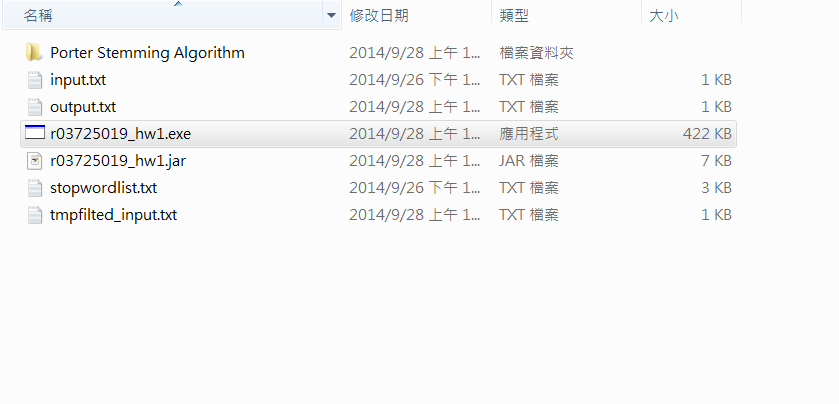
IRTM\_HW1\_Report

1. **Execution Environment**

For your TA convenient, I wrap my code into execute program (r03725019\_hw1.exe).You can easily double click it to run my code on Windows OS.

Note 1 : if you are not using Windows OS, you have to import my source code project (Porter Stemming Algorithm) into Eclipse environment, compile and run it.

Note 2 : I have already run my program and got the output file. You can remove tmpfilted\_input.txt and output.txt and run my program again to get the same result.



1. **What's your design of your program?**

* Input : input.txt , stopwordlist.txt
* Output : tmpfilted\_input.txt , output.txt
* Use Data Structure : StringArray , Arraylist
* Design concept :

According to homework requestment, I tokenize input document, lowercase all word and remove the words which are in stopwordlist. And then save to a temporary file.

Next, using Porter Stemming Algorithm to stem the temporary file to get the output result. Finally save the result to output file.

1. **Do you encounter any difficulty during the task? How do you solve it?**

If you stem word first and then remove stopword, the result may have some strange noise words (e.g. ar,s). These noise words seem have no reference value.

So, I try to remove stopword first, and then do the stemming. The result of this choice get the better output result.

