

review_and_category_analytics_w_rdds_JL

January 25, 2021

0.1 Project 1: Analytics on Glassdoor Reviews and Yelp Category Data

0.1.1 University of California, Santa Barbara

0.1.2 PSTAT 135/235: Big Data Analytics

0.1.3 Last Updated: May 30, 2020

0.1.4 OBJECTIVE

In this assignment, you will perform some basic analytics on review and category data.

This will entail performing operations on *RDDs*, and using *list comprehensions*.

Read in the dataset and perform the steps requested below.

TOTAL POINTS = 10

0.1.5 Config Setup

```
[2]: from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .master("local") \
    .appName("review_and_category_analytics") \
    .config("spark.executor.memory", '8g') \
    .config('spark.executor.cores', '4') \
    .config('spark.cores.max', '4') \
    .config("spark.driver.memory", '8g') \
    .getOrCreate()

sc = spark.sparkContext
```

Read in the dataset

```
[3]: df = sc.textFile("reviews_and_categories.csv")
```

get non-header records

```
[4]: header = df.first()
data = df.filter(lambda n: n != header) \
    .map(lambda row: (row.split(',')[0], row.split(',')[1:])) \
    .map(lambda x: (x[0].split(',')[0], ','.join(x[0].split(',')[1:]), x[1]))
```

print the first 2 records (note: exclude the header in all calculations)

```
[5]: print(data.take(2))
```

```
[('0', '', ['\'point of interest\'', \'mexican\'', \'establishment\'', \'food\'',
\'restaurant\']')], ('1', '', [''])]
```

1) get a record count (2 POINTS)

```
[10]: data.count()
```

```
[10]: 1305
```

store records with non-empty *review_emp_txt*

2) get a count of records with non-missing reviews (2 POINTS)

```
[15]: data_filter = data.filter(lambda r: '' not in r[1]) \
    .filter(lambda r: ' ' not in r[1])

data_filter.count()
```

```
[15]: 504
```

3) Return the count of records where review contains the word *awesome* (1 POINT)

```
[16]: pythonAwesome = data.filter(lambda r: "awesome" in r[1])
pythonAwesome.count()
```

```
[16]: 10
```

Print the records where review contains the word *awesome*

```
[17]: pythonAwesome.take(10)
```

```
[17]: [('280',
    '"Manager Pros Great environment awesome owners! I was happy to come to work
every day and face any challenges presented. It was a pleasant environment with
a lot of opportunity to advance if you worked hard. Cons I had performed the
role of a manager long before I was given the raise to match. There is a high
turn over rate and it s hard to find good team members who want to work
```

hard.""',

['\entertainment\','credit cards\','restaurants\','green smoothie\','colleges and universities\','smoothies and juice bars\','menus\','food, beverages & tobacco\','1\','juice bar\','establishment\','las vegas\','meal takeaway\','price\','juice bars & smoothies\','hospitality\','tallahassee\','point of interest\','shopping\','sandwiches\','health foods\','wheat grass\','other\','food\','smoothies\','hampton university\','franchises\','\$\$\','restaurant chains\','blimey limey\','food and beverages\','dinner, lunch & more\','eating places\','restaurant\','orlando\','outdoor seating\']""'],

('325',

"Lascari s Associate Pros The staff you become close with the cooks are awesome and friendly you begin to have regular customers that come to visit you. Cons Managment is awful not organized only car about family and relatives that work for them dont expect to move up without any type of hassle or an epic load of work.""',

['\whittier\','point of interest\','italian\','restaurant\','pizza\','establishment\','other\','food\']""'],

('382',

"Manager Pros Great environment awesome owners! I was happy to come to work every day and face any challenges presented. It was a pleasant environment with a lot of opportunity to advance if you worked hard. Cons I had performed the role of a manager long before I was given the raise to match. There is a high turn over rate and it s hard to find good team members who want to work hard.""',

['\private lot\','credit cards\','wheelchair accessible\','no reservations\','restroom\','menus\','smoothie shop\','drive through\','1\','dessert, lunch & more\','juice bar\','establishment\','price\','juice bars & smoothies\','slowest drive thru ever\','parking\','point of interest\','shopping\','other\','food\','smoothies\','dining options\','\$\$\','café\','2 get 1 free\','nutrition\','eating places\','table service & take-out\','restaurant\','orlando\','outdoor seating\']""'],

('666',

"Very awesome. Pros They allow the use of flexible schedule. Cons There are too many hostesses.""',

['\buffets\','chinese\','japanese\','point of interest\','eating places\','restaurant\','establishment\','other\','food\']""'],

('698',

"Epidemiology Interviewer II Pros PHI is a great non profit! Great projects awesome people who care for the well being of several different communities. Great benefits too! Cons There are so many different projects that you don t really get to know anyone outside your own.""',

['\health care\','housing\','gym\','immigration\','public health and safety\','health\','philanthropy\','noncommercial research organizations\','journalism\','public health\','point of interest\','california\','other\','establishment\','healthcare reform\']""'],

('714',

'Not a bad place to work Pros The couple that owned the one I worked at were very big on school and very flexible with my school schedule Being a delivery driver is great because the tips are awesome Cons The owner was kind of crazy I m kind of a health freak and just don t like working with pizza"',

['\credit cards\','no reservations\','no outdoor seating\','takeout\','food delivery\','menus\','order food online\','establishment\','pizza delivery\','local pizza\','drinks\','price\','sacramento pizza\','point of interest\','sandwiches\','through the garden\','pizza\','order pizza online\','other\','food\','meal delivery\','pizza menu\','dining options\','beer\','mypizza.com\','pizzeria\','extreme pizza\','\$\$\','salad\','pizza place\','delivery\','gluten-free restaurant\','pints on sunday\','dinner, lunch & more\','restaurant\']"])),

('854',

"Cashier Play Attendant Pros The kids are awesome it feels like a family everyone is willing to help awesome food discount Cons needs new equipment like a new espresso machine coffee maker and blender""',

['\entertainment\','credit cards\','kids activities\','menus\','establishment\','cupcakes\','price\','magazines\','hilarly duff\','point of interest\','reservations\','pesto panini\','other\','food\','dining options\','play area\','american (traditional)\','café\','doulas\','venues & event spaces\','dinner, lunch & more\','eating places\','restaurant\','playground\','outdoor seating\']"])),

('903',

"Manager Pros Great environment awesome owners! I was happy to come to work every day and face any challenges presented. It was a pleasant environment with a lot of opportunity to advance if you worked hard. Cons I had performed the role of a manager long before I was given the raise to match. There is a high turn over rate and it s hard to find good team members who want to work hard.""',

['\breakfast\','entertainment\','credit cards\','restaurants\','no reservations\','no outdoor seating\','colleges and universities\','smoothies and juice bars\','menus\','1\','no wi-fi\','juice bar\','establishment\','wraps and smoothies\','restaurant\','las vegas\','price\','flatbreads\','juice bars & smoothies\','meal takeaway\','hospitality\','cafe\','tallahassee\','point of interest\','shopping\','sandwiches\','health foods\','other\','food\','kiwi citrus green tea\','smoothies\','dining options\','hampton university\','franchises\','restaurant chains\','sandwich place\','kiwi quencher\','café\','food and beverages\','delivery\','juice bars & smoothies\','wraps\','orlando\','dinner, breakfast & more\']"])),

('923',

"Manager Pros Co workers management and overall experience of an awesome brand! Incredible culture awesome food and beautiful diverse customer base. The company is innovative and growing. It is an exciting time to be a part of the team. Cons Need more locations to show the world the brand! As we grow people will see the incredible potential of SkinnyFATS.""',

```
[\'breakfast\', \'lunch\', \'private lot\', \'entertainment\', \'credit
cards\', \'restaurants\', \'no reservations\', \'smart watches\', \'restroom\',
\'burger\', \'menus\', \'pancake\', \'waffles\', \'juices\', \'new american
restaurant\', \'free wi-fi\', \'tacos\', \'juice bar\', \'establishment\',
\'bacon\', \'ahi\', \'cookies\', \'st. lawrence university\', \'las vegas\',
\'cold press juice\', \'price\', \'juice bars & smoothies\', \'sandwich\',
\'djs\', \'tvs\', \'slow boat to china\', \'parking\', \'point of interest\',
\'taco\', \'sandwiches\', \'seafood\', \'eggs\', \'pizza\', \'vegan\',
\'other\', \'food\', \'dining options\', \'truffle fries\', \'juice\',
\'shake\', \'breakfast & brunch\', \'gluten-free\', \'bellagio hotel and
resort\', \'$$\', \'dinner\', \'chix on broadway\', \'breakfast spot\', \'cold-
pressed juice\', \'american (new)\', \'salad\', \'management\', \'delivery\',
\'brunch\', \'burgers\', \'shakes\', \'store\', \'pancakes\', \'salads\', \'wi-
fi\', \'vegetarian\', \'restaurant\', \'healthy\', \'dessert\', \'dinner,
breakfast & more\', \'waffle\', \'fresh juice\', \'outdoor seating\']"]],
('1174',
'"Potato Corner Pros flexible hours awesome coworkers free meal during break
Cons super busy during weekends (make sure you re able to stand for more than 5
hrs straight) sometimes its hard to even go on break because it gets so busy"',
[\'meal takeaway\', \'point of interest\', \'fast food\', \'eating places\',
\'restaurant\', \'establishment\', \'other\', \'food\']"]])]
```

4) Lowercase all reviews, then return the count of records where review contains the word *awesome* (1 POINT)

```
[27]: df3 = data.map(lambda r: r[1].lower()).filter(lambda r: "awesome" in r)
df3.count()
```

[27]: 12

4) Return the top 10 most frequent categories (4 POINTS)

Preprocess the categories by:

- * stripping characters: [] ' "
- * trim spaces before and after words
- * lowercase

NOTE: Be sure to keep terms together, for example 'jet skiing' should not become 'jet', 'skiing'

```
[63]: categories=data.flatMap(lambda r: r[2])
```

```
[195]: wordcount = categories.map(lambda x: x.replace('[', ' ') \
    .replace(']', ' ').replace(' ',' ').strip(' ').lower()) \
    .flatMap(lambda x: x.split(",")) \
    .filter(lambda x: x != '') \
    .map(lambda x: (x,1)) \
    .reduceByKey(lambda x,y:x+y) \
    .map(lambda x: (x[1],x[0])) \
    .sortByKey(False)
```

```
[196]: wordcount.take(10)
```

```
[196]: [(705, " 'establishment'"),
        (685, " 'food'"),
        (671, " 'point of interest'"),
        (642, " 'restaurant'"),
        (473, " 'price'"),
        (471, " 'other'"),
        (311, " 'menus'"),
        (254, " 'eating places'"),
        (253, " 'dining options'"),
        (232, " 'credit cards'")]
```

```
[197]: # Save notebook as PDF document
!jupyter nbconvert --to pdf `pwd`/*.ipynb
```

```
[NbConvertApp] Converting notebook
/home/jovyan/assignments/M2_10/review_and_category_analytics_w_rdds_JL.ipynb to
pdf
[NbConvertApp] Writing 51671 bytes to ./notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', './notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', './notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 59632 bytes to
/home/jovyan/assignments/M2_10/review_and_category_analytics_w_rdds_JL.pdf
```

```
[ ]:
```