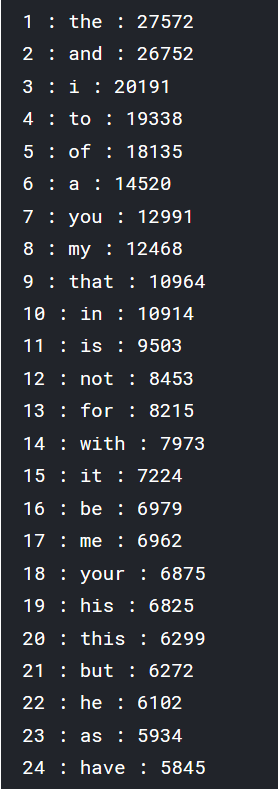
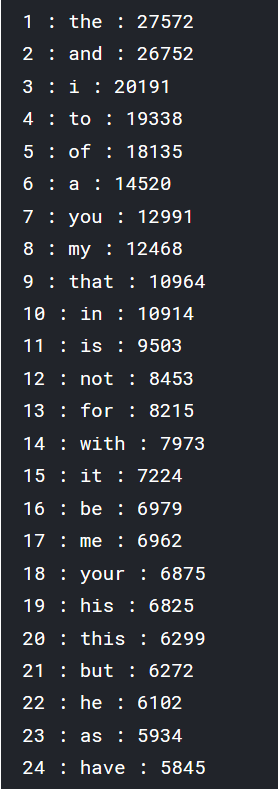
# Task

Load the complete Shakespeare writings, strip the header and search for the #24 most used word in his writings.   
Provide your code in one pdf, txt or by one link!

# The code & the answer

Is public on kaggle: <https://www.kaggle.com/kewagbln/shakespeare-word-count-with-spark-python>



Future improvement: reading the text from line 245, not from 1.

# Historical setup documenation

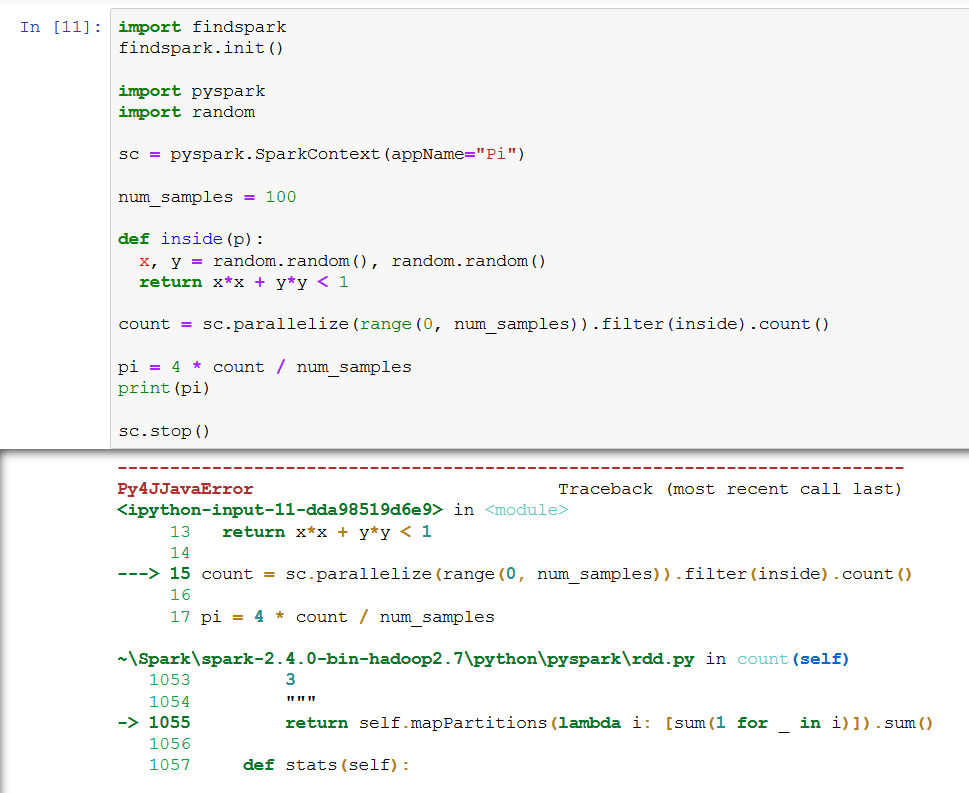
## Setup Spark

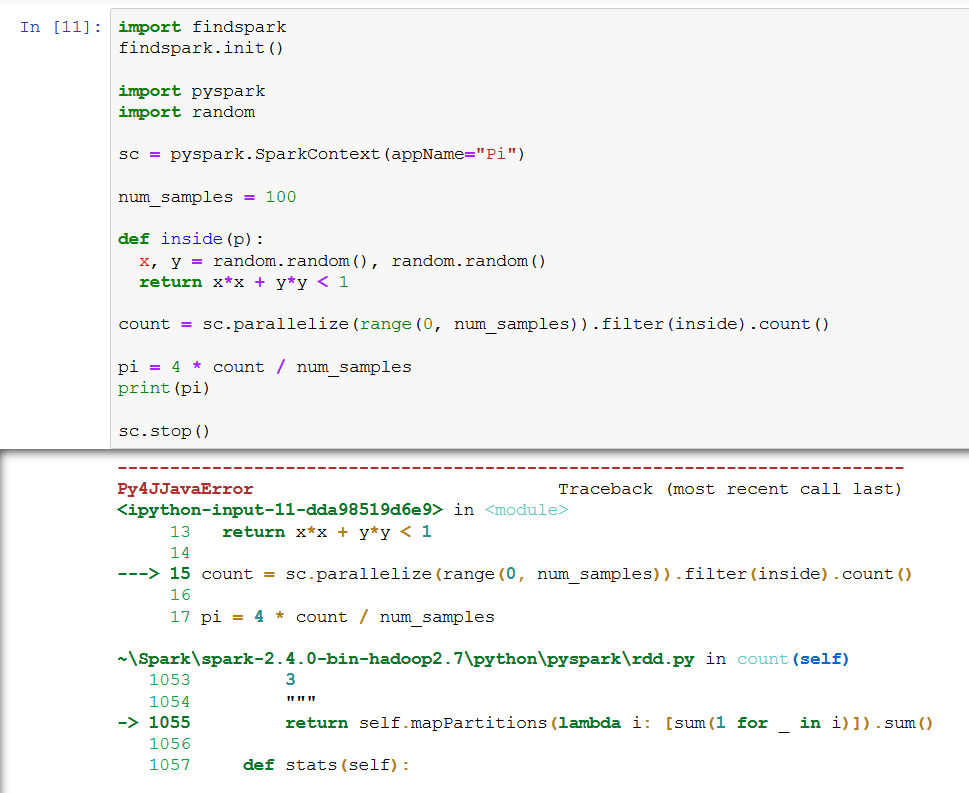
### Try: install on windows

<https://blog.sicara.com/get-started-pyspark-jupyter-guide-tutorial-ae2fe84f594f>

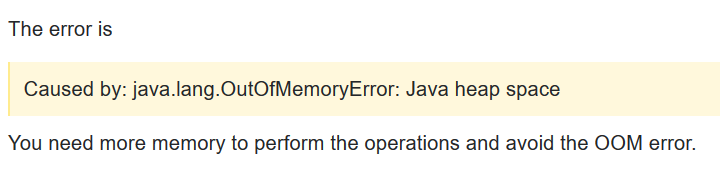
<https://medium.com/@GalarnykMichael/install-spark-on-windows-pyspark-4498a5d8d66c>

Doesn’t work:

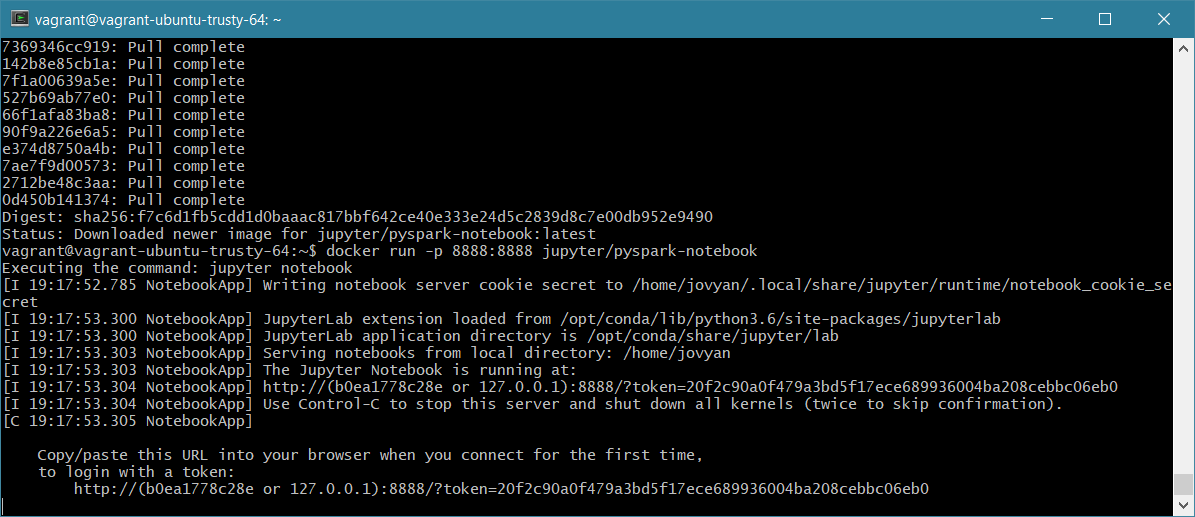
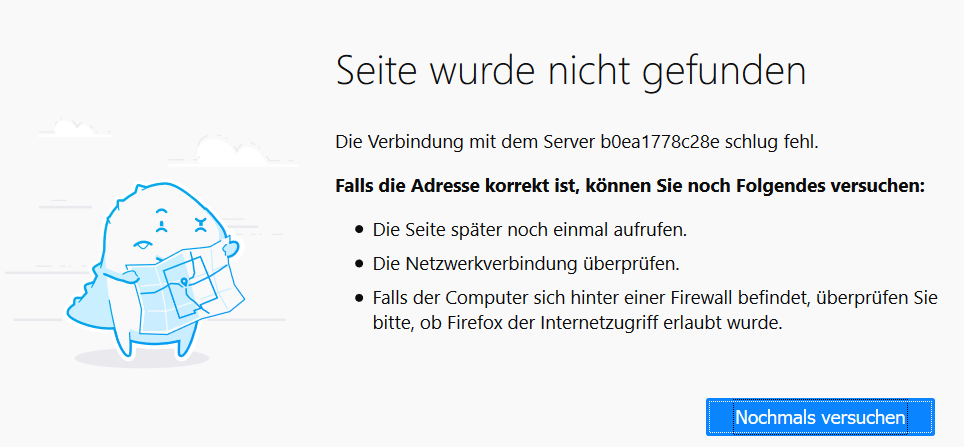




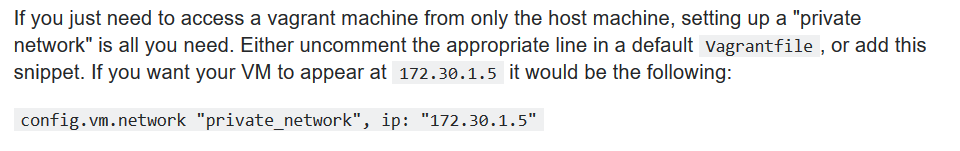
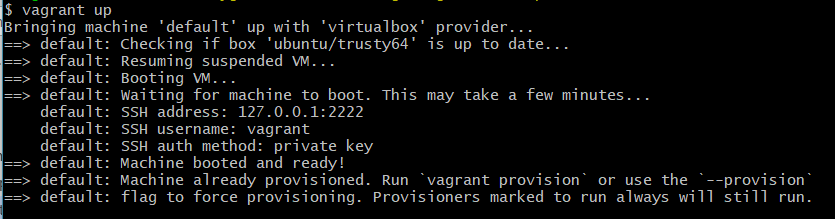
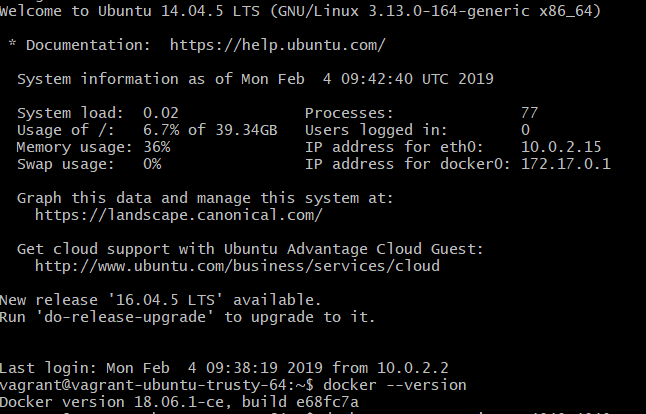
<https://stackoverflow.com/questions/47761758/pyspark-python-issue-py4jjavaerror-an-error-occurred-while-calling-o48-showstr>

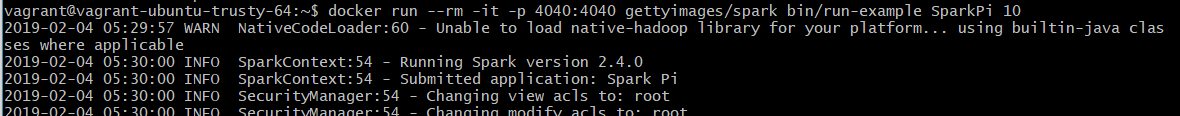


### try: Docker – pyspark

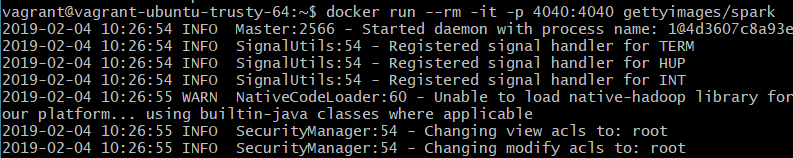
1. Setup a Vagrant VM: vagrant init
2. Setup a docker in the VM: Refresh the vagrantfile / vagrant up / vagrant ssh
3. Get an image: docker pull jupyter/pyspark-notebook  
   <https://hub.docker.com/r/jupyter/pyspark-notebook>   
   More information about the images: <https://jupyter-docker-stacks.readthedocs.io/en/latest/using/selecting.html>
4. Run docker: docker run -p 4040:4040 jupyter/pyspark-notebook  
   Installation takes some time with the first run  
   
5. Open notebook in browser:  
   

### try: Docker – gettyimages/spark

1. Setup a Vagrant VM: vagrant init
2. Setup a docker in the VM: Refresh the vagrantfile  
     
   192.168.3.10  
   <https://stackoverflow.com/questions/14870900/how-to-find-the-vagrant-ip>   
    / vagrant up  
     
   / vagrant ssh  
   
3. Get an image: docker pull gettyimages/spark  
   <https://hub.docker.com/r/gettyimages/spark/>
4. Run an example: docker run --rm -it -p 4040:4040 gettyimages/spark bin/run-example SparkPi 10



Run spark



Run code – doesn’t work:



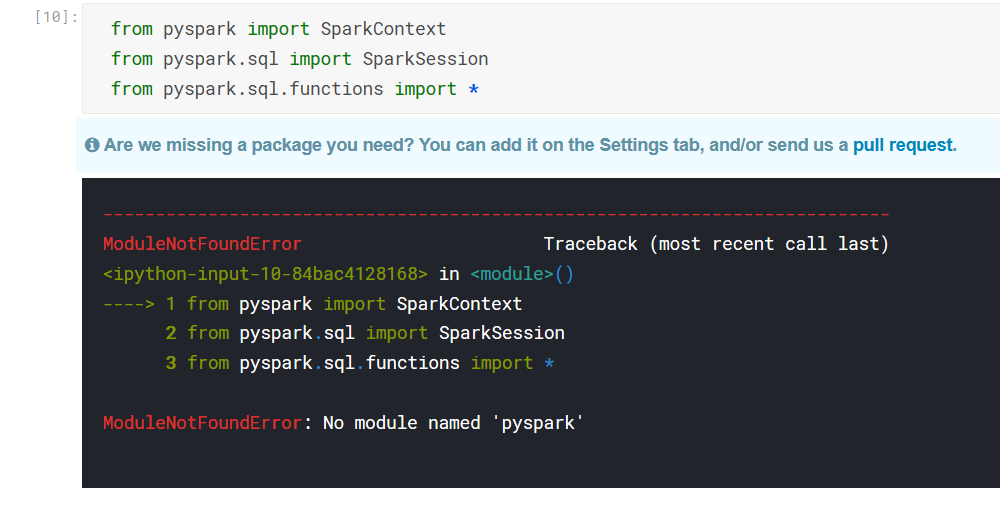
<https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-tips-and-tricks-running-spark-windows.html>

### Try: kaggle

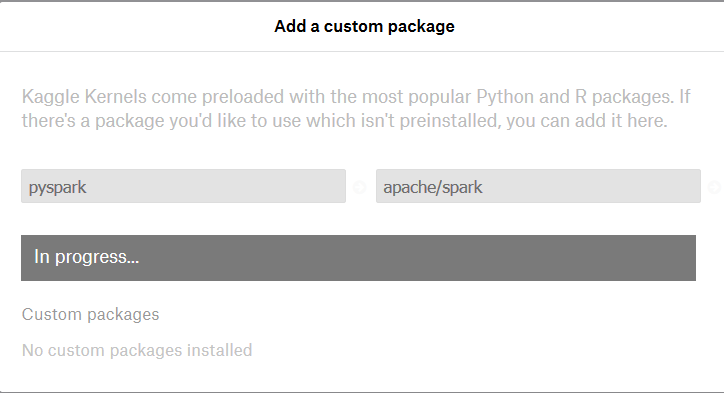
Works in <https://www.kaggle.com/kernels/notebooks/new?forkParentScriptVersionId=5495607>



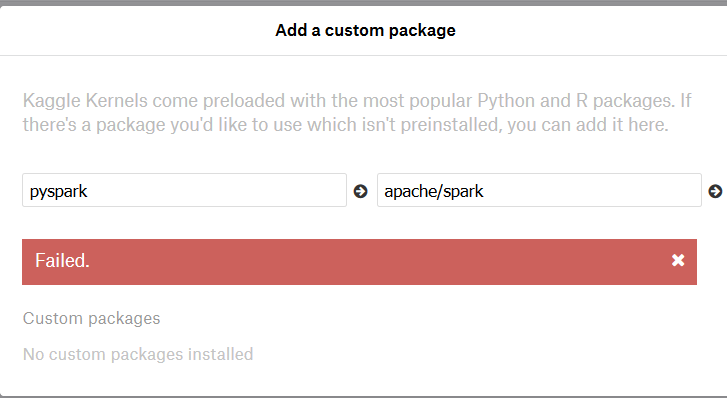
But package missed in my notebook:



Add custom package:



Failed:



Next try – pyspark installation:

