

cs105_lab1b_word_count

databricks

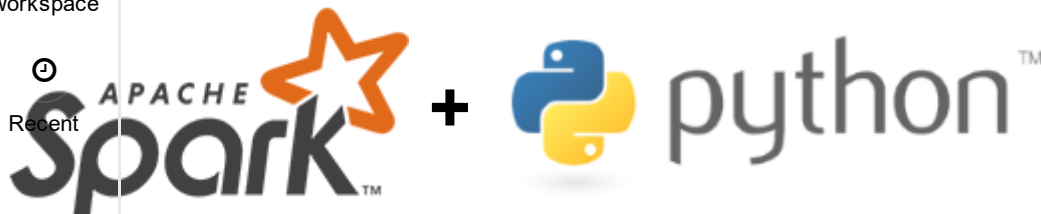


(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Workspace



Tables

Word Count Lab: Building a word count application

Clusters

This lab will build on the techniques covered in the Spark tutorial to develop a simple word count application. The volume of unstructured text in existence is growing dramatically, and Spark is an excellent tool for analyzing this type of data. In this lab, we will write code that calculates the most common words in the Complete Works of William Shakespeare (<http://www.gutenberg.org/ebooks/100>) retrieved from Project Gutenberg (http://www.gutenberg.org/wiki/Main_Page). This could also be scaled to larger applications, such as finding the most common words in Wikipedia.

During this lab we will cover:

- *Part 1:* Creating a base DataFrame and performing operations
- *Part 2:* Counting with Spark SQL and DataFrames
- *Part 3:* Finding unique words and a mean value
- *Part 4:* Apply word count to a file

Send feedback