CS280 Course Project (group of at most 5)

Due:
Code and report: 23:59 Tuesday Dec 24, 9:00
Presentation: Before class Thursday Dec 26

We have learnt we need quality data for data mining, but in reality data is dirty. In this project, you will preprocess the web page documents to remove any html tags and identify content terms of these documents.

You should also remove stop words and perform stemming on each word after identifying all words. Your report and presentation should include the methods, techniques, and programs you developed for the preprocessing and the results:

1. Report should be formally organized, formatted, and written without spelling errors.
2. Report and presentation should address the following:
    1) What libraries and APIs you have used, their websites, and an introduction to them.
        - Provide references at the end of report.
        - You are free to do a research and use any online packages that can help you to finish the job.
    2) Review the key parts of your code to show how your code works on removing html tags and stop words and on performing stemming. What extra measures you have taken to achieve better outcomes.
    3) Any problems you have encountered and solved, and valuable lessons you have learned.
    4) This preprocessing should include documents from both training set and test set as a whole.
    5) Results:
        - Present total number of words identified.
        - Present and discuss top 200 most frequent words identified and their frequencies (including both training set and data set as a whole)
            - Example words that are not properly cleaned, e.g. sle"), il'-10, etc.
                - Why are they difficult to get properly preprocessed?
            - What is the percentage of words that appear in both the training set and the test set (#of common words/total # unique words)?
        - Present and discuss bottom 200 least frequent words identified (including both training and test sets)

The project is evaluated based on four aspects: programs implemented, results presented, report written, and presentation (10 minutes).