cs120_lab1b_word_count_rdd

databricks

Home

Workspace

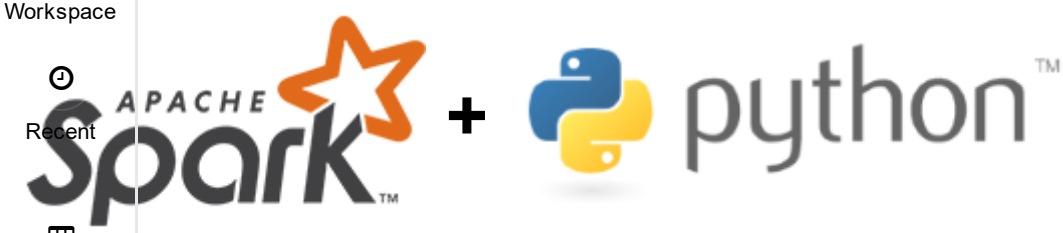
Recent

Tables

Clusters

Jobs

Search



Word Count Lab: Building a word count application

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 find the most common words in Wikipedia.

During this lab we will cover:

- *Part 1:* Creating a base RDD and pair RDDs
- *Part 2:* Counting with pair RDDs
- *Part 3:* Finding unique words and a mean value

Send feedback