LING 570: Hw1
Due date: 11pm on Oct 4, 2018 (Thurs)

For this homework, you are going to write an English tokenizer and a tool that creates a vocabulary from the input text. All the sample files are under
~/dropbox/18-19/570/hw1/examples/.

Q1 (50 points): Implementing an English tokenizer, **eng_tokenizer**.sh

- Format:
  - The command line is: cat input_file | ./eng_tokenizer.sh abbrev_list > output_file
  - abbrev_list is an input file. It contains a list of abbreviations, one abbreviation per line.
  - The input and output files should have the same number of lines, and the k^{th} line in the input corresponds to the k^{th} line in the out file.
  - The tokens in the output lines should be separated by the whitespace.
  - A sample input file is "ex1", and a sample output file is "ex1.tok". The sample output file is meant to show you the format, NOT the gold standard.
- Note:
  - Your tokenizer should not separate numbers, urls, paths, etc. See the slides for 9/27's lecture.
  - You can assume that a token will not cross the line boundary; therefore, your code should process each line independently of other lines.
  - Do not merge the tokens in the input text (e.g., the collocation expression such as "pick up", "because of", "Hong Kong" should NOT be merged into one token).

Q2 (15 points): Writing a tool, **make_voc.sh**, that creates a vocabulary from the input text.

- The command line should be: cat input_file | ./make_voc.sh > output_file
- The tool reads in each line in the input, breaks it into tokens by whitespace only, and output the frequencies of the tokens.
- Each line in the output file is a (token, frequency) pair. The lines are sorted by the frequency of the tokens in descending order.
- A sample input is "ex1", and a sample output is "ex1.voc".

Q3 (10 points): Run the code in Q1 and Q2

- Run the following commands:
    - cat ex2 | ./eng_tokenizer.sh abbrev-list > ex2.tok
    - cat ex2.tok | ./make_voc.sh > ex2.tok.voc
    - cat ex2 | ./make_voc.sh > ex2.voc

- In your note file, write down
    - the numbers of tokens in ex2 and ex2.tok
    - the numbers of lines in ex2.voc and ex2.tok.voc

Submission instruction:

- Submit two files, readme.[txt|pdf] and hw.tar.gz, as specified in the course policy.
- The note file, readme.[txt|pdf], should include the answers to Q3 and any note that you want us to read.
- hw.tar.gz should include all the files specified in ~/dropbox/18-19/570/hw1/submit-file-list, plus any source code (and corresponding binary code) called by the shell scripts.