**Exercise 2**

*Get started with Apache Spark and Python*

**Prior Knowledge**

Unix Command Line Shell

Simple Python

**Learning Objectives**

Understand the Spark system

Understand the Jupyter Notebook model

Submit Spark jobs locally

**Software Requirements**

* Apache Spark 2.3.0
* Python 3.7
* Jupyter notebooks

We are going to do a wordcount against a set of books downloaded from Project Gutenberg. Wordcount is the definitive Big Data program (sort of Hello World for Big Data). We are going to be using Apache Spark’s Python shell, in order to interactively test and run code.

1. You will probably first need to install Apache Spark. Download from here and store the unpacked version in your home directory

<http://archive.apache.org/dist/spark/spark-2.3.0/spark-2.3.0-bin-hadoop2.7.tgz>

For Spark to run you need to have [Java 8](https://www.oracle.com/java/technologies/javase-jdk8-downloads.html) installed on your machine. It should work with more recent versions of Java but some people report problems.

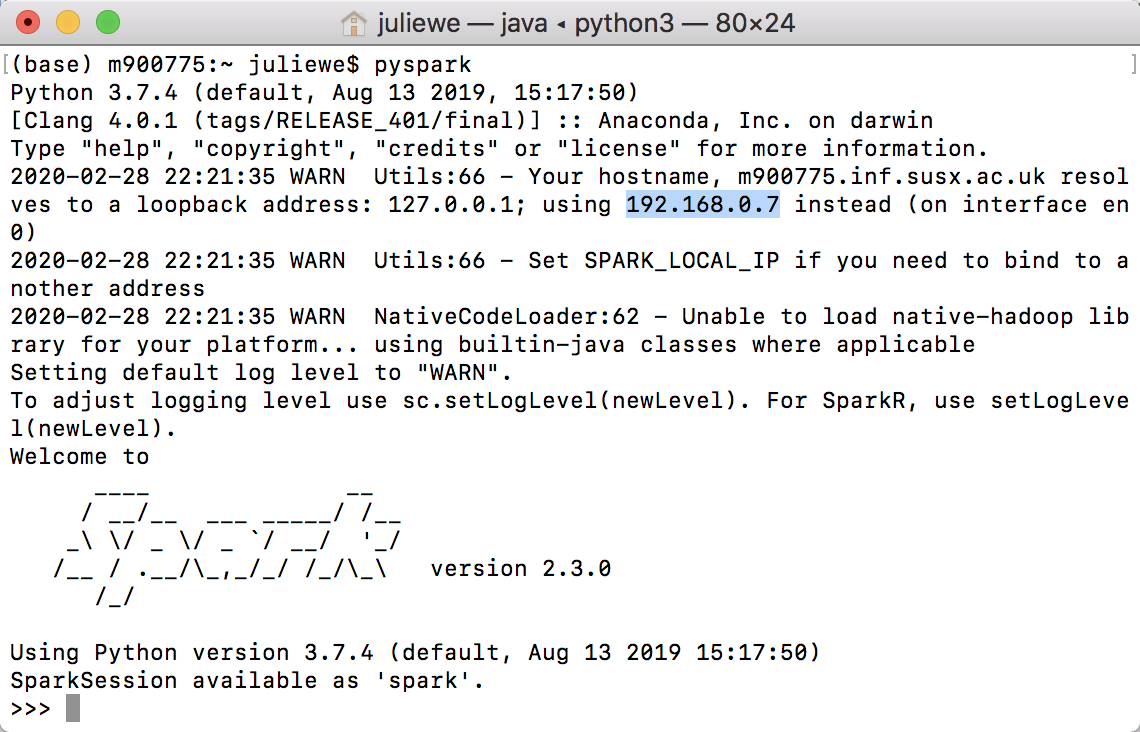
Change .bash\_profile variable settings. Add the following lines to your .bash\_profile in your home directory

export JAVA\_HOME=$(/usr/libexec/java\_home)  
export SPARK\_HOME=~/spark-2.3.0-bin-hadoop2.7  
export PATH=$SPARK\_HOME/bin:$PATH  
export PYSPARK\_PYTHON=python3

Run source ~/.bash\_profile to source this file or open a new terminal to auto-source it.

Now start the Spark Python command line tool by typing pyspark

You should see a lot of log come up, ending in something like:



1. If on mac or linux and you get an error on running pyspark, which is caused by a java.net.BindException, then you may need to update your hosts file as follows:
   1. Get your hostname using the hostname command.
   2. Use sudo nano /etc/hosts and add an entry to this file:

127.0.0.1 your\_hostname

1. Once you have tested that Spark is installed correctly, you can quit the “traditional” Spark Python command line tool as we aren’t going to use this just now. Type quit() to leave
2. You will also need to install the python packages pyspark and findspark into your bigdata environment

conda activate bigdata

conda install pyspark

pip install findspark

1. Now, if you haven’t done so already, make sure you have an up-to-date version of the github repository for this course. At a terminal window:

git clone <https://github.com/julieweeds/BigData.git>

If you have already cloned the repository, you can ensure it is uptodate with:

cd BigData

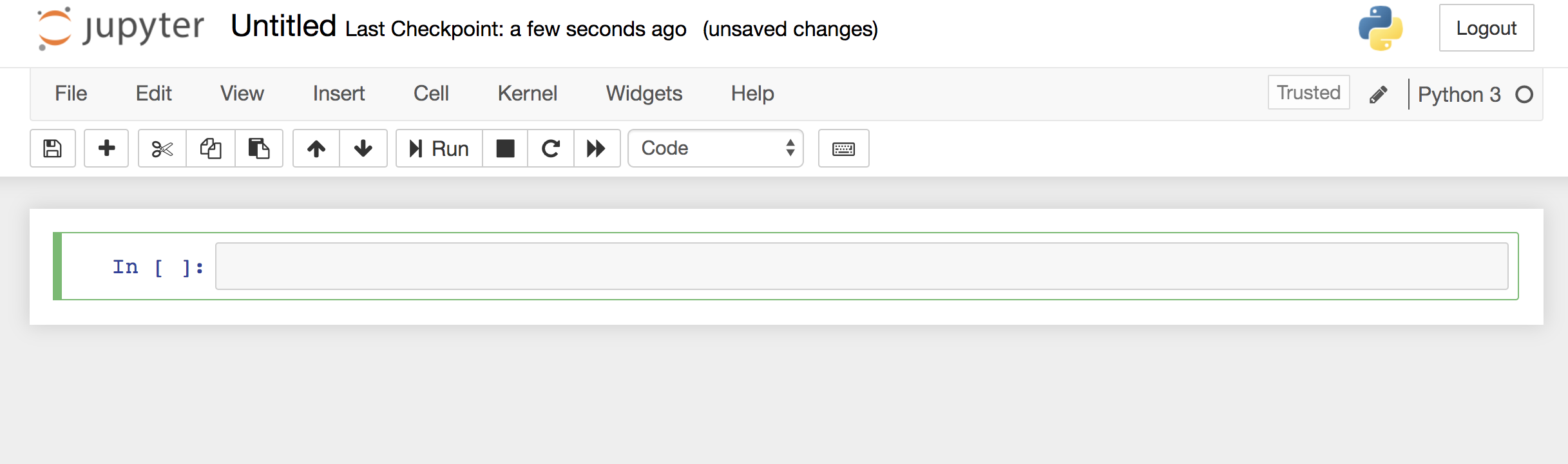
git pull

1. Let’s make a directory to store our code.

mkdir wordcount  
cd wordcount

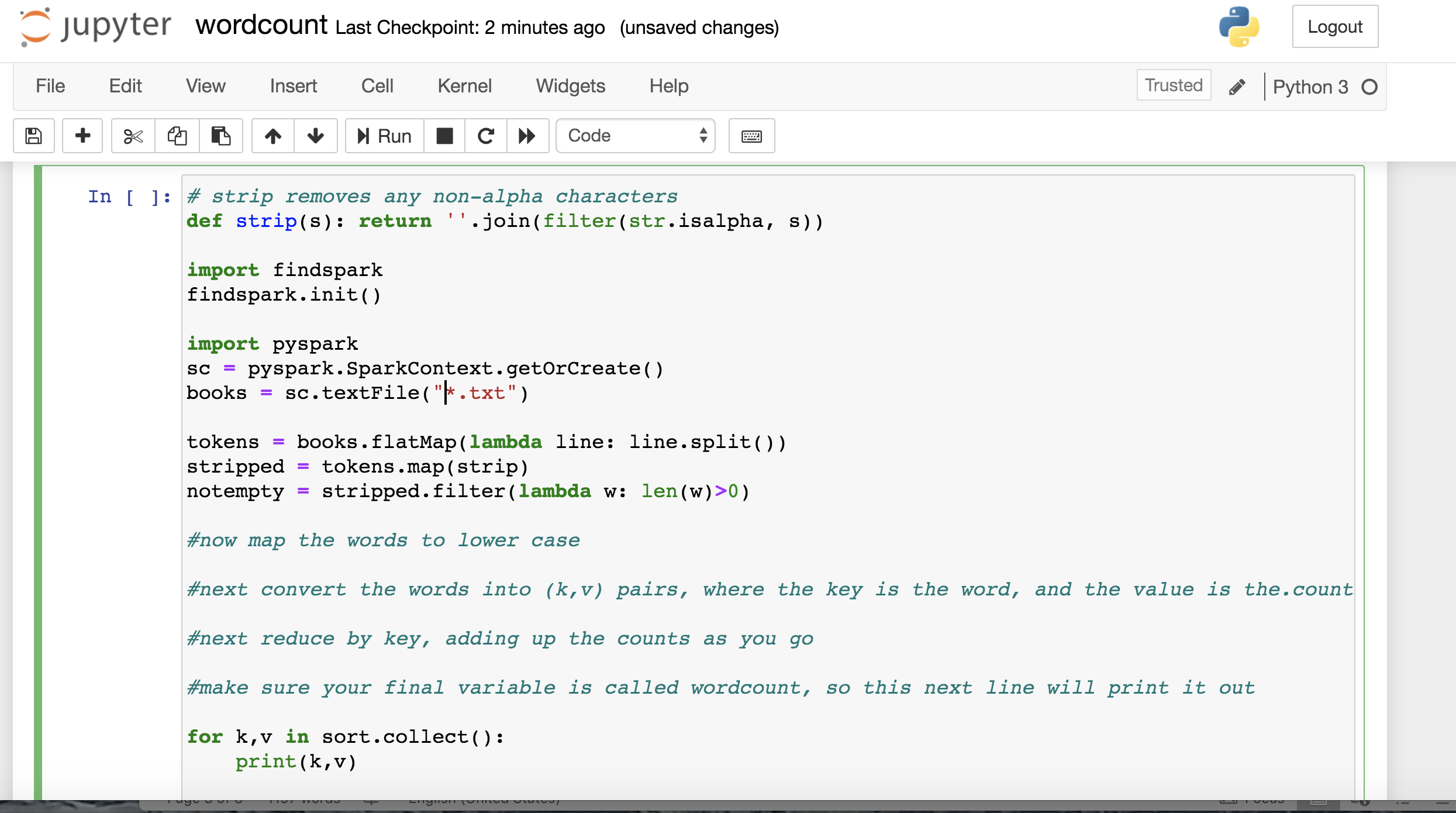
1. We need some books to do a wordcount on. I have included some in the BigData repository. Let’s make symbolic links to make it easier to access them (as if they were in the same directory)

ln -s ~/BigData/datafiles/books/\* ./

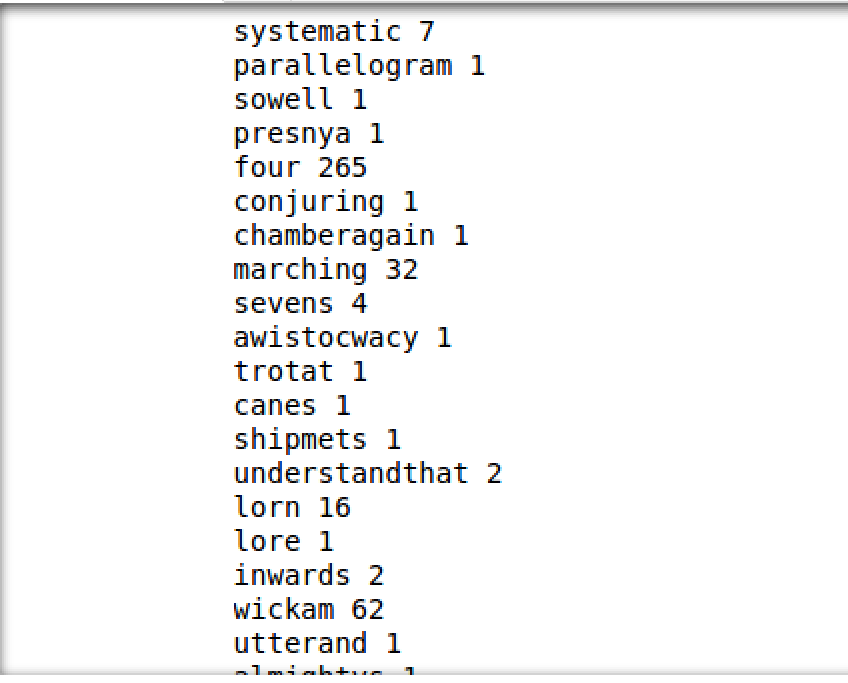
1. Start jupyter notebook and use the **New** button to create a new Python3 notebook:  
   
2. Click on the name of the notebook (currently “Untitled”) and rename it to “wordcount”
3. There is a starter of the code you need in github repository for this course. At the command line:

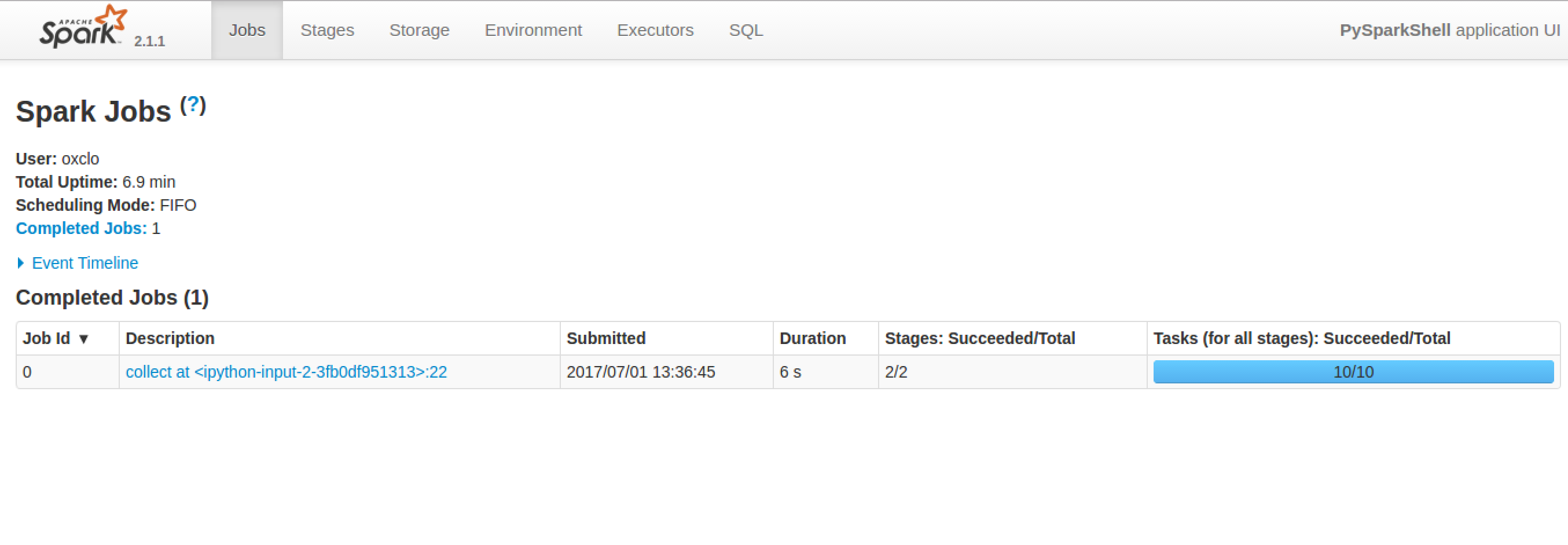
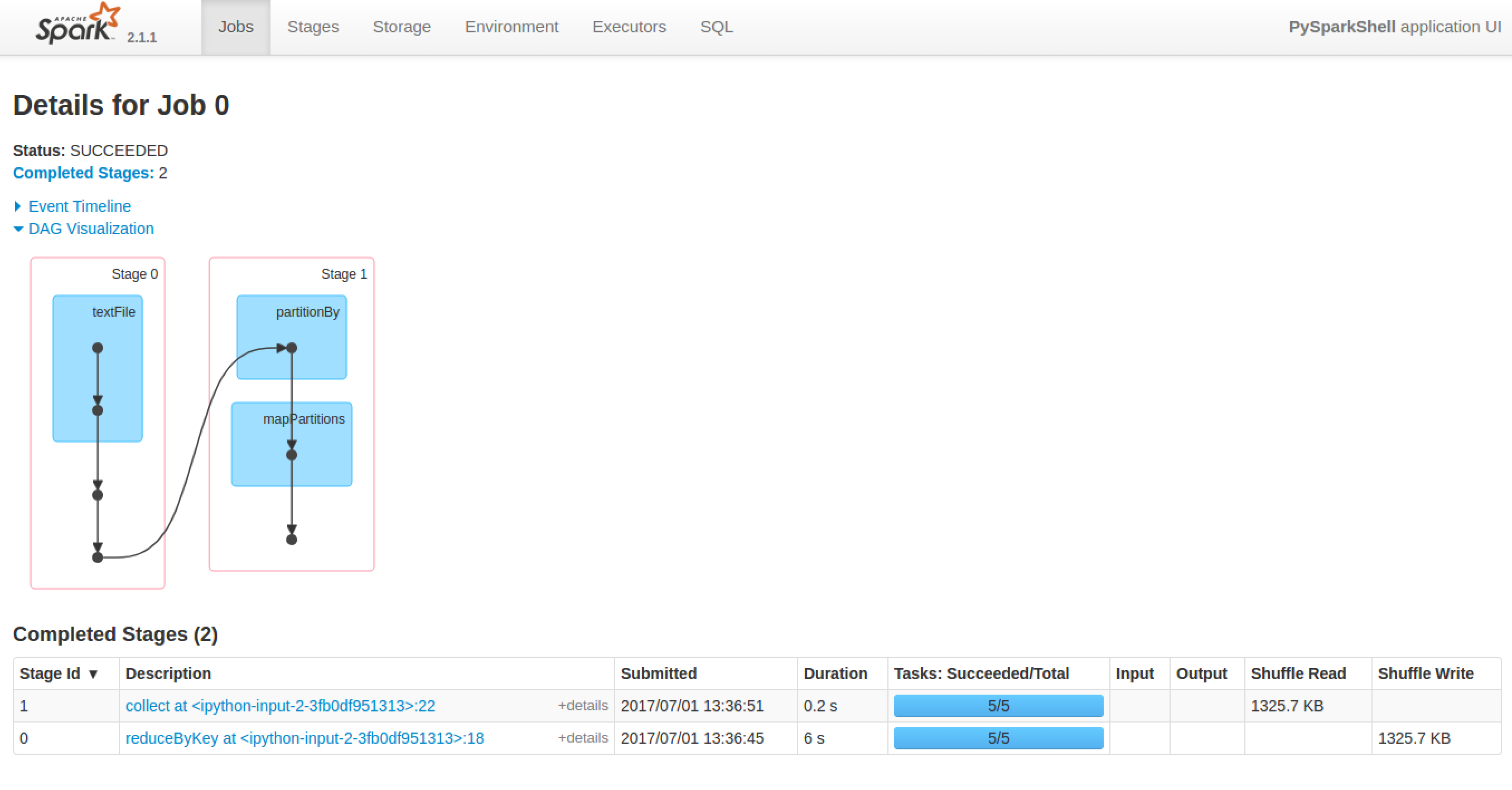
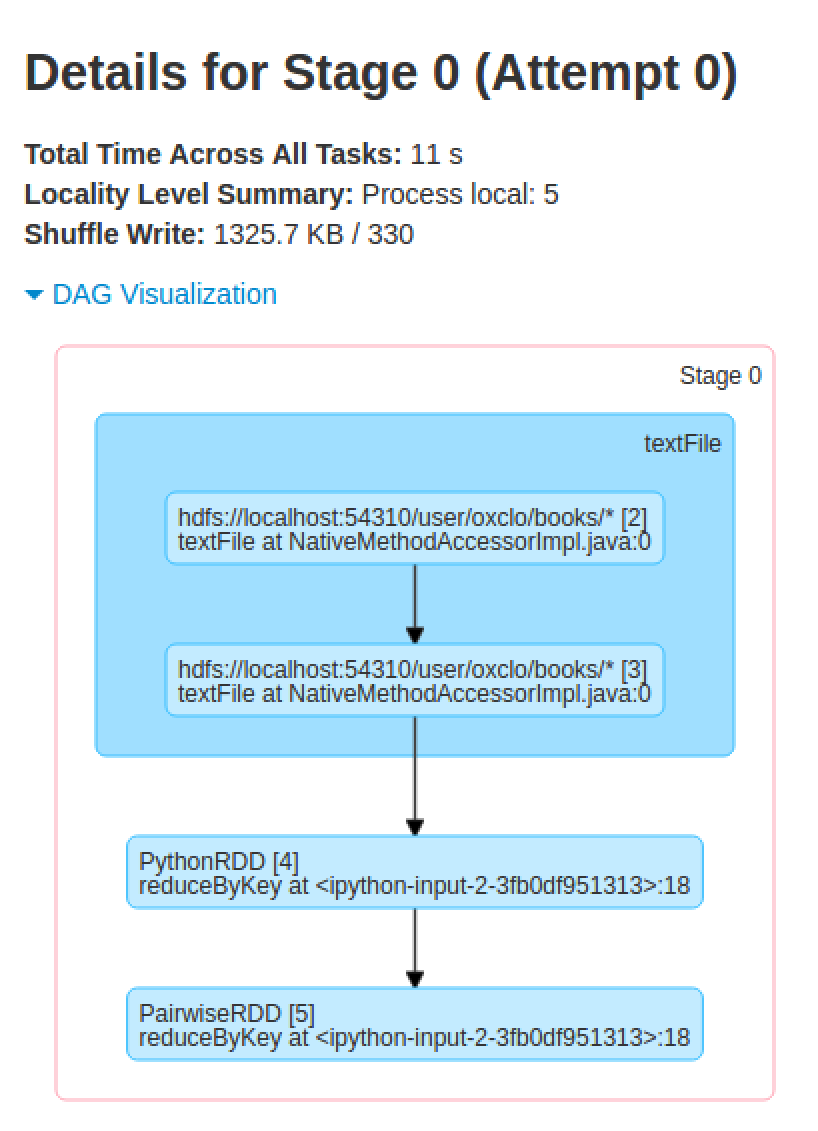
cat ~/BigData/code\_jw/starters/wcnote.py

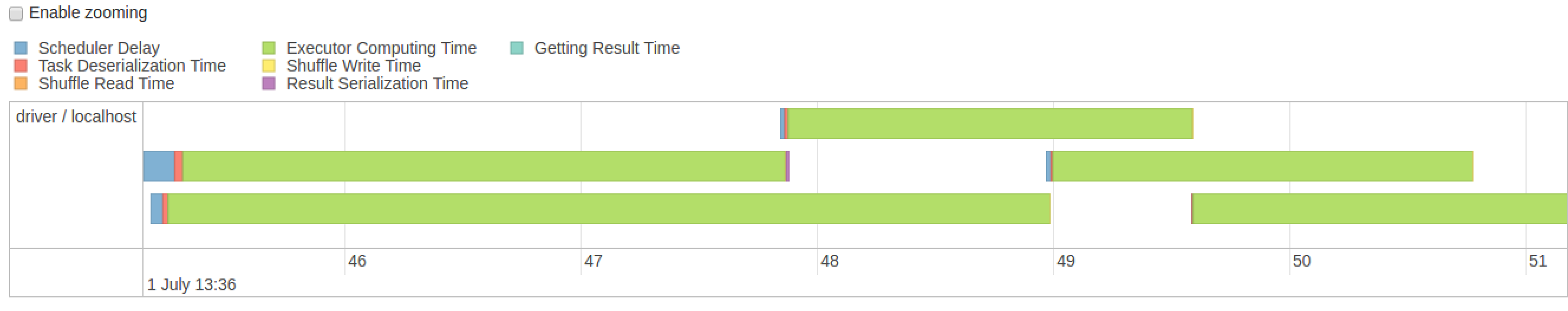
Paste the contents into the cell [1] so it looks like this:



1. There are some aspects that are not filled in that you need to write.   
   Basically, this is a data-processing pipeline (also known as a directed acyclic graph)  
   1. *Let’s look at the parts that are there already.*
   2. We already have a SparkContext object defined in the notebook (in a program you need to define one, which we will see later)
   3. We define a strip function so that we can remove any non-alphanumeric characters:  
        
      def strip(s): return ''.join(filter(str.isalpha, s))
   4. With the preliminaries over, the next line loads the data in:  
      books = sc.textFile("\*.txt")
   5. Then splits the lines into separate words  
        
      tokens = books.flatMap(lambda line: line.split())
   6. Removes non-alpha characters by mapping the strip function onto each token in tokens  
      stripped = tokens.map(strip)  
        
      and removes empty items:  
      notempty = stripped.filter(lambda w: len(w)>0)
2. Now it is time for you to do something!  
     
   Convert all the words to lower case, using a map operation. In python, if *str* is a string, then str.lower() is the same string in lower case.
3. Now you need to get ready for a reduce. In order to do a reduce, we need some form of *key, value* pairs. I recommend using *tuples* which are simply (k,v) in Python (the brackets group the items into a tuple). Remembering how reduce works, we need to map each word to a count. Before reducing, that count is 1. So, we need a lambda that takes a word w and returns (w,1)
4. Now we can do a reduce that adds all those counts together. So that counts accumulated for each key, you can use the .reduceByKey() method.
5. Finally, we need to collect the results and print them. In Spark, they may be distributed across different RDD partitions on different machines, so the collect() method brings them together.  
     
   for k,v in wordcount.collect(): print (k,v)
6. Try running the cell, by clicking 
7. Be patient. I suggest you look at the command window and wait until you see spark start working.
8. You should see a word count appear below cell 1:



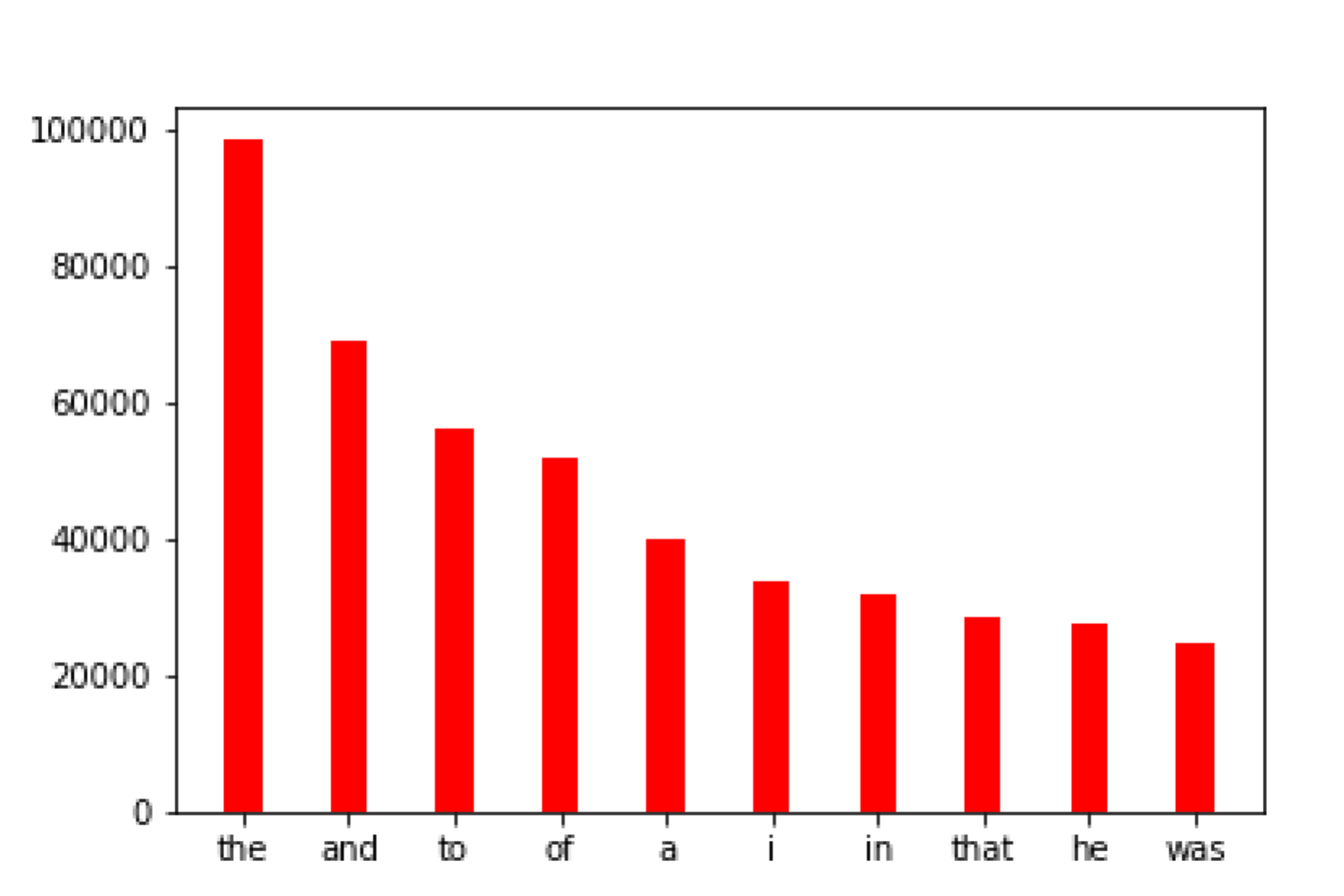
1. While the pyspark is still running browse to <http://localhost:4040>
2. You will see the Spark web console:  
   
3. Click on the blue link “collect at ipython-input”  
   This shows you how Spark converted your code into stages:  
   
4. Click on Stage 0  
   
5. And expand the Event Timeline:



1. Make sure your code is saved from the notebook.
2. Quit the notebook shell by typing Ctrl-C on the command line, and then Y  
   Also close the notebook windows in the browser.
3. Now let’s run the same code as a “job” instead of interactively.
4. Using a text editor, copy and paste that code into a file called bigData/wordcount/wc.py
5. We run jobs locally on a single node directly on Spark:  
   The local[\*] indicates to use as many threads as you have cores on your system:  
   spark-submit --master local[\*] wc.py
6. You can send the output to a file

spark-submit --master local[\*] wc.py > output.txt

However, you will notiice that it is still hidden in all of the logging. Therefore, you will probably want to add a few lines of code to your python program to write the desired output to a file location rather than printing it to stdout.

1. **Congratulations, the lab is complete!**  
     
   **Extensions**
2. Re-load the code into the Jupyter notebook and now improve it to show the wordcount in descending order, starting with the most common words. How many instances of the word ‘the’ are there in the assembled books?
3. Have a look at matplotlib (https://matplotlib.org/users/intro.html)  
   See if you can create the following graph of the top 10 most used words:
4. Adapt your code so that you produce a count of each character rather than of each word. Create a graph of the top 10 most used characters.