Learning Apache Spark with R

Release 1.00

Wenqiang Feng, Wenhao Wang and Zhi Zheng

June 01, 2019

CONTENTS

Welcome to our Learning Apache Spark with R note! In this note, you will learn a wide array of concepts about SparkR in Data Mining, Text Mining, Machine Learning and Deep Learning. Some contents are from [Feng2017] The PDF version can be downloaded from HERE.

CONTENTS

PREFACE

1.1 About

1.1.1 About this note

This is a shared repository for Learning Apache Spark with R. The PDF version can be downloaded from HERE. The first version was posted on Github in ChenFeng ([Feng2017]). This shared repository mainly contains the self-learning and self-teaching notes from Wenqiang during his IMA Data Science Fellowship. The reader is referred to the repository https://github.com/runawayhorse001/LearningApacheSpark for more details about the `dataset` and the `.ipynb` files.

In this repository, I try to use the detailed demo code and examples to show how to use each main functions. If you find your work wasn't cited in this note, please feel free to let me know.

Although I am by no means an data mining programming and Big Data expert, I decided that it would be useful for me to share what I learned about PySpark programming in the form of easy tutorials with detailed example. I hope those tutorials will be a valuable tool for your studies.

The tutorials assume that the reader has a preliminary knowledge of programing and Linux. And this document is generated automatically by using sphinx.

1.1.2 About the authors

- **Wenqiang Feng**

    - Data Scientist and PhD in Mathematics

    - University of Tennessee at Knoxville

    - Email: von198@gmail.com

    Wenhao Wang

    - and PhD in

    - The University of Kansas

> – Email:

James Zheng

> – Sr. Data Scientist and PhD in

> – University of Tennessee at Knoxville

> – Email:

- **Biography**

  Wenqiang Feng is Data Scientist within DST's Applied Analytics Group. Dr. Feng's responsibilities include providing DST clients with access to cutting-edge skills and technologies, including Big Data analytic solutions, advanced analytic and data enhancement techniques and modeling.

  Dr. Feng has deep analytic expertise in data mining, analytic systems, machine learning algorithms, business intelligence, and applying Big Data tools to strategically solve industry problems in a cross-functional business. Before joining DST, Dr. Feng was an IMA Data Science Fellow at The Institute for Mathematics and its Applications (IMA) at the University of Minnesota. While there, he helped startup companies make marketing decisions based on deep predictive analytics.

  Dr. Feng graduated from University of Tennessee, Knoxville, with Ph.D. in Computational Mathematics and Master's degree in Statistics. He also holds Master's degree in Computational Mathematics from Missouri University of Science and Technology (MST) and Master's degree in Applied Mathematics from the University of Science and Technology of China (USTC).

- **Declaration**

  The work of Wenqiang Feng was supported by the IMA, while working at IMA. However, any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the IMA, UTK and DST.

1.2 Motivation for this tutorial

I was motivated by the IMA Data Science Fellowship project to learn PySpark. After that I was impressed and attracted by the PySpark. And I foud that:

1. It is no exaggeration to say that Spark is the most powerful Bigdata tool.

2. However, I still found that learning Spark was a difficult process. I have to Google it and identify which one is true. And it was hard to find detailed examples which I can easily learned the full process in one file.

3. Good sources are expensive for a graduate student.

1.3 Copyright notice and license info

This Learning Apache Spark with R PDF file is supposed to be a free and living document, which is why its source is available online at https://runawayhorse001.github.io/LearningApacheSpark/pyspark.pdf. But this document is licensed according to both MIT License and Creative Commons Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0) License.

When you plan to use, copy, modify, merge, publish, distribute or sublicense, Please see the terms of those licenses for more details and give the corresponding credits to the author.

1.4 Acknowledgement

At here, I would like to thank Ming Chen, Jian Sun and Zhongbo Li at the University of Tennessee at Knoxville for the valuable disscussion and thank the generous anonymous authors for providing the detailed solutions and source code on the internet. Without those help, this repository would not have been possible to be made. Wenqiang also would like to thank the Institute for Mathematics and Its Applications (IMA) at University of Minnesota, Twin Cities for support during his IMA Data Scientist Fellow visit.

A special thank you goes to Dr. Haiping Lu, Lecturer in Machine Learning at Department of Computer Science, University of Sheffield, for recommending and heavily using my tutorial in his teaching class and for the valuable suggestions.

1.5 Feedback and suggestions

Your comments and suggestions are highly appreciated. I am more than happy to receive corrections, suggestions or feedbacks through email (von198@gmail.com) for improvements.
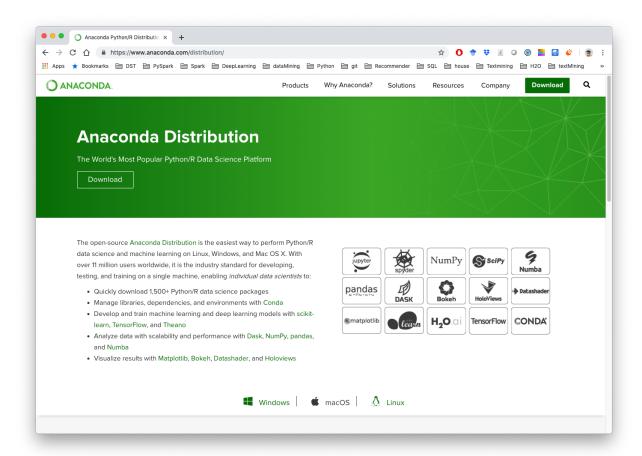
TWO

SPARKR INSTALLATION

Note: This Chapter *SparkR Installation* is for beginner. If you have some `Python` programming experience, you may skip this chapter.

No matter what operator system is, I will strongly recommend you to install `Anaconda` which contains `Python`, `Jupyter`, `spyder`, `Numpy`, `Scipy`, `Numba`, `pandas`, `DASK`, `Bokeh`, `HoloViews`, `Datashader`, `matplotlib`, `scikit-learn`, `H2O.ai`, `TensorFlow`, `CONDA` and more.

Download link: https://www.anaconda.com/distribution/

THREE

MAIN REFERENCE

BIBLIOGRAPHY

[Feng2017] Wenqiang Feng. Learning Apache Spark with Python, 2017.