[P] /Spark/ Capstone Project

/ Submit Requirement

本项目要求的是将文件提交到 Git 上，并给出链接。课程中的要求是：

> Submission Instructions
> Create a GitHub repository for this project, containing your notebook and README file. Once your project is finished, submit the URL of this repository.

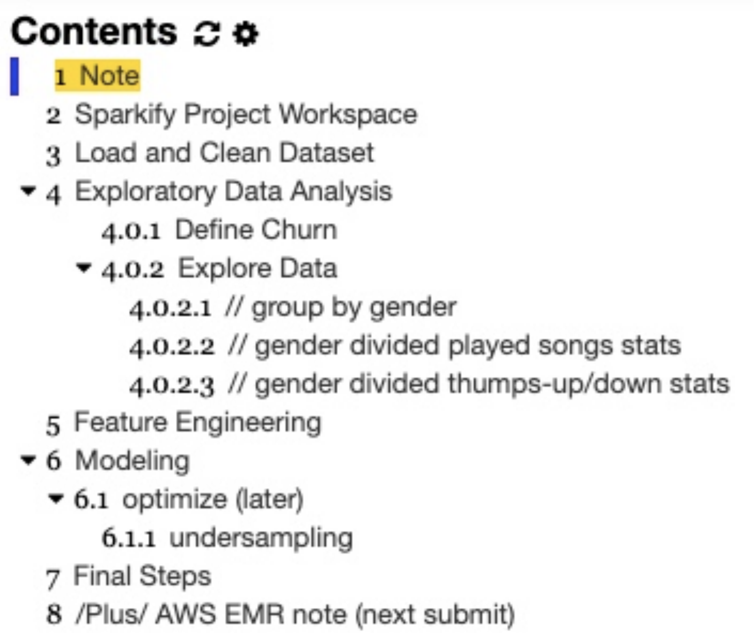
/ Note

作为 Git 的默认 README.md 文件，可以对项目进行简单的介绍和指导，便于评审老师进行审阅。我的分为以下几个部分（个人经验，不是参考答案）：

Hi Reviewer,

Thanks for your time to review my capestone project. I want to make some note at top to save your time on review my project.

- structure



Contents

1 Note
2 Sparkify Project Workspace
3 Load and Clean Dataset
4 Exploratory Data Analysis
4.0.1 Define Churn
4.0.2 Explore Data
4.0.2.1 // group by gender
4.0.2.2 // gender divided played songs stats
4.0.2.3 // gender divided thumps-up/down stats
5 Feature Engineering
6 Modeling
6.1 optimize (later)
6.1.1 undersampling
7 Final Steps
8 /Plus/ AWS EMR note (next submit)

- request for coaching
    - **Q1:How to practice Sprak at real world?** I am new to spark, I manage about 3 week to finish this project, and still get some unclear on part of it. So where can I get some real big data to explore the skill of spark (on AWS, not local). Can you share something?
    - **Q2:Spark vs Flink** I do some research that both Spark and Flink are super-stars nowadays. While Spark come from batch to stream, and Flink come from stream to batch. How shoud I choose from them. Or for the short, what are the pro and cons for Spark and Flink on differed situation?
- notice

    - data file: for the git limit < 100MB, I use a compressed file to complete(also the same with 128MB at local, if you want to r

- version
    - version1 -
    - version2 - finish uda workplace run out.
    - version3 - finish local run out. (mac env)
    - version4 - (finished at next submit) - aws emr run out, and take a new chaper at the end `/Plus/AWS EMR note`

Thanks a lot again, waiting for you reply.
Meng

/ Project Overview (Uda)

// What will I learn?

You'll learn how to manipulate large and realistic datasets with Spark to engineer relevant features for predicting churn. You'll learn how to use Spark MLlib to build machine learning models with large datasets, far beyond what could be done with non-distributed technologies like scikit-learn.

// Career Relevance

Predicting churn rates is a challenging and common problem that data scientists and analysts regularly encounter in any customer-facing business. Additionally, the ability to efficiently manipulate large datasets with Spark is one of the highest-demand skills in the field of data.

// Essential Skills

- Load large datasets into Spark and manipulate them using Spark SQL and Spark Dataframes
- Use the machine learning APIs within Spark ML to build and tune models
- Integrate the skills you've learned in the Spark course and the Data Scientist Nanodegree program

// Take the Spark Course (Pre Study Requirement)

You can find the Spark course in your Extracurriculars section here.

/ Project Instructions

The full dataset is 12GB, of which you can analyze a mini subset in the workspace on the following page. Optionally, you can choose to follow the instructions in the Extracurricular course to deploy a Spark cluster on the cloud using AWS or IBM Cloud to analyze a larger amount of data. Currently we have the full 12GB dataset available to you if you use AWS. If you use IBM, you can download a medium sized dataset to upload to your cluster.

Details on how to do this using AWS or IBM Cloud are included in the last lesson of the Extracurricular Spark Course content linked above. Note that this part is optional, and you will not receive credits to fund your deployment. You can do the IBM portion for free. Using AWS will cost you around $30 if you run a cluster up for a week with the settings we provide.

Once you've built your model, either in the classroom workspace or in the cloud with AWS or IBM, download your notebook and complete the remaining components of your Data Scientist Capstone project, including thorough documentation in a README file in your Github repository, as well as a web app or blog post explaining the technical details of your project. Be sure to review the Project Rubric thoroughly before submitting your project.

/ Plus.AWS

As mentioned in the class, After finish the cape stone project. I am still trying do the work at full data (12GB) on AWS EMR. The effort will later be noted.