Sparkify

October 11, 2019

1 Sparkify Project Workspace

This workspace contains a tiny subset (128MB) of the full dataset available (12GB). Feel free to use this workspace to build your project, or to explore a smaller subset with Spark before deploying your cluster on the cloud. Instructions for setting up your Spark cluster is included in the last lesson of the Extracurricular Spark Course content.

You can follow the steps below to guide your data analysis and model building portion of this project.

In [1]:
```
# import libraries
# import libraries
from pyspark.sql import SparkSession
from pyspark.sql.functions import avg, col, concat, desc, explode, lit, min, max, split,
from pyspark.sql.types import IntegerType

from pyspark.ml import Pipeline
from pyspark.ml.classification import LogisticRegression, RandomForestClassifier, GBTCla
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml.feature import CountVectorizer, IDF, Normalizer, PCA, RegexTokenizer, St
from pyspark.ml.regression import LinearRegression
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder

import re
import datetime
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [2]:
```
# create a Spark session
spark = SparkSession \
    .builder \
    .appName("Sparkify") \
    .getOrCreate()
```

2 Load and Clean Dataset

In this workspace, the mini-dataset file is `mini_sparkify_event_data.json`. Load and clean the dataset, checking for invalid or missing data - for example, records without userids or sessionids.

```
In [3]: # Read in full sparkify dataset
        # event_data = "s3n://dsnd-sparkify/sparkify_event_data.json"
        sparkify_data = 'mini_sparkify_event_data.json'
        df = spark.read.json(sparkify_data)
        df.persist()

Out[3]: DataFrame[artist: string, auth: string, firstName: string, gender: string, itemInSession
```

2.0.1 Clean the NA value in userId and sessionId

```
In [4]: df.count()

Out[4]: 286500

In [5]: df = df.dropna(how = "any", subset = ["userId", "sessionId"])
        df.count()

Out[5]: 286500
```

2.0.2 Clean the empty userId

```
In [6]: df = df.filter(df.userId!="")
        df.count()

Out[6]: 278154
```

3 Exploratory Data Analysis

When you're working with the full dataset, perform EDA by loading a small subset of the data and doing basic manipulations within Spark. In this workspace, you are already provided a small subset of data you can explore.

3.0.1 Define Churn

Once you've done some preliminary analysis, create a column `Churn` to use as the label for your model. I suggest using the `Cancellation Confirmation` events to define your churn, which happen for both paid and free users. As a bonus task, you can also look into the `Downgrade` events.

3.0.2 Explore Data

Once you've defined churn, perform some exploratory data analysis to observe the behavior for users who stayed vs users who churned. You can start by exploring aggregates on these two groups of users, observing how much of a specific action they experienced per a certain time unit or number of songs played.

3.0.3 We can see different pages

```
In [7]: df.select("page").dropDuplicates().show()
```

```
+--------------------+
|                page|
+--------------------+
|              Cancel|
|    Submit Downgrade|
|         Thumbs Down|
|                Home|
|           Downgrade|
|         Roll Advert|
|              Logout|
|       Save Settings|
|Cancellation Conf...|
|               About|
|            Settings|
|     Add to Playlist|
|          Add Friend|
|            NextSong|
|           Thumbs Up|
|                Help|
|             Upgrade|
|               Error|
|      Submit Upgrade|
+--------------------+
```

3.0.4 Let's see one user who did "Cancellation Confirmation"

```
In [8]: df.filter(df.page=="Cancellation Confirmation").select("userId").dropDuplicates().show(1
```

```
+------+
|userId|
+------+
|   125|
|    51|
|    54|
|100014|
|   101|
|    29|
|100021|
|    87|
|    73|
|     3|
|    28|
|100022|
```

```
|100025|
|300007|
|100006|
+------+
only showing top 15 rows
```

```
In [9]: # add time to see the time clear
        get_time = udf(lambda x: datetime.datetime.fromtimestamp(x / 1000.0).strftime("%Y-%m-%d
        df = df.withColumn("time", get_time(df.ts))
```

```
In [10]: df.select(["userId", "page", "time", "level", "song", "sessionId"]).where(df.userId ==
```

```
+------+---------------+-------------------+-----+--------------------+---------+
|userId|           page|               time|level|                song|sessionId|
+------+---------------+-------------------+-----+--------------------+---------+
|    30|       NextSong|2018-10-01 00:01:57| paid|           Rockpools|       29|
|    30|       NextSong|2018-10-01 00:06:34| paid|   Time For Miracles|       29|
|    30|       NextSong|2018-10-01 00:11:16| paid|Harder Better Fas...|       29|
|    30|       NextSong|2018-10-01 00:14:59| paid|Passengers (Old A...|       29|
|    30|Add to Playlist|2018-10-01 00:15:05| paid|                null|       29|
|    30|       NextSong|2018-10-01 00:18:04| paid|          Fuck Kitty|       29|
|    30|       NextSong|2018-10-01 00:20:18| paid|                Jade|       29|
|    30|       NextSong|2018-10-01 00:24:01| paid|   So-Called Friends|       29|
|    30|       NextSong|2018-10-01 00:28:07| paid|           Represent|       29|
|    30|       NextSong|2018-10-01 00:31:49| paid|           Here I Am|       29|
|    30|       NextSong|2018-10-01 00:35:32| paid|Rebirthing (Album...|       29|
|    30|       NextSong|2018-10-01 00:39:25| paid|Dog Days Are Over...|       29|
|    30|       NextSong|2018-10-01 00:43:04| paid|Tomorrow Is A Lon...|       29|
|    30|       NextSong|2018-10-01 00:46:46| paid|    Halloween Spooks|       29|
|    30|       NextSong|2018-10-01 00:49:05| paid|            Stronger|       29|
|    30|       NextSong|2018-10-01 00:54:16| paid|   Dis Iz Brick City|       29|
|    30|       NextSong|2018-10-01 00:57:53| paid|          Move Along|       29|
|    30|       NextSong|2018-10-01 01:01:51| paid|           Manhattan|       29|
|    30|       NextSong|2018-10-01 01:05:15| paid|                Undo|       29|
|    30|       NextSong|2018-10-01 01:11:03| paid|     The Big Gundown|       29|
|    30|       NextSong|2018-10-01 01:15:23| paid|          Black Bird|       29|
|    30|    Thumbs Down|2018-10-01 01:15:24| paid|                null|       29|
|    30|       NextSong|2018-10-01 01:18:27| paid|              Nausea|       29|
|    30|       NextSong|2018-10-01 01:21:43| paid|           Matricide|       29|
|    30|       NextSong|2018-10-01 01:27:03| paid|             Valerie|       29|
|    30|       NextSong|2018-10-01 01:30:52| paid|           Margarita|       29|
|    30|       NextSong|2018-10-01 01:34:08| paid|   Le Jardin d'Hiver|       29|
|    30|      Thumbs Up|2018-10-01 01:34:09| paid|                null|       29|
|    30|       NextSong|2018-10-01 01:39:50| paid|Soon As I Get Hom...|       29|
|    30|      Thumbs Up|2018-10-01 01:39:51| paid|                null|       29|
|    30|       NextSong|2018-10-01 01:45:14| paid|    Vamos a la Playa|       29|
```