pig_final_assignme...

```
%sh
ls /shared/d/yg/Downloads/dataset/
```

READY

```
business.json
checkin.json
elephant-bird-core-4.15.jar
elephant-bird-hadoop-compat-4.15.jar
elephant-bird-pig-4.15.jar
json-simple-1.1.1.jar
photos.json
piggybank.jar
plant_locations.csv
plants-locations_reactors.csv
review.json
tip.json
Untitled.ipynb
user.json
```

```
%pig
REGISTER '/shared/d/yg/Downloads/dataset/elephant-bird-core-4.15.jar'
REGISTER '/shared/d/yg/Downloads/dataset/elephant-bird-hadoop-compat-4.15.jar'
REGISTER '/shared/d/yg/Downloads/dataset/elephant-bird-pig-4.15.jar'
REGISTER '/shared/d/yg/Downloads/dataset/json-simple-1.1.1.jar'
REGISTER '/shared/d/yg/Downloads/dataset/piggybank.jar'
```

READY

```
%pig
a = LOAD '/shared/d/yg/Downloads/dataset/business.json' USING com.twitter.elephantbird
```

READY

```
--1. The average # of reviews and the average # of stars grouped by city and business
```

READY

```
%pig
businessData = FOREACH a GENERATE (int)json#'review_count' as review_count,(double
    )json#'stars' as stars, json#'city' as city, json#'categories' as categories;
flattenedBusinessData = FOREACH businessData GENERATE review_count,stars, city, FLATTEN
    (categories);
groupedBusinessData = GROUP flattenedBusinessData BY (city,categories);
```

READY

```
(,Pizza,4.0,3.5)
(,Fashion,5.0,4.0)
(,Italian,4.0,3.5)
(,Shopping,6.0,3.25)
(,Restaurants,4.0,3.5)
(,Sports Wear,5.0,4.0)
(,Sporting Goods,5.0,4.0)
(,Shopping Centers,7.0,2.5)
(Oka,Food,3.0,4.5)
(Oka,Thai,3.0,2.5)
(Oka,Parks,10.0,4.0)
(Oka,Beaches,10.0,4.0)
(Oka,Active Life,10.0,4.0)
(Oka,Campgrounds,10.0,4.0)
(Oka,Restaurants,3.0,2.5)
(Oka,Specialty Food,3.0,4.5)
(Oka,Hotels & Travel,10.0,4.0)
(Oka,Fruits & Veggies,3.0,4.5)
```

Output exceeds 102400. Truncated.

```
--2. Pivot the business categories as columns, and show the average # stars for each
```

READY

```
%pig
q2_3 = FOREACH a GENERATE (double)json#'stars' as stars, json#'city' as city,
    json#'state' as state, json#'categories' as categories;
q2_2 = FOREACH q2_3 GENERATE stars, city, state, FLATTEN(categories);
q2_1 = GROUP q2_2 BY (city,state,categories);
q2 = FOREACH q2_1 GENERATE group.city as city,group.state as state, group.categories as
```

READY

```
(Oka,QC,Thai,2.5)
(Oka,QC,Parks,4.0)
(Oka,QC,Beaches,4.0)
(Oka,QC,Active Life,4.0)
(Oka,QC,Campgrounds,4.0)
(Oka,QC,Restaurants,2.5)
(Oka,QC,Specialty Food,4.5)
(Oka,QC,Hotels & Travel,4.0)
(Oka,QC,Fruits & Veggies,4.5)
(Phx,AZ,Bars,3.5)
(Phx,AZ,Cafes,3.5)
(Phx,AZ,Tacos,3.5)
(Phx,AZ,Mexican,3.5)
(Phx,AZ,Spanish,3.5)
(Phx,AZ,Beer Bar,3.5)
(Phx,AZ,Nightlife,3.5)
(Phx,AZ,Restaurants,3.5)
(Ajax,ON,Bars,3.086956521739134)
```

```
Output exceeds 102400. Truncated.
```

```
--3. What is the average rank (# stars) for businesses that are ‘Mexican’ category, AND
```
READY

```
%pig
q3_1 = FOREACH a GENERATE (double)json#'stars' as stars,json#'attributes' as attributes
    , json#'categories' as categories:bag{a:tuple(b:chararray)} ;
q3_2 = FOREACH q3_1 GENERATE stars as stars,FLATTEN(attributes#'RestaurantsTakeOut') as
    takeout, FLATTEN(categories) as categories;
q3_3 = FILTER q3_2 BY (categories=='Mexican') AND (takeout matches '.*true.*');
q3_4 = GROUP q3_3 BY (takeout,categories);
```
READY

```
(true,Mexican,3.436754507628294)
```

```
--4. For businesses within 15km of Toronto center, show the average # stars and average #
```
READY

```
%pig
q4_1 = FOREACH a GENERATE (double)json#'stars' as stars, json#'latitude' as latitude,
    json#'longitude' as longitude, (double)json#'review_count' as reviews,
    json#'categories' as categories:bag{a:tuple(b:chararray)} ;
```
READY

```
%pig
q4_3 = FOREACH q4_2 GENERATE *, (111.045* org.apache.pig.piggybank.evaluation.math
    .toDegrees(ACOS(COS(org.apache.pig.piggybank.evaluation.math.toRadians(43.6532))*
    COS(org.apache.pig.piggybank.evaluation.math.toRadians(latitude))* COS(org.apache
    .pig.piggybank.evaluation.math.toRadians(-79.3832) - org.apache.pig.piggybank
    .evaluation.math.toRadians(longitude))+ SIN(org.apache.pig.piggybank.evaluation
    .math.toRadians(43.6532))* SIN(org.apache.pig.piggybank.evaluation.math.toRadians
    (latitude))))) as distance:double;
```
READY

```
(British,3.5,60.39473684210526)
(Buffets,3.0657894736842106,53.89473684210526)
(Burgers,3.12291169451074,37.76133651551313)
(Burmese,3.3333333333333335,15.333333333333334)
(Butcher,4.071428571428571,13.742857142857142)
(Cabaret,4.0,10.0)
(Casinos,3.2,10.9)
(Chinese,3.28743961352657,41.08212560386473)
(DUI Law,3.5,7.5)
(Dim Sum,3.336734693877551,64.0)
(Doctors,3.4685534591194966,8.10691823899371)
(Donairs,3.25,27.125)
(Embassy,2.75,8.0)
(Falafel,3.6296296296296298,18.296296296296298)
(Fashion,3.44361833952912,8.863692688971499)
(Fishing,4.0,6.0)
(Framing,4.0,7.117647058823529)
(Grocery,3.443820224719101,15.960674157303371)
```

```
--Q5. For the top 10 and bottom 10 food businesses near Toronto (ranked by stars),
```
READY

```
%pig
q5_1 = FOREACH a GENERATE (double)json#'stars' as stars,json#'business_id' as id,
    json#'latitude' as latitude, json#'longitude' as longitude, json#'categories' as
    categories:bag{a:tuple(b:chararray)} ;
q5_2 = FOREACH q5_1 GENERATE stars,id,latitude,longitude, FLATTEN(categories) as
    categories;
q5_3 = FOREACH q5_2 GENERATE *, (111.045* org.apache.pig.piggybank.evaluation.math
    .toDegrees(ACOS(COS(org.apache.pig.piggybank.evaluation.math.toRadians(43.6532))*
    COS(org.apache.pig.piggybank.evaluation.math.toRadians(latitude))* COS(org.apache
    .pig.piggybank.evaluation.math.toRadians(-79.3832) - org.apache.pig.piggybank
    .evaluation.math.toRadians(longitude))+ SIN(org.apache.pig.piggybank.evaluation
```
READY

```
(-0DwB6Swi349EKfbBAOF7A,3.5)
(-0NrB58jqKqJfuUCDupcsw,3.5)
(-25X5v1q3WU6s-craJSvTw,3.5)
(-4eA7Um2eiOKsGLmcXNx_w,4.0)
(-6CGECRbeyTCeyU4oHeXHQ,2.5)
(-76didnxGiiMO80BjSpYsQ,3.0)
(-9zPSrzbZ81FismxDSGLtA,5.0)
(-BAUrljU90RNV-hkokLhXA,3.5)
(-BJ0Z28LoETB_ZsdA4Ikeg,3.0)
(-BSf1LtnOrtCWAzmumxUpw,4.0)
(-Btu8zliXgeSH4eBm1u9Rw,4.0)
(-BvRroH7Q2CbJnp-IygX3Q,4.5)
(-EFDz-s9QUWJbFIp16O_3g,2.0)
(-Ej5bLw_bYuKt20kBFQi3w,4.0)
(-FHjXYCSizyNgUv-EXn6Yg,4.0)
(-IXNFjtECsn8fqF047tYFw,3.5)
(-IvATB9k2qNz19Gy0Q2NQw,3.5)
(-IzV9Xtwdnh8CCvTPwdIfw,3.0)
```

```
%pig                                                                    READY
q5_top = ORDER q5_6 BY stars DESC;
top10 = limit q5_top 10;
dump top10;
```

```
(xUsYf7lB0bi1zGEw41-aTw,5.0)
(oN8pqCTXY4ac4DmUXUfdvQ,5.0)
(aEA010Lba1mOCq976XThyw,5.0)
(6AeBlimS00y7CdhuhjpjRg,5.0)
(6T8YFkN7xkGlBfLMqdp28w,5.0)
(6aOnrzf15RMqFNOQ-_ElIA,5.0)
(V92rbUoSYcebJx42d10GZw,5.0)
(0j3ScXP2pii16Y4ojtKdSQ,5.0)
(6kIlmP82sIq2jxhNGUkEtg,5.0)
(DwG7_vYztZP-AMXEIvGgFA,5.0)
```

```
%pig                                                                    READY
q5_bottom= ORDER q5_6 BY stars ASC;
bottom10 = limit q5_bottom 10;
dump bottom10;
```

```
(V4226pZ4bN0mtEGeT7xkuQ,1.0)
(ldFrsUNIkDbSPYnspn8Pxg,1.0)
(WvMkxBdYLTh8ikog84ghCA,1.0)
(wl5lXES4GqDKJ00Z_c7ZVg,1.0)
(CHf_Uk6x6pF740PA6amvXw,1.0)
(u7bjH0lJcE7Q4BFlKTPJcg,1.0)
(85atsQTkgBz5tl_F2M4ZtA,1.0)
(UaoAGXDJPcpP7yObGZGD_Q,1.0)
(OaKWXPZl3yfEbhcGWFGTCw,1.0)
(7T1XVTSocHcOu2H5v2Iqog,1.0)
```

```
%pig                                                                                    READY
r = LOAD '/shared/d/yg/Downloads/dataset/review.json' USING com.twitter.elephantbird
    .pig.load.JsonLoader('-nestedLoad') as (json:map[]);
r1= FOREACH r GENERATE json#'stars' as stars,json#'business_id' as id, json#'date' as
    date;
```

```
(aEA010LbaImUCq976XInyw,5.0)
(6AeBlimS00y7CdhuhjpjRg,5.0)
(6T8YFkN7xkGlBfLMqdp28w,5.0)
(6aOnrzf15RMqFNOQ-_ElIA,5.0)
(V92rbUoSYcebJx42d10GZw,5.0)
(0j3ScXP2pii16Y4ojtKdSQ,5.0)
(6kIlmP82sIq2jxhNGUkEtg,5.0)
(DwG7_vYztZP-AMXEIvGgFA,5.0)
(V4226pZ4bN0mtEGeT7xkuQ,1.0)
(ldFrsUNIkDbSPYnspn8Pxg,1.0)
(WvMkxBdYLTh8ikog84ghCA,1.0)
(wl5lXES4GqDKJ00Z_c7ZVg,1.0)
(CHf_Uk6x6pF740PA6amvXw,1.0)
(u7bjH0lJcE7Q4BFlKTPJcg,1.0)
(85atsQTkgBz5tl_F2M4ZtA,1.0)
(UaoAGXDJPcpP7yObGZGD_Q,1.0)
(OaKWXPZl3yfEbhcGWFGTCw,1.0)
(7T1XVTSocHcOu2H5v2Iqog,1.0)
```

```
%pig                                                                                    READY
joined_r = JOIN r1 by id, top10bottom10 by id;
final_required_Data = FOREACH joined_r GENERATE top10bottom10::id as bid, (double)r1
    ::stars as star,SUBSTRING(r1::date,5,7) as month;
filtered_data_by_month = FILTER final_required_Data BY (month matches '01|02|03|04|05'
```

```
(85atsQTkgBz5tl_F2M4ZtA,1.0,04)
(85atsQTkgBz5tl_F2M4ZtA,1.0,05)
(V4226pZ4bN0mtEGeT7xkuQ,1.0,02)
(V4226pZ4bN0mtEGeT7xkuQ,1.0,02)
(oN8pqCTXY4ac4DmUXUfdvQ,5.0,05)
(oN8pqCTXY4ac4DmUXUfdvQ,5.0,01)
(OaKWXPZl3yfEbhcGWFGTCw,2.0,04)
(OaKWXPZl3yfEbhcGWFGTCw,1.0,03)
(OaKWXPZl3yfEbhcGWFGTCw,1.0,05)
(UaoAGXDJPcpP7yObGZGD_Q,1.0,03)
(UaoAGXDJPcpP7yObGZGD_Q,1.0,02)
(UaoAGXDJPcpP7yObGZGD_Q,1.0,02)
(u7bjH0lJcE7Q4BFlKTPJcg,1.0,03)
(wl5lXES4GqDKJ00Z_c7ZVg,1.0,05)
(0j3ScXP2pii16Y4ojtKdSQ,5.0,02)
(0j3ScXP2pii16Y4ojtKdSQ,5.0,02)
(6aOnrzf15RMqFNOQ-_ElIA,5.0,02)
(6aOnrzf15RMqFNOQ-_ElIA,5.0,01)
```

```
%pig
grouped_data_by_business = GROUP filtered_data_by_month by bid;
avg_rating = FOREACH grouped_data_by_business GENERATE group, AVG
    (filtered_data_by_month.star) as avg_stars;
```

READY

```
(0j3ScXP2pii16Y4ojtKdSQ,5.0)
(6T8YFkN7xkGlBfLMqdp28w,5.0)
(6aOnrzf15RMqFNOQ-_ElIA,5.0)
(7T1XVTSocHcOu2H5v2Iqog,1.0)
(85atsQTkgBz5tl_F2M4ZtA,1.0)
(CHf_Uk6x6pF740PA6amvXw,1.5)
(OaKWXPZl3yfEbhcGWFGTCw,1.3333333333333333)
(UaoAGXDJPcpP7yObGZGD_Q,1.0)
(V4226pZ4bN0mtEGeT7xkuQ,1.0)
(WvMkxBdYLTh8ikog84ghCA,1.0)
(aEA010Lba1mOCq976XThyw,5.0)
(ldFrsUNIkDbSPYnspn8Pxg,1.0)
(oN8pqCTXY4ac4DmUXUfdvQ,5.0)
(u7bjH0lJcE7Q4BFlKTPJcg,1.0)
(wl5lXES4GqDKJ00Z_c7ZVg,1.0)
```

```
%pig
```

READY