

**School of Information Technology**

**FYPJ 2013 P3**

**Social Media and Analytics**

**(Topic and User Profile)**

**Project Documentation**

**Hong Kay San 116853D**

**Lim Boon Leng 112723X**

Contents

[1. Project Objective 3](#_Toc372808006)

[2. Project Specifications 5](#_Toc372808007)

[3. Project Implementation 10](#_Toc372808008)

[4. Sources and Manuals 12](#_Toc372808009)

[5. Supervisor Sharing 21](#_Toc372808010)

[6. Project Team Sharing 22](#_Toc372808011)

[7. How it works 23](#_Toc372808013)

[8. Results Analysis 26](#_Toc372808014)

[9. Discoveries 31](#_Toc372808015)

# 1. Project Objective

**FYPJ Period:** 2013 Period 3

**Location:** L.336 [Seats 13 – 14]

**Duration:** 2 Sept 2013 – 22 Nov 2013

**Supervisor:** Mr James Goh

**About Us:**

**Team Members:**

* Hong Kay San 116853D
* Lim Boon Leng 112723X

**Introduction**

Creating an automated cluster evaluation program that allows users to find out which clustering schema or algorithm is more preferable for documents clustering.

As a continuation from the previous group, by making use of Mahout, an open-source program that contains many clustering algorithms, the process of clustering the forum posts / documents will be handled by it.

In order to create an accurate and efficient evaluation program that works well with Mahout, a lot of research into Clustering, Cluster Evaluation and Mahout is required, in order to integrate both together. The best cluster evaluation metric would have to be determined, and making Mahout’s output suitable for input into the evaluation program would be required.

To get us started with the project, a good understanding of how clusters work is required. The research documentation can be located in the document ‘Clustering Research Documentation’.

The research results and details for the best Cluster Evaluation Metric can be located in the document ‘Metrics Research Documentation’, including how it works, its computation algorithm, and reasons of its suitability and superiority amongst other available evaluation metrics.

Thereafter, a results analysis is done on the clustering algorithms, to finalize the best clustering algorithm for documents clustering, by making use of the results computed by the cluster evaluation metric. The report is available at ‘Results Analysis Documentation’.

**Objectives:**

To help organizations do their target audience research more effective and efficiently, aiding in their ability to identify key topic trend. Traditionally in marketing, researches are done through surveys; questionnaires information is freely available in the internet but why?

The aim of this project is to make use of the World Wide Web (WWW) as our target market audience to find the latest trends in the real world. In the cyber world, there are chat rooms, forums and other social media platforms. However, there's always an 'anonymous wall' between the users and the computers. This causes users to experience the *online disinhibition effect.*

What is the online disinhibition effect?

"The online disinhibition effect is a loosening (or complete abandonment) of social restrictions and inhibitions that would otherwise be present in normal face-to-face interaction during interactions with others on the Internet." & "Because of this loss of inhibition, some users may exhibit benign tendencies such as: becoming more affectionate, more willing to open up to others, less guarded about emotions, all in an attempt to achieve emotional catharsis." - Source: Wikipedia.org

Therefore, we can conclude that the internet is the best place to obtain feedbacks from users from all age groups, specifically forums websites.

In order to identify topic trends or hidden topic, by using the method of machine learning clustering algorithms to re-organize and re-categorizes all the post in the forum into topics, topic names generated by topic modeling algorithms i.e. Unsupervised Clustering.

The tool that was used to identify topic trends and hidden topics was Apache Mahout using two of its clustering algorithm **Latent Dirichlet allocation** (LDA) and **K-means**.

Author clustering is another type of clustering. It can be another choice of clustering for user to view in forum. Instead of cluster by topic and threads, it can cluster by author’s name. It is to allow user to have another view of clustering in forums instead of topic titles. This is done by manipulating the crawl data from the database using SQL statements.

In order to find out which clustering algorithm is more suitable / efficient in clustering the forum posts, an internal cluster evaluation metric, **S\_Dbw Validity** Index was used. Thorough research revealed that this metric provides the most accurate results, as compared to other internal evaluation metrics.

Internal evaluation metrics is preferred in our project scope, as it was stated as a requirement that no human intervention/manual work should be done for evaluating the clusters.   
(External evaluation metrics require a ‘benchmark/gold standard’ set by a human)

To integrate the final product of the evaluation program and Mahout’s output vectors together, some manipulation and formatting is required to be done to the output of Mahout, before using it as input for the evaluation.

# 2. Project Specifications

**Introduction to Cluster Analysis**

Cluster analysis / Data clustering

The task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Document clustering (or Text clustering)

Automatic document organization, topic extraction and fast information retrieval or filtering.   
It is closely related to data clustering.

Research Outcome:

What we are focusing on: **Document Clustering**

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster.

In general, there are two common algorithms.

The first one is the hierarchical-based algorithm, which includes single link, complete linkage, group average and Ward's method. By aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing.

Another representation of Document Clustering:

Flat Algorithms

* Usually start with random (partial) partitioning
* Refine it iteratively
  + K-means clustering

Hierarchical Algorithms

* Agglomerative (Bottom-up)
* Divisive (Top-down)

Hard clustering

* Each document belongs to exactly one cluster

Soft clustering

* Document can belong to more than one cluster

Hard vs. Soft clustering

Soft clustering is more relaxed, allowing the values to be part of more than one cluster whereas hard clustering approach only assigns a value of 1 or 0.

The higher the value is, the higher the degree of confidence that the objects belong to that cluster.

Flat vs. Hierarchical

Flat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other. Hierarchical clustering creates a hierarchy of clusters.

Flat clustering, each cluster does not relate to each other. Hierarchical is clustering that has sub-clustering of another clustering. Eg: A main thread on healthcare with sub post on nurse and doctor.

**\*\*The full detailed documentation is located in ‘FYPJ 2013 P3/Documentations/2013 P3/Clustering Research Documentation’\*\***

**Introduction to Cluster Evaluation**

Evaluation of Cluster results

Evaluation of clustering results is sometimes referred to as **cluster validation**.

|  |  |
| --- | --- |
| **Cluster Validation** | |
| **Internal Evaluation (Internal Criterion)** | **External Evaluation (External Criterion)** |
| Based on *data clustered itself* | Based on data *not used for clustering*   * + Using known class labels / external benchmarks   + Consists of pre-classified items   + Created by humans (experts) |

Internal Evaluation

Evaluating the cluster results based on data that was clustered itself. These methods assign the best score to the algorithm that produces clusters with high intra-cluster similarity and low inter-cluster similarity.

External Evaluation

Evaluating the cluster results based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by human (experts).

What we want to achieve / do:

*Automate the evaluation of clustering algorithms’ performance through use of suitable metrics.*

Typical objective of clustering is to attain high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar).  
This is an ***internal criterion***for the clustering quality.

Review of the different internal clustering validation metrics

Each metric has their own different formulae or ways of calculations to determine the optimal clustering algorithm. By reading through several research papers and external study resources, it was found that the papers have conducted researches and their results were crucial, in that they have proven that the **S\_Dbw Validity Index** is better than the other metrics.

Thus, focus is shifted onto this metric to facilitate implementation into our project.

**S\_Dbw Index**

**Referenced from [7]:**

S\_Dbw is a validity index based on *Scattering and Density* between clusters.

Its objective: A definition of a relative algorithm-independent validity index, for assessing the quality of partitioning for each set of input values.

Main features of S\_Dbw:

* Evaluates resulting clustering schemes as defined by the algorithm under consideration
* Selects for each algorithm the optimal set of input parameters with regards to specific data set

Why is ‘S\_Dbw’ superior compared to other metrics?

**Referenced from ‘Understanding of Internal Clustering Validation Measures’ [4]:**

From the research paper, ‘Understanding of Internal Clustering Validation Measures’, a detailed study is conducted to evaluate 11 widely-used internal validation measures.

Investigation was done on the metrics’ validation properties in five different aspects: monotonicity, noise, density, sub-clusters and skewed distributions.

**\*\*The full detailed documentation is located in ‘FYPJ 2013 P3/Documentations/2013 P3/Metrics Research Documentation’\*\***

**Resources**

Operating Systems

* Microsoft Windows 7 Enterprise (Service Pack 1)
* Linux

Servers

* MySQL Server 5.6

Computer Languages Used

* Java
* R Programming

Browsers Used

* Internet Explorer
* Google Chrome

Programs Used

* Eclipse Juno
* MySQL Workbench
* Visual Paradigm
* Microsoft Office
* Cygwin Terminal
* R Studio

APIs Used

* Apache Mahout

**UI Features**

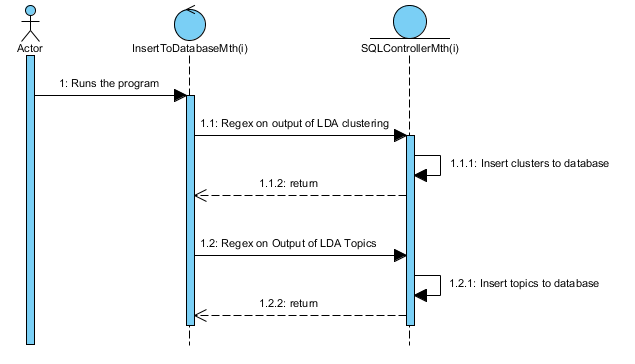
- Not Applicable -

# 3. Project Implementation

**Clustering the data (using LDA with TFIDF vectors)**

Using Apache Mahout to handle the clustering task

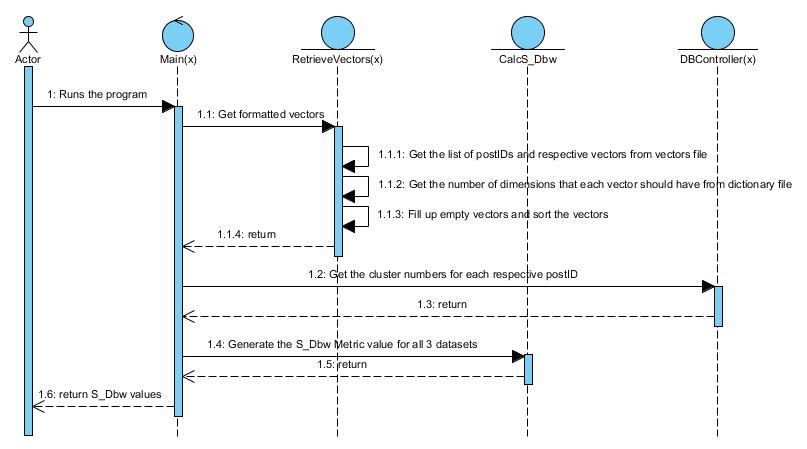
**Storing generated clusters (and topics) into the database**

 *Where the value of (i) is Mth1 / Mth6 / Mth12.*

**Retrieving Mahout’s output vectors**

Using Apache Mahout to convert the SequentialFile to human-readable text file   
(seqdumper command)

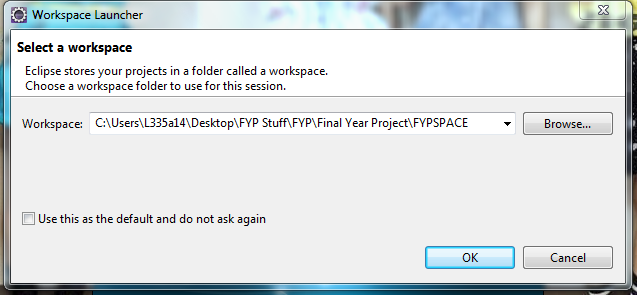
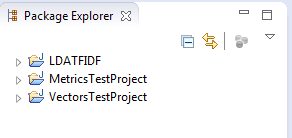
**Generating the S\_Dbw Metric results (after formatting Mahout’s output)**



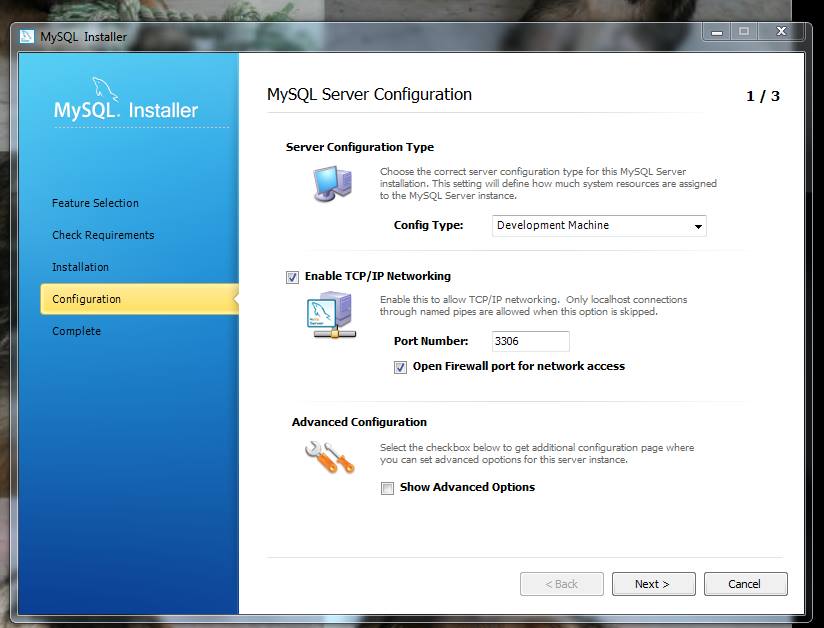
*Where the value of (x) is KMeans / LDA*

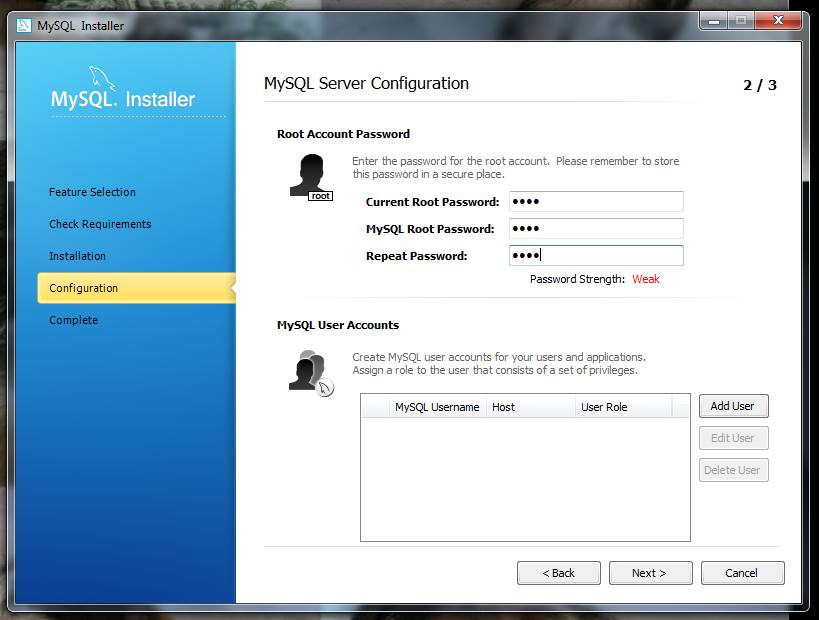
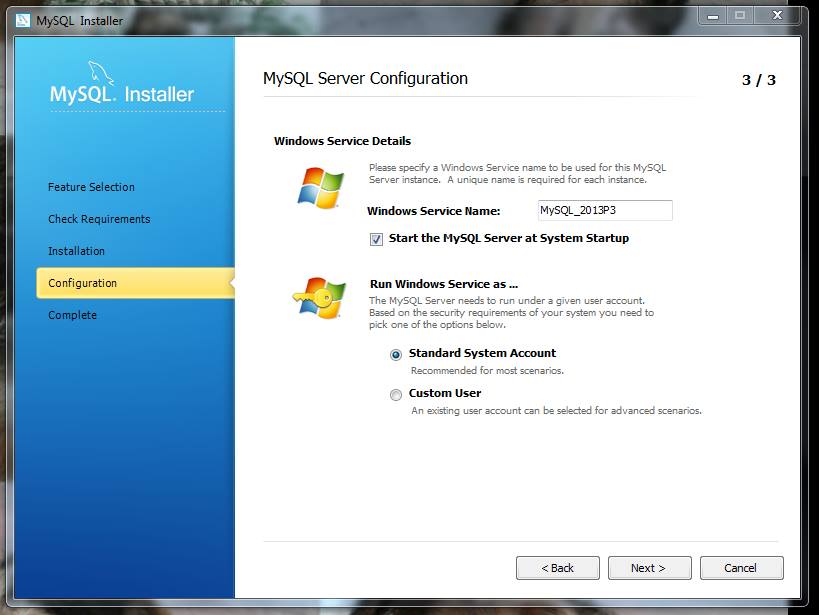
# 4. Sources and Manuals

**Setting up Eclipse Juno**

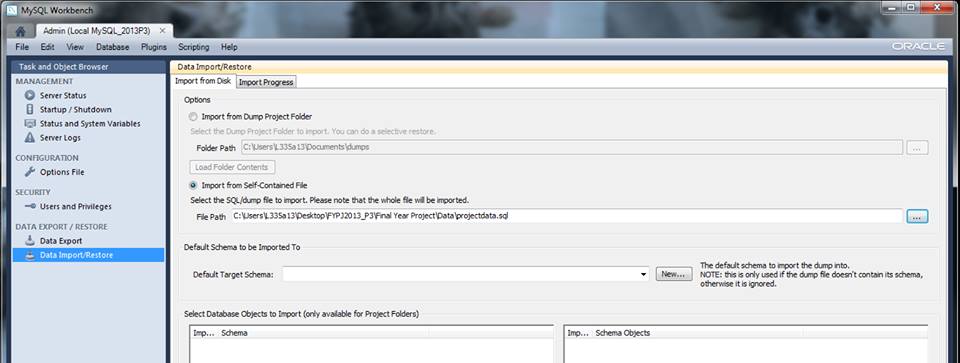
1. Run eclipse.exe from the folder ‘FYPJ 2013 P3\Installation Files\eclipse-jee-juno-SR1-win32-x86\_64\eclipse’  
     
   
2. Press the “Browse” button and browse to the ‘FYPJ 2013 P3\FYPJ’ folder. Thereafter, Press “OK”.
3. You will be able to see the Project Explorer:   
   
4. Select and re-import all the necessary JAR files that have error (due to different directory issues) (The JAR files should be in the folder of the respective project folder)

**Setting up Database**

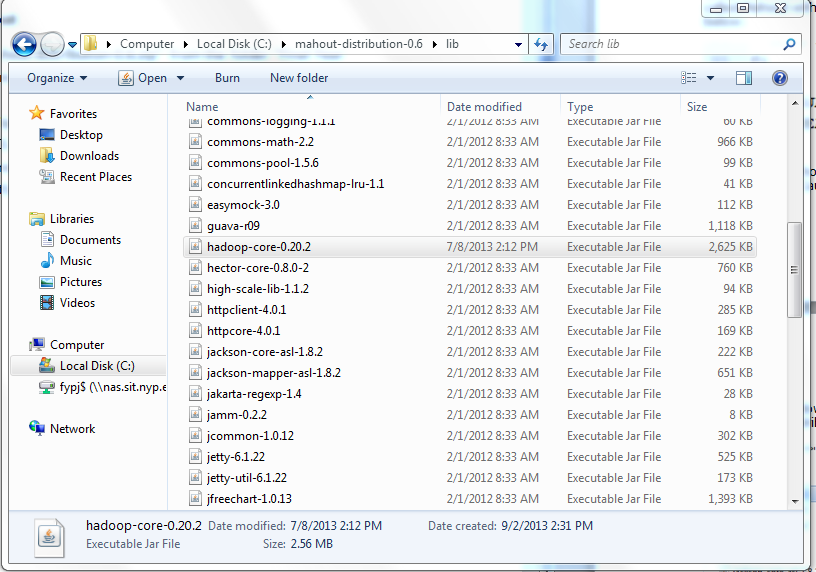
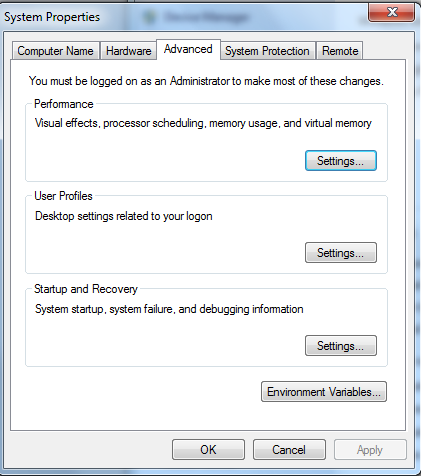
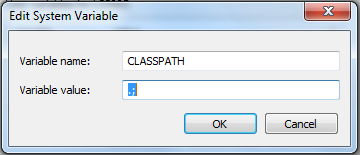
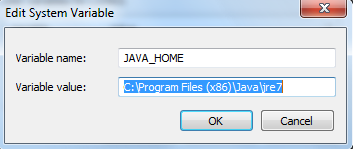
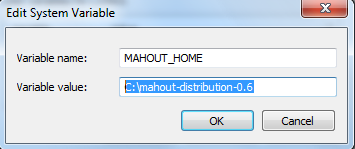
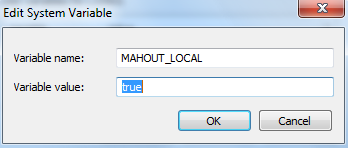
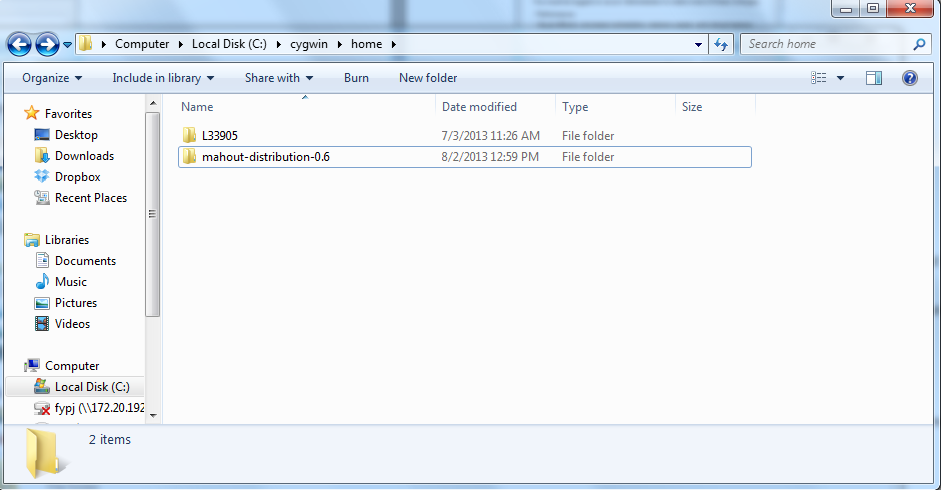
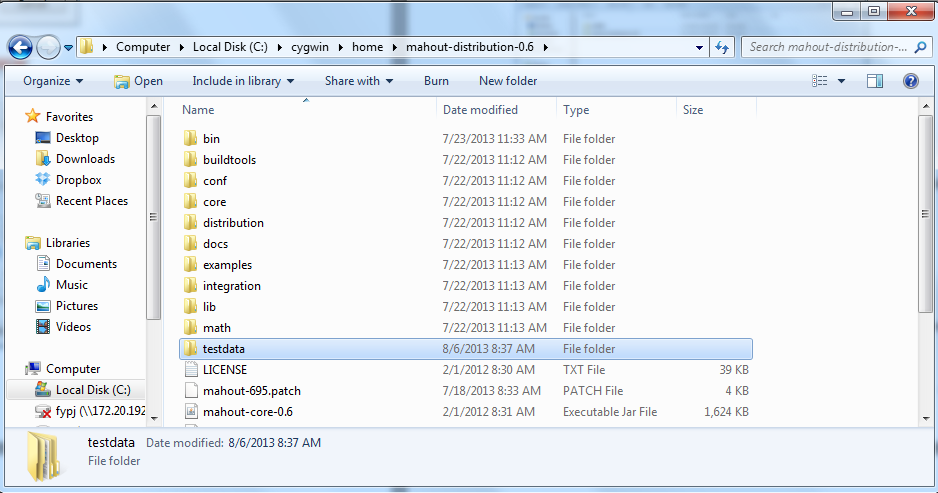
1. Run and install “mysql-installer-community-x.x.xx.x.exe” from the folder “FYPJ 2013 P3\Installation Files” (where x’s are the version numbers)
2. Set-up the installation as follows: (with username: “root”, password: “root”).   
     
   

1. Launch “MySQL Workbench” and click on “Manage Import / Export”.   
   Select the “Import from self-contained file” option, browse for SQL schema file from the folder ‘FYPJ 2013 P3\Database’.  
   Press the “Start Import” button. (May take up to 5 – 10 minutes)

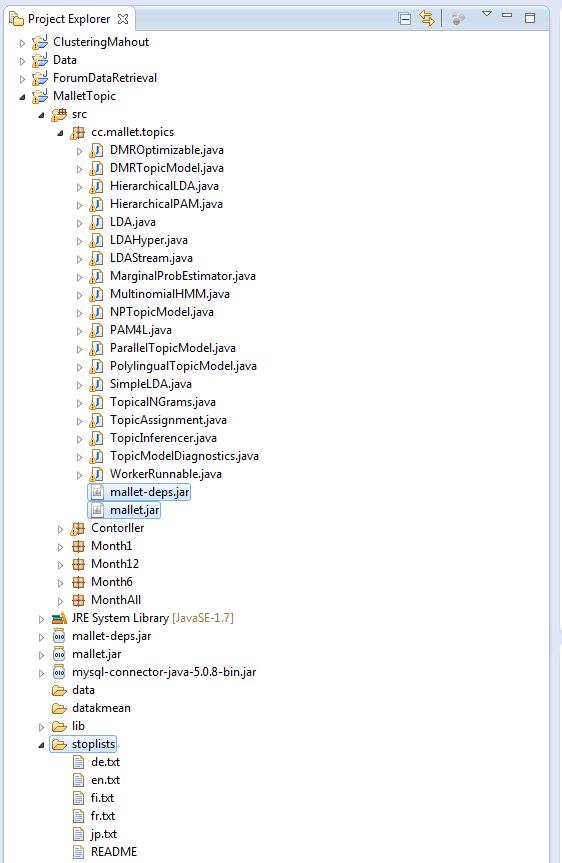


**Setting up Apache Mahout**

1. Unzip the file “mahout-distribution-0.6.zip” from the folder ‘FYPJ 2013 P3\Installation Files’.
2. Browse to the unzipped folder and enter the “lib” folder. Locate the jar file named “hadoop-core-0.20.204.0.jar” and DELETE IT (It will cause problems when trying to run Mahout in Windows, thus have to downgrade its version). Browse to the folder ‘FYPJ 2013 P3\Installation Files’ and replace the DELETED file with “hadoop-core-0.20.2.jar”.   
   
3. Go to Computer > Properties and select “Advanced System Settings” from the left panel.   
   
4. In the “System Properties” window, select the “Environment Variables…” button.   
   
5. Follow the screenshots and add the following Environment variables: (Select the “New” button at the System Variables section)  
     
     
     
     
     
     
     
     
   Once complete, press “OK”, and “OK” again for the windows that appear.
6. Mahout requires Cygwin to be used. Please Refer to ‘FYPJ 2013 P3\Documentations\2013 P2\Mahout Setup in Windows’ to install Cygwin.   
   Once your done copy the “mahout- distribution-0.6” folder to “Cygwin” folder’s home folder.  
   
7. Create a new folder called ‘testdata’ in ‘C:\cygwin\home\mahout-distribution-0.6’  
     
   

**Setting up Apache Mahout**

1. Go back to Eclipse and browse to the “MalletTopic” project folder.   
   Import the following jar files located in the project folder. Also, check that the stoplists folder is located in the project folder.



# 5. Supervisor Sharing

# 6. Project Team Sharing

# 112723X\_Lim Boon Leng

Machine learning; documents clustering is a very fresh topic, having heard it for the first time. During the weeks of research, many interesting things were found; how actual clustering works, the wonders of clustering, and the complex algorithms involved. All these required a lot of time digesting in order to fully understand how they work, and why such algorithms really work during clustering.

As an Information Security student, working on clustering was really weird, as no form of security; things I have learnt for the past few semesters can somehow be applied here. Working with my partner, who is from another course, was a good experience, getting to know how do people from other courses work.

From this project, I’ve learnt and have a slight grasp on the topic of clustering, although there were slight difficulties in understanding and programming the formulae. Time management was also one of the problems.

All in all, having such an experience was really cool and enriching, I feel that FYPJ will really help a lot when working in the industries, with similar happenings, such as working with people you never knew before, something new that you never heard of.

**116853D\_Hong Kay San**

The project is based on clustering topics and this topic is a very dry and broad topic. So I’m having a hard time to get to know about the clustering algorithms and how it works.

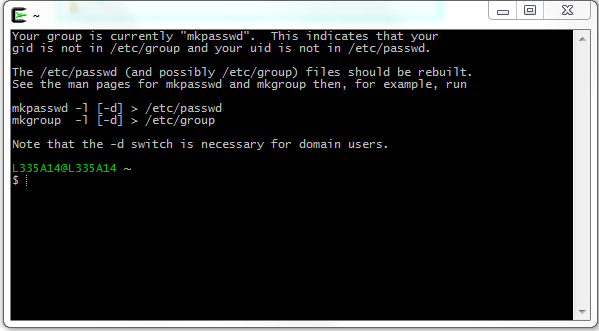
After researching, I’m still unable to understand some of the clustering algorithms and how they work. So I read the research over and over again, reading on notes that were done by my partners and finally I get to understand how they work.

During the research on the S\_DBW R Studio, there were some difficulties that we faced it and it was not solved so in the end we have to give up on the R Studio and expand the excel spreadsheet.

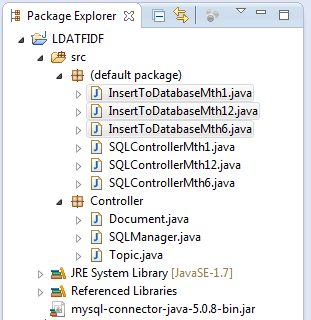
From this project, I’ve learnt a lot that I will never forget. It's a great experience for me.

# 7. How it works

# Generating clusters, vectors (LDA TFIDF)

1. Go to the start menu and launch Cygwin Terminal as administrator.   
     
   
2. Run the commands from the text file located in the folder ‘FYPJ 2013 P3\Vectors, Dictionaries’.   
   (Required to edit the command as necessary, according to where your files are located, for each dataset) (An explanation of the commands is available at ‘FYPJ 2013 P3\Documentations\2013 P3\Results Analysis Documentation’.
3. After finish generating the clusters, dump the vectors and dictionary file generated to a text file using Mahout’s seqdumper command. (A copy of all the vectors and dictionary files is available at ‘FYPJ 2013 P3\Vectors, Dictionaries’.

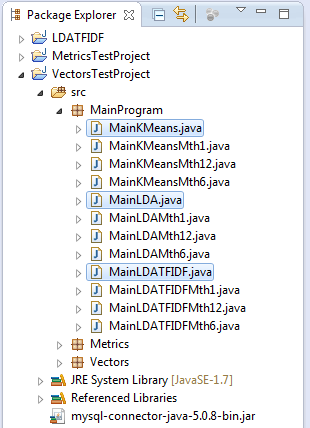
# Inserting the cluster results into the database



Make sure to edit the program such that the file paths point to where your files are located.   
Also remember to change the ‘SQLManager’ and ‘SQLControllers’ if you wish to change database or table names.

Thereafter, run the 3 highlighted programs to insert them into the database.

# Running the main program



The above 3 highlighted programs were split into 9 (3 for each dataset), in order to allow Eclipse to run the programs in parallel, to save time. (Runtime of Mth12 program takes up to a week)

Make sure to edit the program such that the file paths point to where your files are located, for each program.

The above program will do all the processing at once; Formatting Mahout’s output vectors, retrieving the respective clusters, and lastly, calculating the S\_Dbw Index value.

# 8. Results Analysis

**Objective**

To identify the more efficient, suitable clustering algorithm for clustering documents/forum posts in order to create a clustering document engine, to uncover hidden and or trending topics within a dataset/forum.

**Introduction**

Latent Dirichlet Allocation (LDA) is a possible clustering algorithm that can meet the objective. LDA started out being a Topic Modeling algorithm and has improved to become a clustering algorithm. Now LDA is considered to be one of or the best clustering algorithms around; a simple Google search will tell you that LDA is useful and accurate when it comes to clustering documents/posts. Hence there is a reason to use LDA and evaluate its results to verify its suitability.

In order to do so, we will be using other an established clustering algorithm, KMeans, to compare their results and determine if LDA is a suitable algorithm.

**Input Dataset**

There are 3 datasets that are used in this comparison, which are forum posts from the sub-forum ‘Local & Foreign Affairs’ from the forum ‘SGClub’.

Raw Dataset 1 (1 Month)

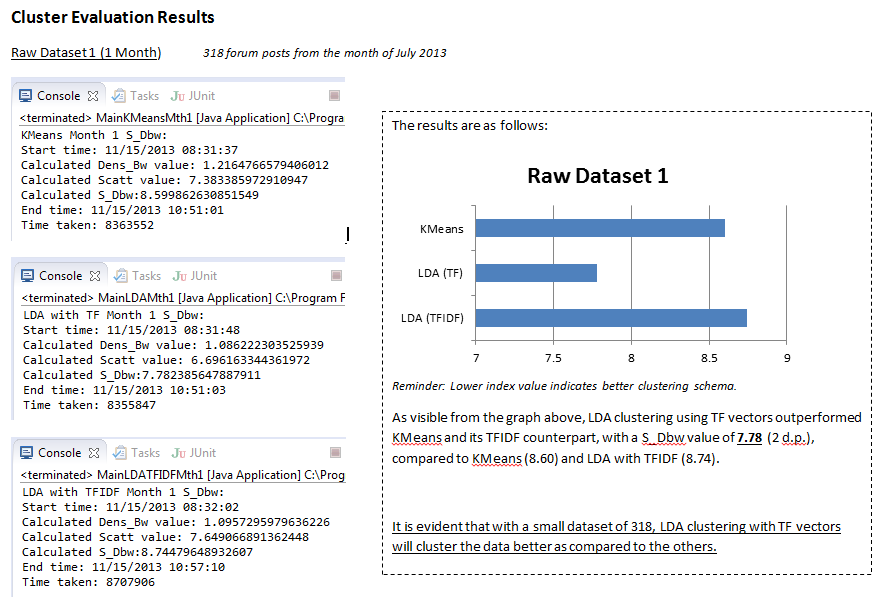
318 forum posts from the month of July 2013

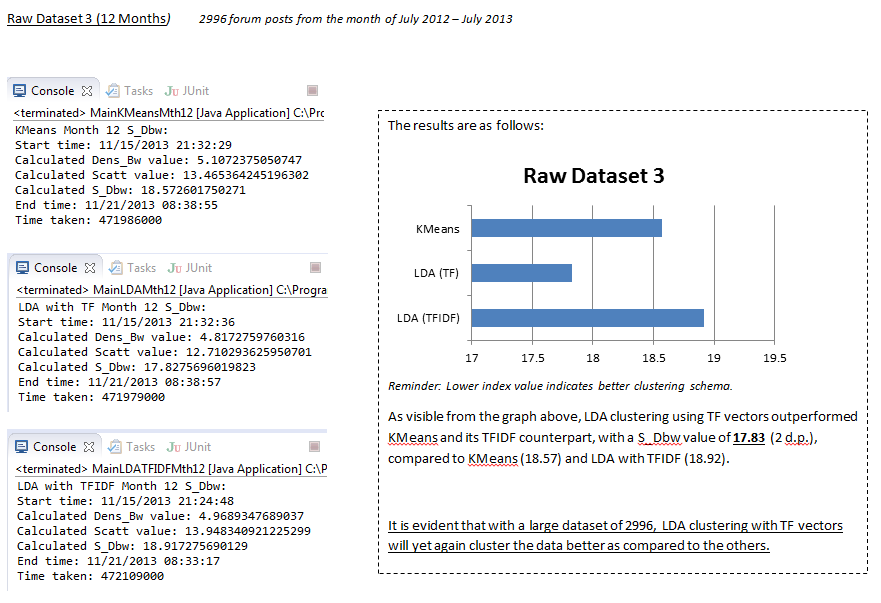
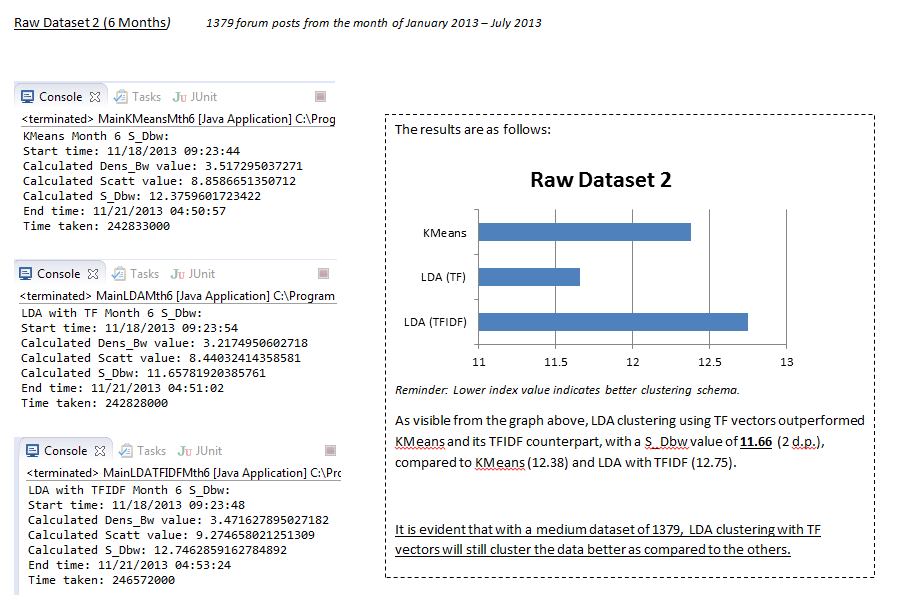
Raw Dataset 2 (6 Months)

1379 forum posts from the month of January 2013 – July 2013

Raw Dataset 3 (12 Months)

2996 forum posts from the month of July 2012 – July 2013





**Conclusion**

As visible from the above results of clustering from all three datasets, it is safe to assume that LDA using TF vectors is the best clustering algorithm, for any type of dataset be it small or large.

***LDA with TF vectors outperformed both its counterpart of TFIDF vectors and KMeans by achieving the lowest S\_Dbw index value for all three datasets.***

**This concludes that LDA is a suitable algorithm for documents clustering, as compared to KMeans.**

From the results, it also concludes the fact that LDA using TFIDF vectors will underperform, just as pointed out by the previous’ group research that using LDA with TF will be *more efficient*. (This was their reason for using LDA with TF initially).

Both the Scatter and Dens\_Bw values of LDA clustering with TF vectors are **always the lowest** compared to the others, thus resulting in a lower value of S\_Dbw.

Something interesting that can be concluded from the results is also that the Scatter value for LDA with TF vectors will always be the **lowest**, which is also the main reason of it having a lower S\_Dbw index value, as the Dens\_Bw value for all three clustering results are **around the same**, not having much of a big difference.

**This concludes that LDA is a suitable algorithm for documents clustering, as compared to KMeans. Using TF vectors during clustering will also achieve better results rather than making use of TFIDF vectors for the LDA algorithm.**

**\*\*The full detailed documentation is located in ‘FYPJ 2013 P3/Documentations/2013 P3/Results Analysis Documentation’\*\***

# 9. Discoveries

* During the processing of raw data to sequence file to vectors, there was a phenomenon whereby there were posts without any vectors (null values for all dimensions)
  + Omitted / ignored temporarily
  + Discovery: Were forum posts with minimal text , words of no ‘importance’
* Processing of raw data to sequence file to vectors is deterministic, but   
  for the same set of vectors, LDA will produce different cluster results
  + LDA is not deterministic.
  + With reference to WWW, KMeans is also not deterministic.