1. I implemented a Latent Aspect Rating Analysis from Hongning Wang, Yue Lu, Chengxiang Zhai in python. The code is provided in a zip file. The dataset used is Yelp Restaurant Review corpus. Latent Aspects are calculated and new aspects from the reviews is generated. The model was fine tuned empirically to get more than 25 aspects from every review. Every aspect was provided a rating.

2. The training phase on more than 4000 reviews took time. I would wish to use a better GPU for this purpose.

3. The dataset is freely available from https://www.yelp.com/dataset

3. CODE:

Dependencies
It requires the nltk dataset and additional packages for running. Use nltk.download() and download all the given packages. Also required is the vader package from nltk.
It requires python3.

Running Code
The code can be run manually by using main.py

The newly acquired aspects are stored in the "output" folder."**final_aspect_words.txt**
" text file holds the newly acquired aspects mined. Corresponding ratings are allocated to every aspect mined and stored in review_data.txt.

4. Theory and Observations:

In this paper, authors identified and analysed a new problem of opinionated text data analysis called Latent Aspect Rating Analysis (LARA), which aims to analyse opinions expressed about an object in an online review at the level of topical aspects to discover the latent opinion of each individual reviewer on each aspect as well as the relative focus on various aspects when forming the overall judgement To solve this new text mining issue in a general manner, they proposed a novel probabilistic rating regression model. Empirical studies on a data set for a hotel review show that the proposed latent rating regression model can effectively solve the LARA problem and that a thorough analysis of opinions at the level of topical aspects allowed by the proposed model can help a broad range of application tasks, such as overview of aspect opinion, ranking of individuals based on aspect ratings, and reviewer analysis.

In the code I went through the following steps to replicate the paper:

I. Create Vocabulary from the dataset.

II. I used porter stemmer on the dataset to create a vocabulary (stemmed) corpus of the reviews. This stemmed corpus will then be used for aspect mining and rating.

III. Initially defined aspects (10 aspects) are read from the file init_aspect_word.txt"

IV. The aspects are then mined using LARAM bootstraping , which uses the following two classes for Restaurant reviews:

- class Review:
- class Restaurant

V. The the w matrix is calculated using ci-square metrics to calculate the ratings per mined aspect.

VI. The results are the saved in output folder.

NOTE: The original paper used regression to calculate weight for each aspect. But I also used sentiment analysis module from nltk to rate individual aspects. The Regression.py does not work, I just left it there for reference. Use mainpy to run the actual code.