**OUR APPROACH:**

**TASK1)** Generate a training set using business attributes and review data and using classification algorithms from weka we can predict if we start a new business with with given attributes what can be the rating for the new business.

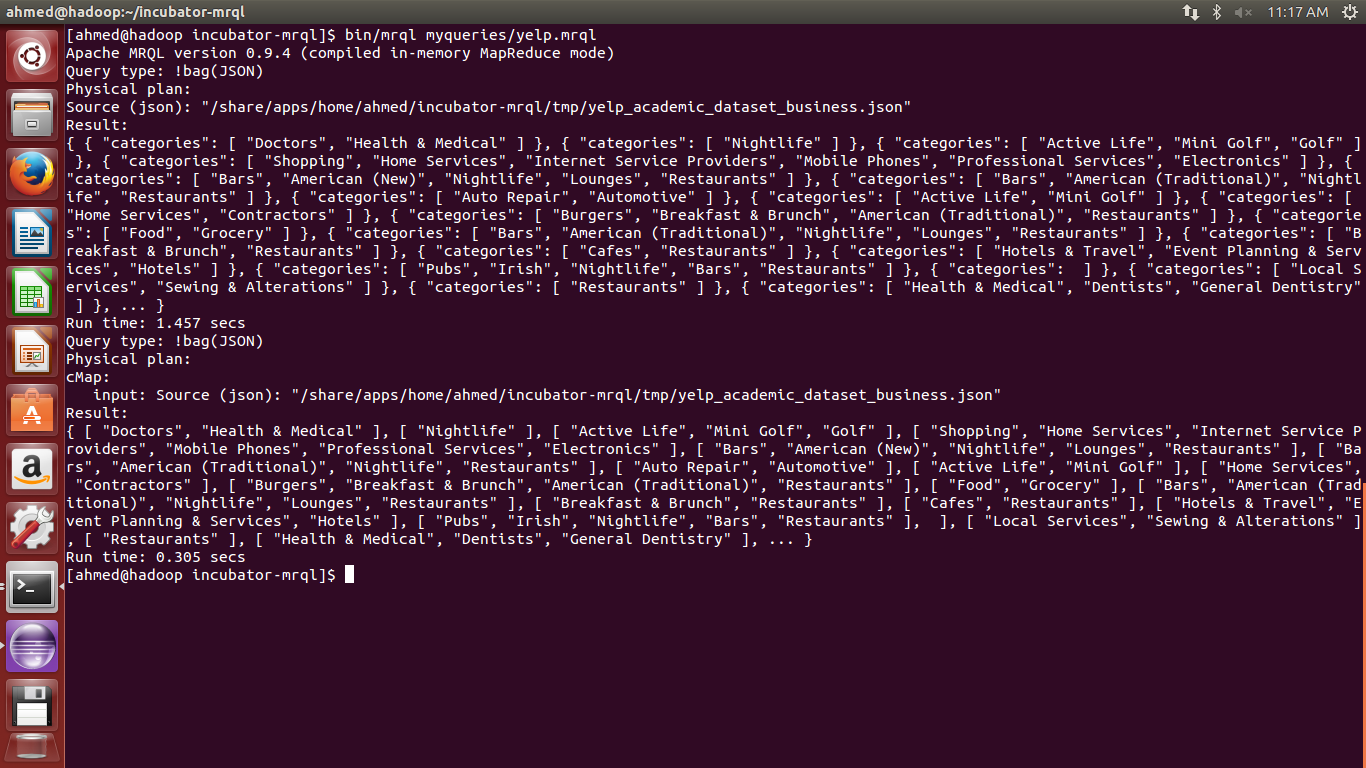
**TASK2)** Each file consists of either userid or business id using this key idea we are trying to come up with a graph and then use HITS algorithm. This will be done using MRQL.

**CURRENT STATUS:**

We have decided to use map reduce approach to solve this challenge. The dataset given is in json format. As an initial step we had decided to try various tools to see how we can access data to create a graph as well as training data for applying data mining algorithm using map reduce approach. We were successfully able to write MRQL query for querying dataset and moving ahead with further implementation.

TASKS ACHIEVED:

* Installation of mrql.
* Understanding of tools like mrql, weka and pandas.
* Successfully executed one mrql query.
* Next step is generation of graph from given dataset
* Failed to process json using pandas but still working on it.



CHALLENGES FACING CURRENTLY:

While working with dataset we realized that dataset in csv has a lot of columns and using json was a better and optimized way. Also we tried to write a python script to extract data for creating our own training set for using classifiers from weka but json parser from python, pandas library seem to throw error while parsing json file.

JSON file seems to have mixed datatypes and so it breaks the json parser implementation in various libraries and tools.

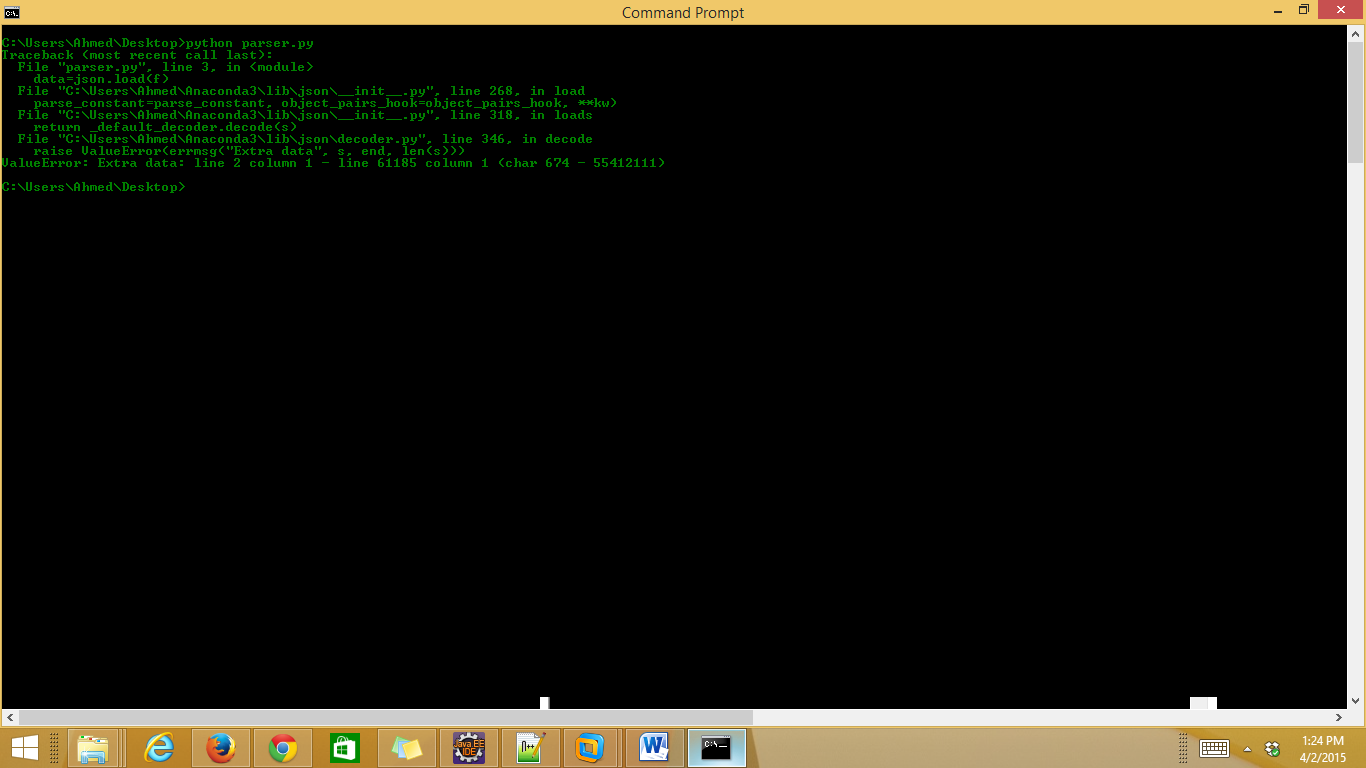
Below is a screen shot of the challenges faced after running the code below.

import csv

import pandas

csv\_reader=pandas.read\_json('yelp\_academic\_dataset\_business.json')

print(csv\_reader)



**Change of plan since the project proposal**

Since the proposal things have become more clear and specific and we are moving towards clarifying our ideas as we progress. Tasks mentioned in proposal have reduced to two specific and focused tasks. Since many options were open initially, now we are working hard to find a best option by the of project submission.

**Remaining work**

Since we have decided to reduce this data by generating our training set from this humungous dataset and then use weka and try different algorithms and evaluate different possibilities. More time will be spent on creating our training set due to the challenge mentioned above. Once it is done classification can be achieved in a short time.

Also we are using a very optimistic and difficult approach for coming up with a graph. This part has lot of work to be done. In this part we will have to various challenge like define a good graph and then creating it from dataset. Also we need to implements the HITS algorithm using MRQL which itself will take a lot of time.