Pig Latin and Hadoop File System (HDFS) to derive some statistics from **Yelp Dataset.**
The dataset files are as follows and columns are separate using '**::**'
business.csv.
review.csv.
user.csv.

Dataset Description.
The dataset comprises of **three** csv files, namely user.csv, business.csv and review.csv.

Business.csv file contain basic information about local businesses.
Business.csv file contains the following columns "business_id"::"full_address"::"categories"

'business_id': (a unique identifier for the business)
'full_address': (localized address),
'categories': [(localized category names)]

review.csv file contains the star rating given by a user to a business. Use user_id to associate this review with others by the same user. Use business_id to associate this review with others of the same business.

review.csv file contains the following columns "review_id"::"user_id"::"business_id"::"stars"
'review_id': (a unique identifier for the review)
'user_id': (the identifier of the reviewed business),
'business_id': (the identifier of the authoring user),
'stars': (star rating, integer 1-5),the rating given by the user to a business

user.csv file contains aggregate information about a single user across all of Yelp
user.csv file contains the following columns "user_id"::"name"::"url"
user_id': (unique user identifier),
'name': (first name, last initial, like 'Matt J.'), this column has been made anonymous to preserve privacy
'url': url of the user on yelp

Write efficient Pig Latin program in to find the following information. Load the files in HDFS and read it in your Pig Latin program.

NB: :: is Column separator in the files.

Q1.

List the business id , full address and categories of the **Top 10 businesses** located in **Palo Alto, CA** using the average ratings. This will require you to use review.csv and business.csv files.

Please answer the question by **calculating the average rating**s given to each business using the review.csv file. Do not use the already calculated ratings (average stars) contained in the business entity rows.

Q2

List the business id , full address and categories of the **Top 10 businesses** located in **CA** but not in **Palo Alto**, **CA** using the average ratings. This will require you to use review.csv and business.csv files.

Please answer the question by **calculating the average rating**s given to each business using the review.csv file. Do not use the already calculated ratings (average stars) contained in the business entity rows.

Q3:

Using Pig Latin script, Implement co-group command on business_id for the datasets review and business. Print first 5 rows.

Q4:

List the 'user id' and 'rating' of users that reviewed businesses located in Stanford

Required files are 'business' and 'review'. Print first 10 rows.