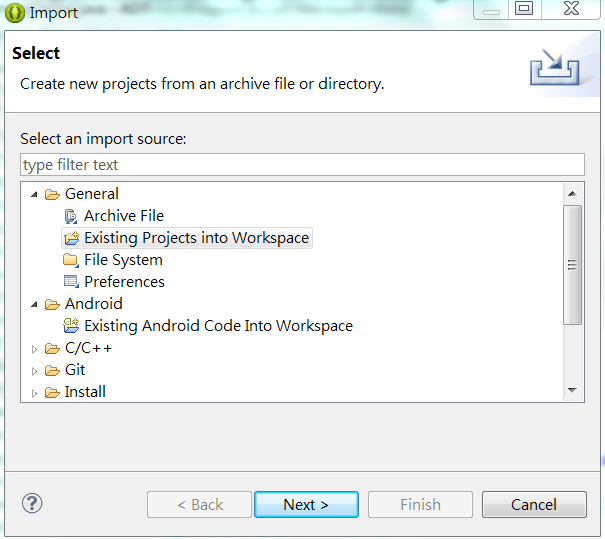
**Information Retrieve Homework 4 Report**

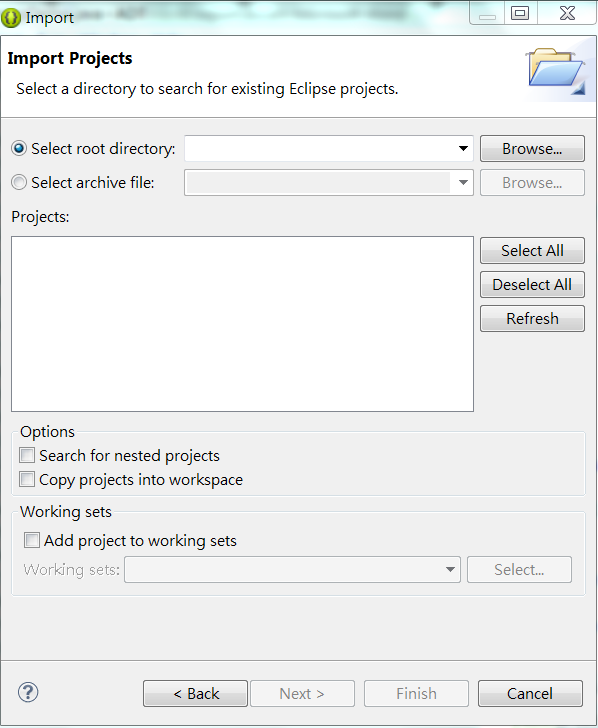
**R03725019 李士暄**

1. **Execution Environment**

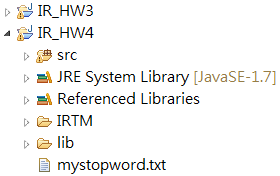
* **System requirement**
  + OS : Windows 7
  + JAVA JDK : 1.7.0\_25
  + IDE : Eclipse
* **How to execute my program**
* Open Eclipse IDE，Choose 【File】🡪【import】，Choose【Existing Projects into Workspace】



* Press【Browse】to select project【IR\_HW4】



* Please press 【ctrl+F11】to run program.



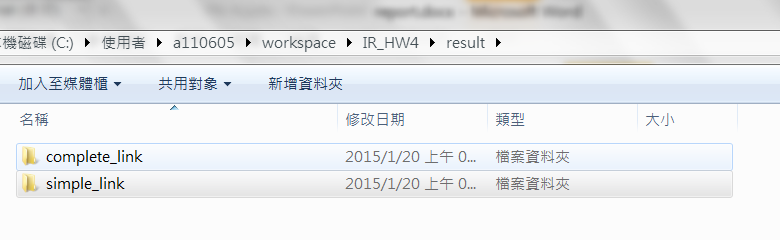
1. **What's your design of your program?**

* **Input :** IRTM file collection with 1095 documents
* **Output :** 8.txt、13.txt、20.txt
* **Programming Language :** java
* **Data Structure** : ArrayList、Hashmap
* **Design Thinking :**

First of all, I prepare the preprocessed documents which are stemmed and removed stopword. Then use HAC algorithm to cluster documents from bottom to up. According to PA4 requirements,

* Documents are represented as normalized TF-IDF vectors.
* Cosine similarity for pair-wise document similarity.

I have tested the similarity by **single-link** and **complete-link**. Finally, save to file when clusters number is 8、13、20.(Under *IR\_HW4\result*)



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

When I cluster documents , I use single-link first. But the result shows that it becomes a huge cluster and the others tiny clusters which are few documents in it.

So, I try another method -- complete-link. The result shows more perfectly. Each cluster is about 50~70 pieces of documents. It seems average to every clusters.