Assignment 3

This is individual assignment, you required to do following:

- Retrieving data from http://scrapsfromtheloft.com/2017/05/06/louis-ck-oh-my-god-full-transcript/, this is one example, please grab at least 5 URLs from same website, extra data from <div class=”post_content”>, there will be list of paragraph with html tag <p>
- Save retrieved data into local disk with pickle library of Python
- Clean data by performing following tasks:
    - Make text all lower case
    - Remove punctuation
    - Remove numerical values
    - Remove common non-sensical text like /n
    - Tokenize text
    - Remove stop words
- Create a document-term matrix using CountVectorizer from sklearn and exclude common English stop words.

Submission:

- Source code
- Result
- PPT

Deadline: Oct-18-2019 midnight, random selected students will do presentation on Oct-19-2019 classes.

Dr. Zhijiang Chen
BJU@CHEN.ME