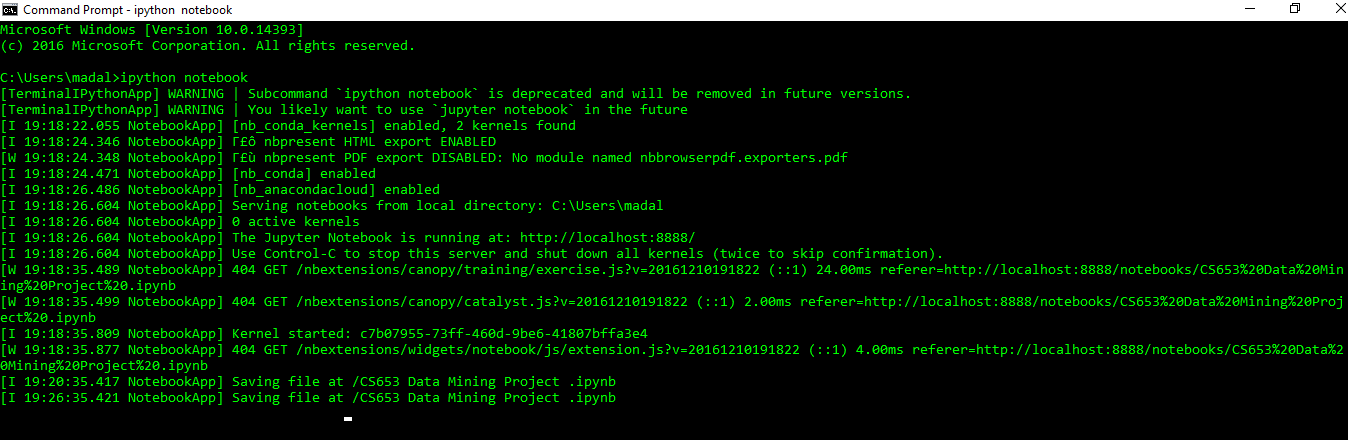
CS-653 Data mining Project

Instructions and manual how this works:

We used the ipython note book for sentiment analysis using naïve bayes.



Next1: we have uploaded our data set into the [http://localhost:8888/tree with extension of.txt](http://localhost:8888/tree%20with%20extension%20of.txt).

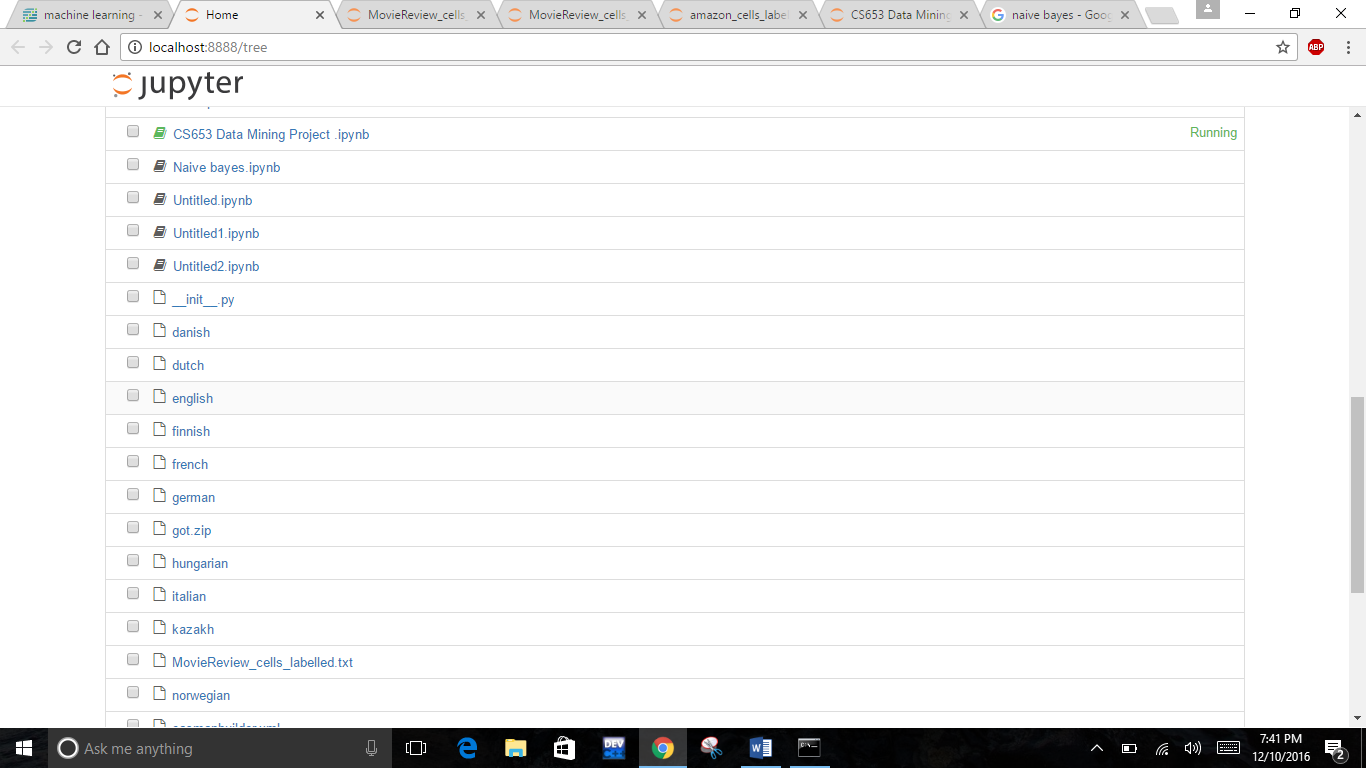
[MovieReview\_cells\_labelled.txt](http://localhost:8888/edit/MovieReview_cells_labelled.txt)

Next2: We have uploaded the packages such as NLTK 3.0 documentation: The Natural Language Toolkit (NLTK) is an open source Python library for Natural Language Processing. The packages named as follows

from nltk.corpus import stopwords http://www.nltk.org/api/nltk.html

Next3: the number of observations are shown as follows 

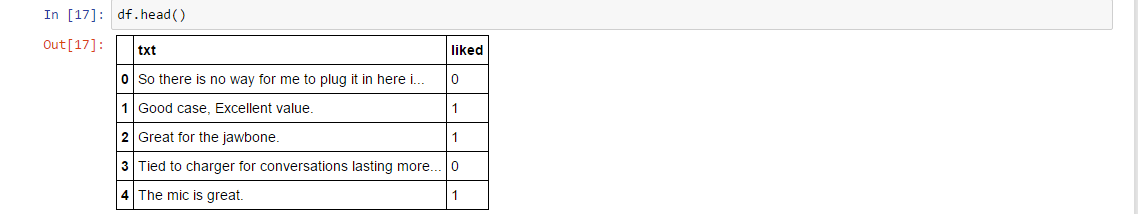
This shows the home page that give the view of the data set name and the other imported files in this following one



We read the data set as csv file separated by tab spaced

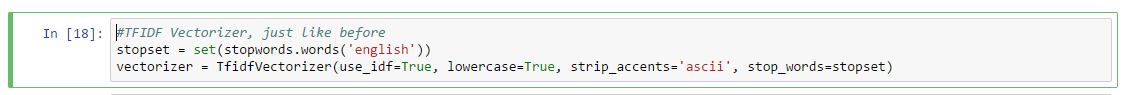
C:\Users\madal\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Capture2.png

The data set is classified as follows



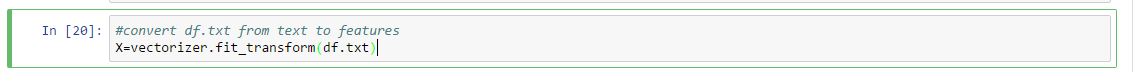
We vectorize the text by using TFID

from sklearn.feature\_extraction.text import TfidfVectorizer



The featuring the text is shown in this

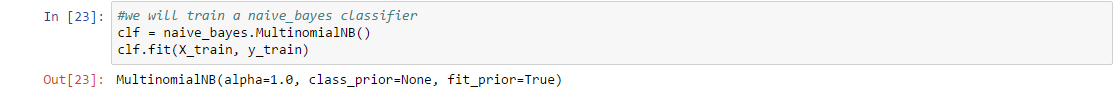
#convert df.txt from text to features



Next4:As we mentioned in our project proposal we have used the #test train split set as usual

C:\Users\madal\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Capture6.png

Next5: we have trained it using naïve\_bayes classifier

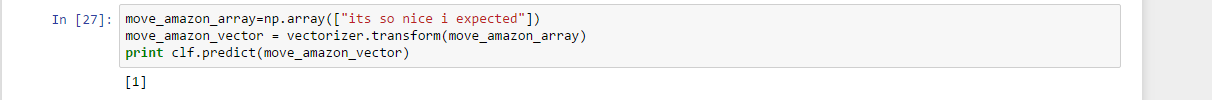


Next6: We have also calculated the accuracy of the Sentiment analysis using naïve bayes using the probability function

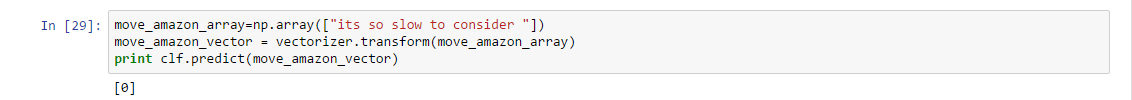


We evaluate the sentence as follows

1. Let us suppose we give an positive review it shows 1.
2. If at all we give the negative review it shows 0.

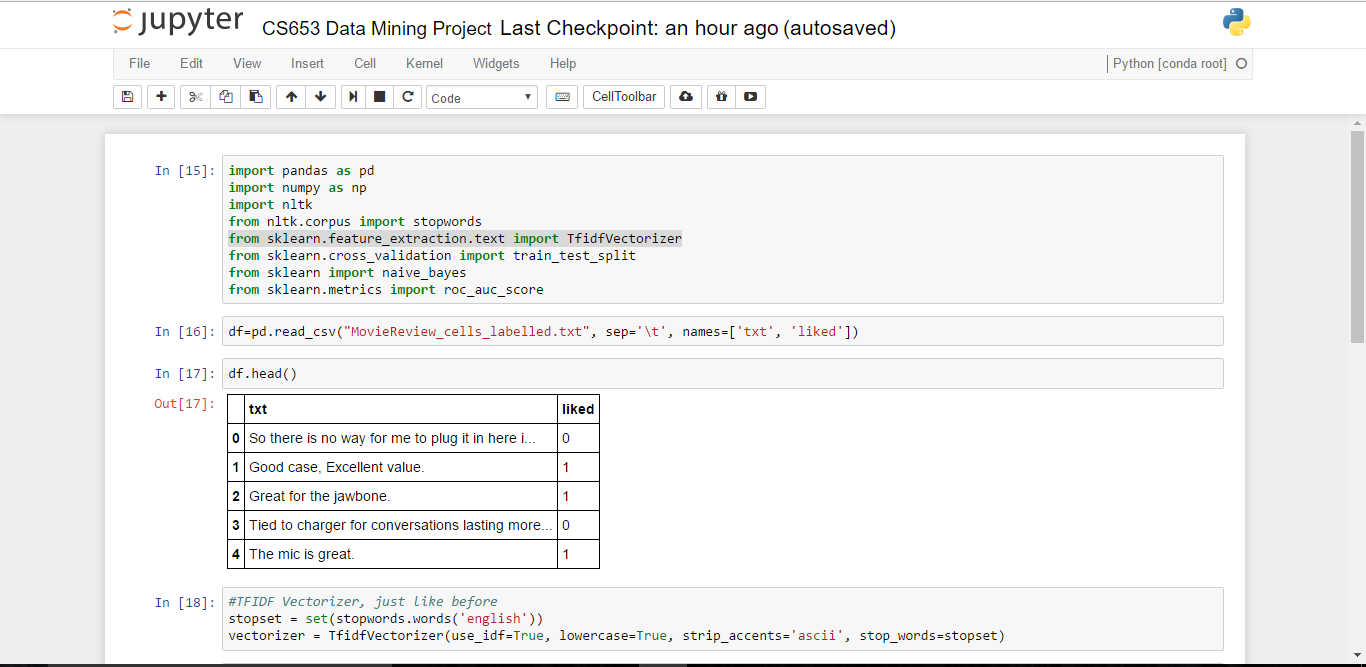


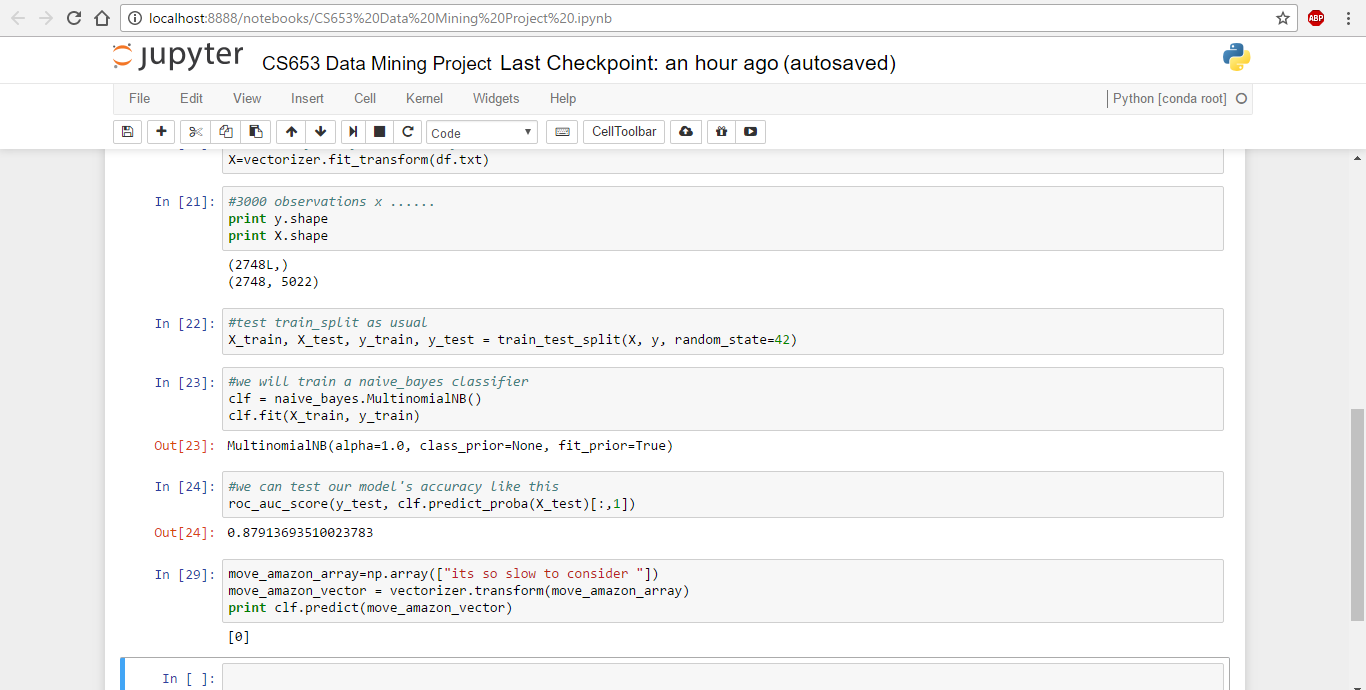
As we give the positive review the naïve bayes analysis give us 1 conforming that its one

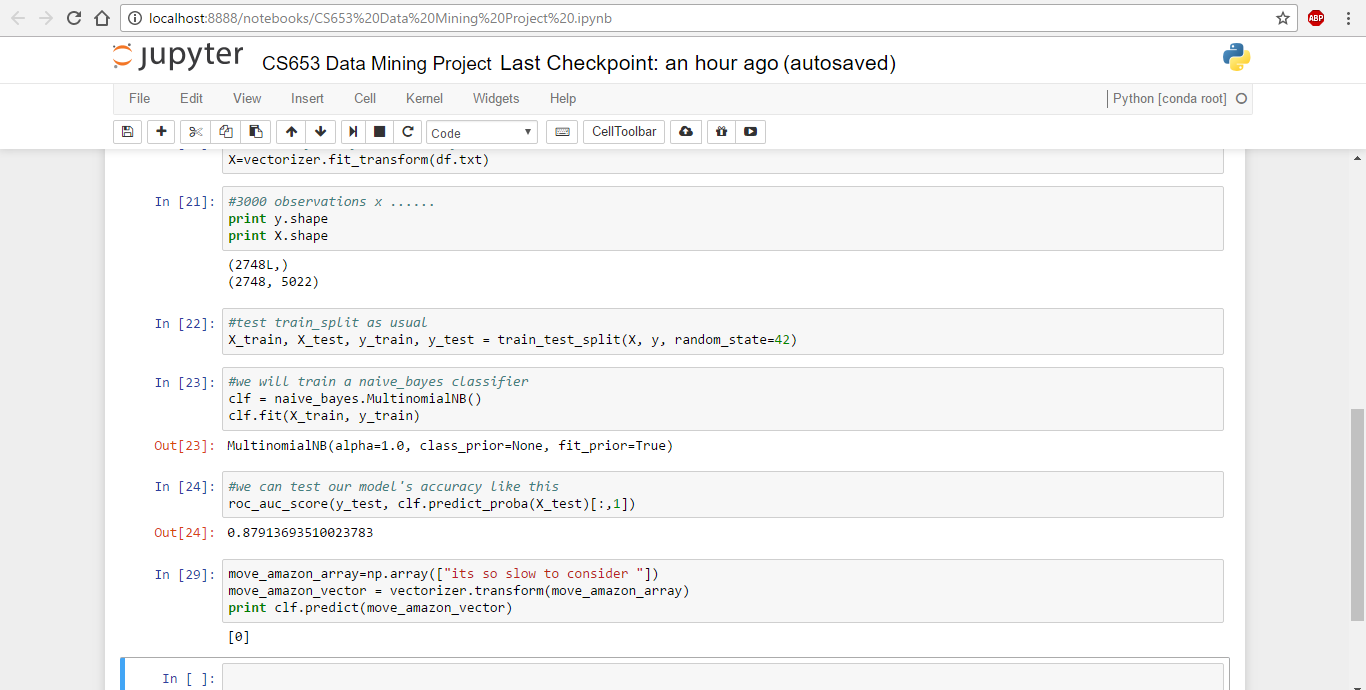


As we give the negative review the naïve bayes analysis gives us 0 conforming that its zero

Here are the screen shots of the source code







**Challenges we came across:**

While collecting the data we first used the twitter API to get the movie reviews so as we extracted the data from that we found most of them are irrelevant tweets which we does n’t give any information from which does n’t say anything about the movie review.

After that we have used the data set for IMDB movie review and we used help form the tensorflow

https://www.tensorflow.org/

# API Documentation

TensorFlow has APIs available in several languages both for constructing and executing a TensorFlow graph. The Python API is at present the most complete and the easiest to use, but the C++ API may offer some performance advantages in graph execution, and supports deployment to small devices such as Android.

Additionally, the TensorFlow maintainers intend to include APIs for [Java](https://github.com/tensorflow/tensorflow/issues/5). We hope that the TensorFlow community will develop front ends for other languages like JavaScript, Lua, R and perhaps others, building on the [approach recommended by the TensorFlow maintainers](https://www.tensorflow.org/versions/r0.12/how_tos/language_bindings/index.html).

We got the data set of 3000 reviews from IMDB movie review and we had done the featuring and labeling of the data and done the classification under Naïve bayes analysis : using training and testing model.

**Data Extraction**

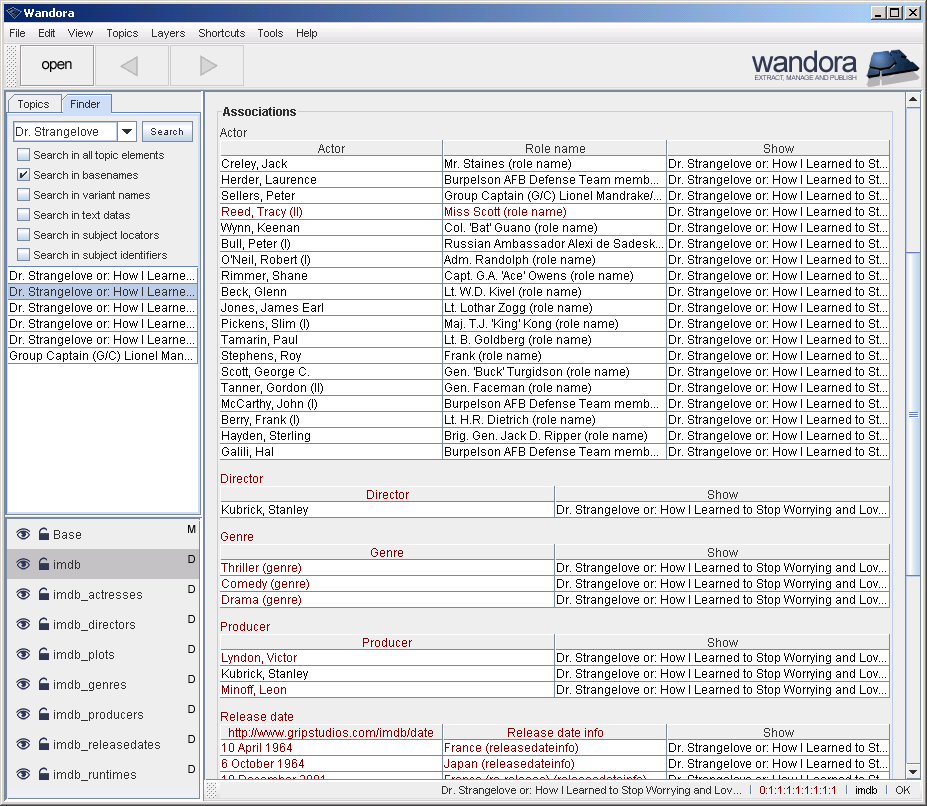
IMDB extractor transforms [Internet Movie Database](http://www.imdb.com/) data files into a topic map browsable with Wandora. Extractor has been created for demonstration purposes only. Wandora does not contain any IMDB data files. Also, be aware that Wandora or Wandora authors have no rights to give you any permission to use IMDB data. If you plan to use IMDB topic maps beyond personal usage, you should contact [IMDB Licensing department](http://www.imdb.com/licensing/).

As datafiles are extremely large you can't extract data to [memory topic maps](http://www.wandora.org/wiki/Memory_topic_map) but have to use [database topic maps](http://www.wandora.org/wiki/Database_topic_map). Wandora does not transfer all IMDB files. Current extractor transfers only

* actors
* actresses
* keywords
* countries
* language
* locations
* genres
* movies
* biographies
* producers
* directors
* plot summaries
* running times
* release dates
* **reviews**

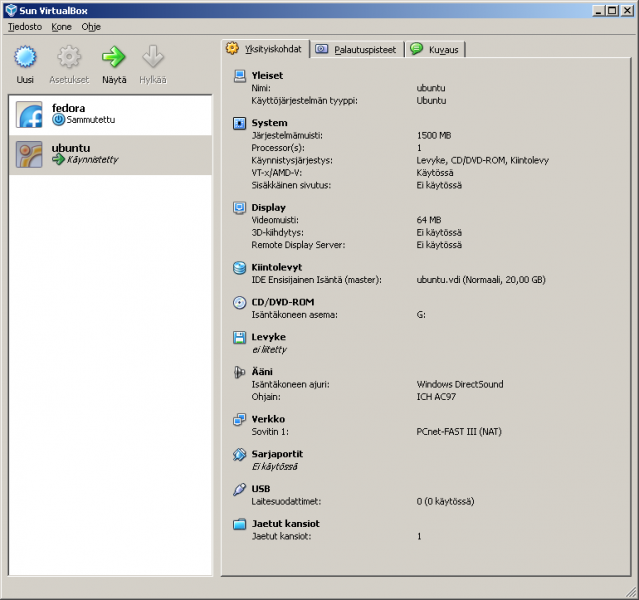
To prepare the extraction download all required data files and unpack them to your local file system. Then create a database topic map and start extractor with **File > Extract > Media > IMDB Extractor**. Wandora requests a folder containing IMDB data files or a single data file and starts the extraction after successful data file or folder identification. IMDB data files are very large and you should be patient as the extraction may take a while.

Below is a screenshot of Wandora viewing associations of movie **Dr. Strangelove...**. Notice the layer structure. Each IMDB datafile has been extracted to a separate database topic map.



**Step by step example of extracting IMDB with Wandora**

Next screen shot views system properties of the Ubuntu Linux used for IMDB extractions. Notice the memory amount given for the Linux. We gave the Ubuntu 1500 MB of memory. Our experiences suggest you should give Linux memory as much as possible. With small memory footprints the IMDB extraction fails after heavy swapping. Now start Ubuntu Linux and log in.

[](http://www.wandora.org/w/images/Imdb_09.png)

**Setting up databases for IMDB topic maps**

As stated in the beginning of IMDB extractor documentation above, you need a database topic map to store extracted topic map as it is very large. To prepare database topic map start another terminal window in Ubuntu with option **Applications > Accessories > Terminal**.

In terminal

Install MySQL server with command **sudo apt-get install mysql-server**.

Log into the MySQL server with command **mysql --user=<your-username> --password=<your-password>**

Create empty databases with MySQL command **create database <database-name>;** (notice ending semicolon) for next database names:

imdb\_reviews

Prepare each created database with Wandora specific database table structures in **wandora/build/resources/conf/database/db\_mysql.sql**. In detail:

Select database with MySQL command **use <database-name>;**, for example **use imdb\_reviews;** (notice ending semicolon).

Read database table creation clauses from external file with MySQL command **source wandora/build/resources/conf/database/db\_mysql.sql;** (notice ending semicolon). Notice that you may have to change the path of **db\_mysql.sql** depending on you Wandora installation directory and your current directory.

**Extracting IMDB with Wandora**

Go back to the Wandora application started earlier and select menu option **Layers > New layer**. A dialog window opens. Select **Database.**

Select **MySQL test** in database settings list and click **Edit** button. Another dialog opens for database settings (see image below). In this dialog you can enter database's name, user name, and password. Change database name to **imdb\_reviews**

Now click **OK** button and database configuration window closes reveling previous dialog window. Enter name for the layer, say **imdb\_reviews**, keep the MySQL test database configuration selected, and click **OK** button. Wandora creates a new topic map layer and shows it left bottom corner of Wandora application window (see below). Now select the created layer by clicking it. Selected layer is little darker than unselected. Now all "write" operations go to the selected database topic map layer.

Next we are going to start the IMDB extraction. Select menu option **File > Extract > Media > IMDB extract...**. Wandora opens a Files/Urls/Raw selector. Keep the Files tab open and click **Browse** button. A file selector opens. Go to the directory you uncompressed IMDB data files and select **reviews.list**. To start extraction press **Extract** button. As IMDB data files are extremely large, it is not very surprising the extraction takes several hours.