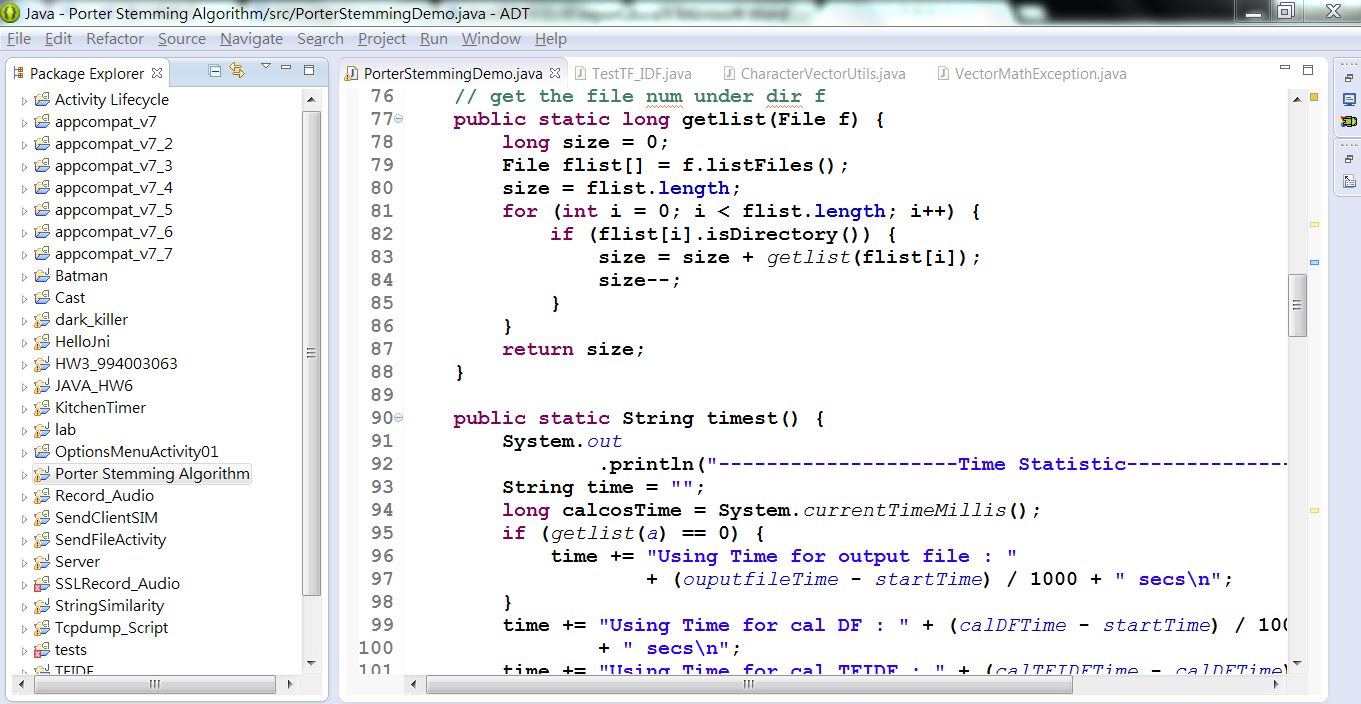
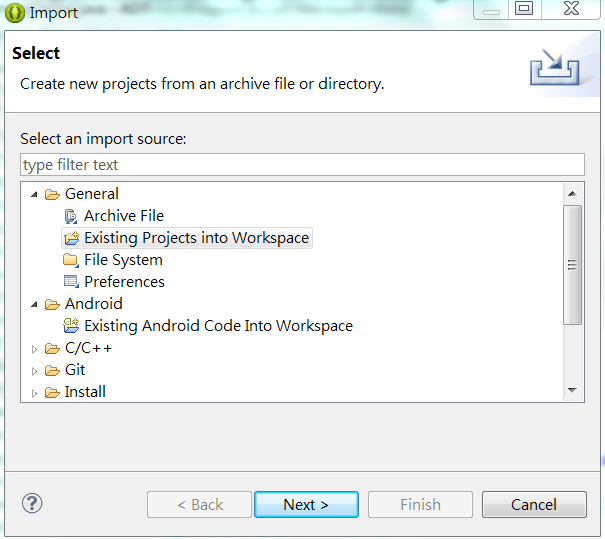
IRTM\_HW2\_Report

1. **Execution Environment**

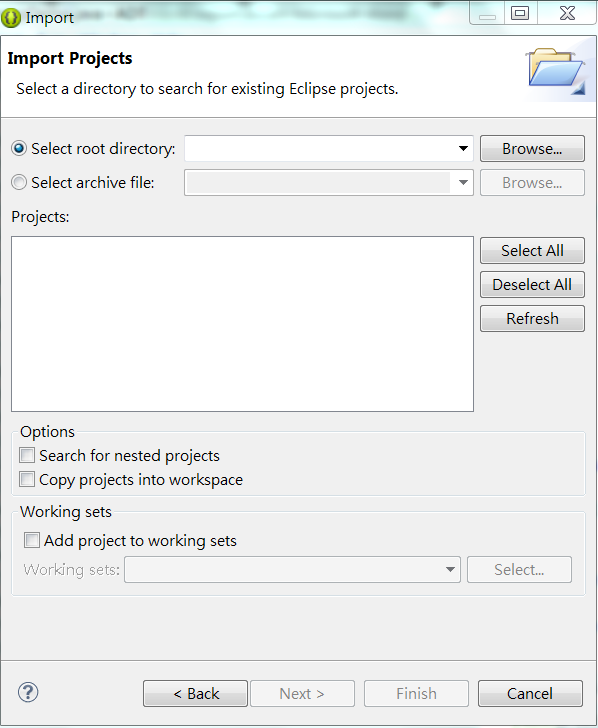
Please import my project directory “Porter Stemming Algorithm” into your eclipse IDE. The following picture shows you how to import my project:



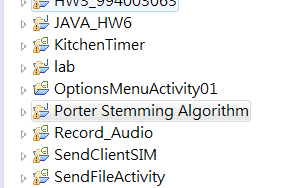
Open Eclipse IDE，Choose 【File】🡪 【import】



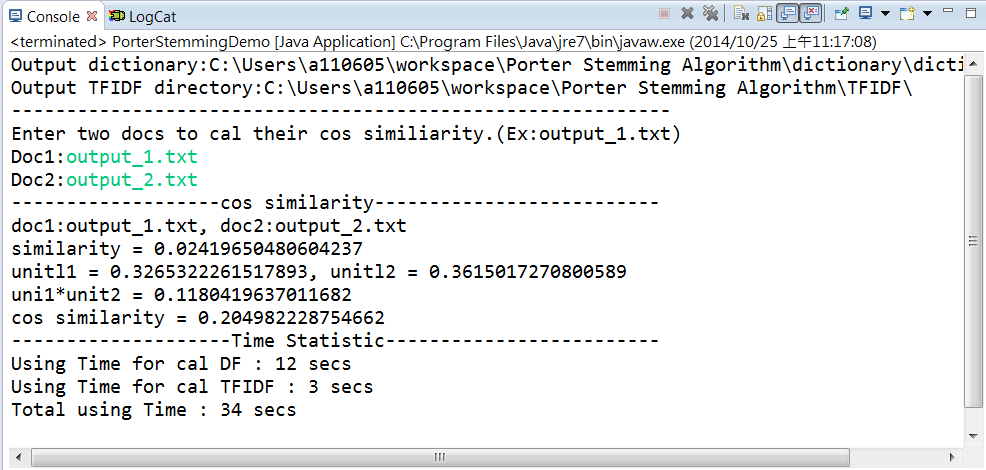
Choose【Existing Projects into Workspace】



At 【Select root directory】press【Browse】to select my project 【Porter Stemming Alogrithm】



Then my project will import into your workspace , please press 【ctrl+F11】to run my program.



The result is on console.

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

Input : IRTM file collection

Output :

* Dictionary.txt , which contains Documents Frequency
* TFIDF files , which contains every term’s TFIDF in each document.
* Cos similarity, calculate the cos similarity between two docs which you entered.

Programming Language : java

Data Structure : Vector、Treemap、Hashmap

Design Thinking : using porter stemming algorithm to stem the input documents, save them into vector, calculate all document’s DF and TFIDF of each doc. Finally, you could enter two docs to compare their cos similarity.

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

When calculating TDIDF , I use the formula on teacher’s PPT.

IDF=TF \* IDF, where TF= the number of occurrences of term *t* in document *d*)

But the output result seems a little strange. So I choose to use the formula on Website

TF (term) = frequency of the term in the document / no. of terms in the document

Then the final TFIDF results get pretty perfect.