Assignment 1

Counting and sorting

[Description]

In this assignment, you need to write a program to sort the lines in ascending order according to their numbers of words. Each line may include both Chinese and English words. Each Chinese character is counted as one word and each English word is counted as one word. English words are separated by spaces and Chinese characters. Spaces are characters but not counted as words, and you cannot remove them in your output file.

Here are some examples:

The word count of “看NBA” is 2;
The word count of “看N B A” is 4;
The word count of “再send email” is 3;
The word count of “google forms 查詢” is 4;
The word count of “linkedin的創始人” is 5;
The word count of “爱行山 愛健身” is 6;

When two lines are having the same word count, they should be sorted in ascending order according to the Unicode of their first characters. If the first characters are the same, then compare the second characters…. etc.

You can assume all the input text is in UTF8 format.

Please remember you have to sort the lines without modifying the content of each line.

[Instruction]

Most probably you will need python to finish your task, but you are free to use any Linux command to ease your work.

Here are some common Linux commands you may find useful for text processing:

grep
awk
sort
uniq
xargs
sed
cat

Pay attention to the options of these commands.

You have to package your work into a shell script named “sort.sh”.
Your script should be used with the following usage:

```
./sort.sh  input_text  output_text
```

where input_text and output_text are the input and output files respectively.

There are a sample input file and a sample output file for you to verify your work.
There are placed at /data/cs310/baibing/assignment1
Your program is allowed to create any intermediate files. The running time of your script should be within a minute under our server configuration.

[Submission]

You have to create a folder named “assignment1” under your working directory at /data/cs310/XXX/. Thus, the path of your folder should be /data/cs310/XXX/assignment1/, where XXX is your login name.
Place the script file in the folder, then use chmod to turn off the read access of your folder so that others can’t read your solutions.