HomeWork 1:

1. Github

Please follow the instructions:
1.1 in your github directory (***/Tian), create a .txt file, type anything in the file, save and exit.

1.2 use git-bash, cd to the github folder (***/Tian)

1.3 in git-bash, use the following command to check the git status(do not include the ':') : git status
you will find git recognize the new file, it suggest you to add the new file to index

1.4 in git-bash (without :) : git add yourfile.txt
1.5 in git-bash, check your status: git status
1.6 in git-bash, commit to repository: git commit -m 'anycomments'
1.7 push to remote: git push
1.8 in the github website, check your repository, find the new .txt file

here, we only have one branch. You can always add your own branch, edit it, review it, and then merge it to the master branch. We will get to it later in the class.

reference:
A. github flow: https://guides.github.com/introduction/flow/

2. Data Cleaning
In https://www.lintcode.com/ai/movie-review-recognition/overview, download the data

Please finish the following task, with labeledTrainData.tsv:
2.1 in the same folder of your data, open a jupyter notebook using anaconda prompt

2.2 in the notebook: import pandas, numpy, nltk, string (I'll talk about the pandas in detail next week, you can try it a little bit)

2.3 use the same method I used in class, remove the stopwords and punctuation in the movie review

2.4 save the rest in any datastructure (dataframe prefered, pure text is ok).

2.5 please report, how many comments you have (hint, use pandas dataframe, you can check the shape of the dataset by using the .shape method)

3. Data frequency
use the review in the labeledTRainData.tsv (only need the reviews), use methods in nltk,
3.1 create the text corpus (remember the trump.text?)
3.2 use nltk.FreqDist(text) to create the table of word frequency, report the top 10 word2vec (most_common)
3.3 use nltk.collocations to get the most frequently used 3 words phrase.

Please finish the report, and submit it to Di Rang.