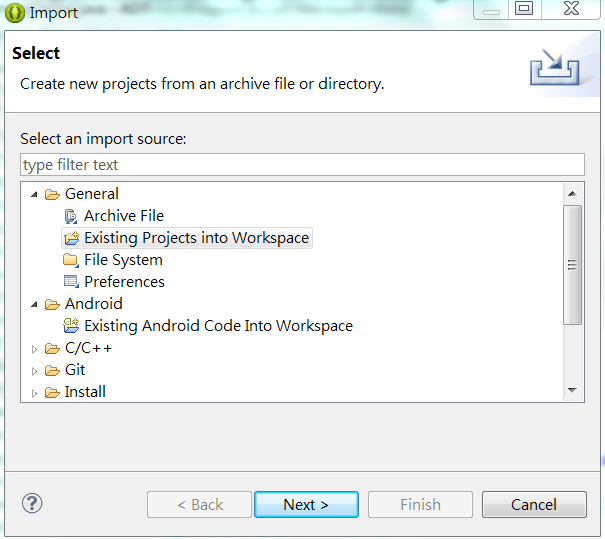
**Information Retrieve Homework 3 Report**

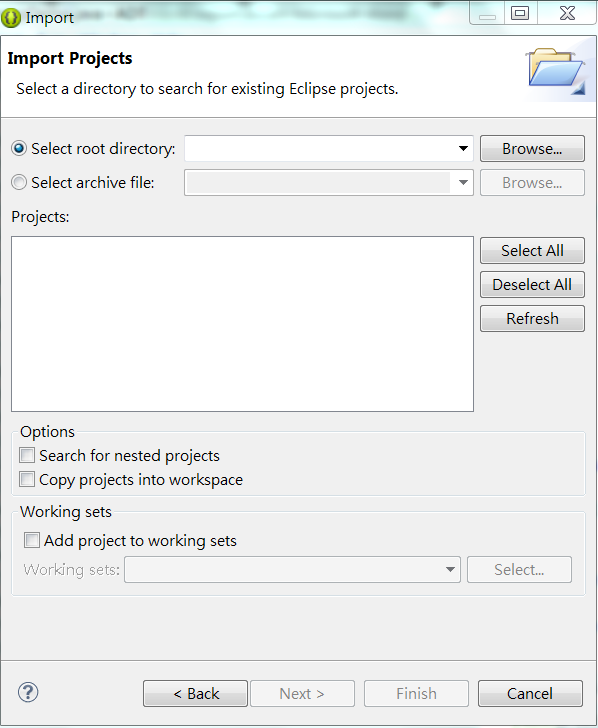
**R03725019 李士暄**

1. **Execution Environment**

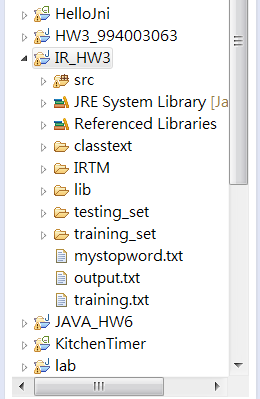
* System requirement
  + OS : Windows 7
  + JAVA JDK : 1.7.0\_25
  + IDE : Eclipse
* How to execute my program
* Open Eclipse IDE，Choose 【File】🡪【import】，Choose【Existing Projects into Workspace】



* Press【Browse】to select project【IR\_HW3】



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



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

* **Input :** IRTM file collection with 1095 documents
* **Output :** output.txt
* **Programming Language :** java
* **Data Structure** : Vector、Treemap、Hashmap、Documents、featureStat、NaiveBayesKnowledgeBase
* **Design Thinking :**

　First of all, I stem and do stopword removal to all documents and save them into new document collections. I separate 1095 docs into “training\_set” and “testing\_set” .Then using “chi-square” method to do feature selection in training\_set. You can see selected feature in my project (selected\_feature.txt).Second, I train all docs in training\_set to get log prior and log P(x|c).And save the information into data structure. Third, I test testing\_set by calculating C MAP probability for each testing doc to get the argmaxscore of classes. And save results into output.txt.

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

During getting selected feature, I used “chi-square method” to calculate their X^2 measure. But most of them get the same scores. It seems that something goes wrong in my chi-squared procedure

. Finally, I cannot solved this problem. So I decide to choose 2833 features to do training process. The final classification results seems great by doing so.