you can remove it from you vocabulary, but you need put all the removed words in an independent file named "remove_word.txt". Please write it clearly in your report why you remove these words.
   - Your program should be able to output a file named "vocabulary.txt", which should contain all the words in your vocabulary after optimization.
   - Please note your training and testing datasets should use the same optimizations.

2. **For task1, the results of probabilistic is too small, using what format to represent it in the file?**

   For the result of task1, please write the frequency of each word in your vocabulary and the probability using $\log_{10}$ leave 10 digits after decimal no rounding off. Your program only need to output **one** file named "model-2018.txt" NOT three files.

3. **During demonstration, will the testing samples have the same format with the same type?**

   Yes, the testing sample will have the same format with the same four class types.

4. **Is there a limitation for the running time of my program?**

   Due to the size of the datasets provided, there is no limitation when you train the dataset of 2018 and test the dataset 2019. But during your demonstration, we will give you a small datasets, which will contains less than 100 samples and your program should be able to analyze it within 1 min on lab machines.

5. **Which libraries you can use for project2?**

   All the libraries listed in the project2 description you can use to implement your code.
   Please note re is not in the description of libraries.

Last Modified: Monday November 11, 2019.