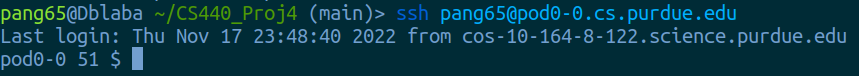
# This tutorial is based on Purdue University pod server.

# PySpark Environment Setup

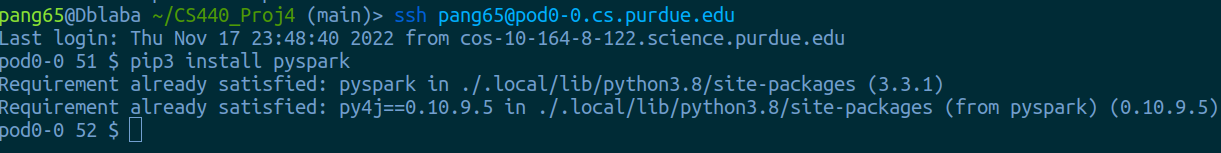
### Use ssh to login purdue pod server

Use command: ssh [username@pod0-0.cs.purdue.edu](mailto:username@pod0-0.cs.purdue.edu)



### Install PySpark lib in pod0-0 server

Command: pip3 install pyspark



### Setup $JavaHome environment variable

echo "export JAVA\_HOME='/usr/lib/jvm/java-1.17.0-openjdk-amd64'" >> ~/.bashrc; source ~/.bashrc



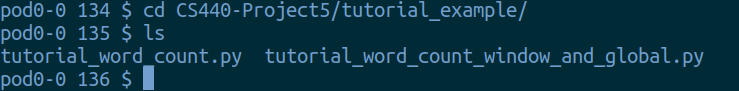
Now, you have setted up the environment successfully. You don’t need to set it up again when you next time login into pod0-0 server.

# Run tutorial examples in pod0-0

### Git clone project5 from github

Command: git clone https://github.com/px1900/CS440-Project5.git

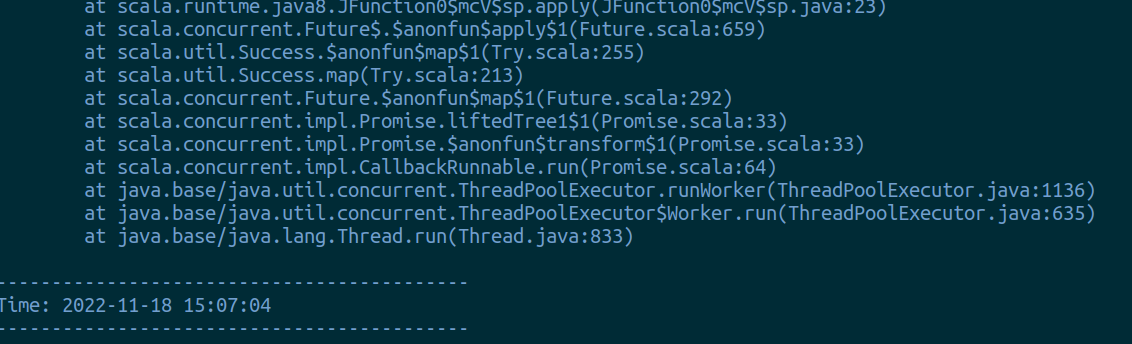
### Get into tutorial example directory



### Run tutorial\_word\_count.py

Command: python3 tutorial\_word\_count.py

It will keep showing error message, just ignore them. The error message is caused by connect error with localhost:9009. It will disappear after we connect with it.



### 

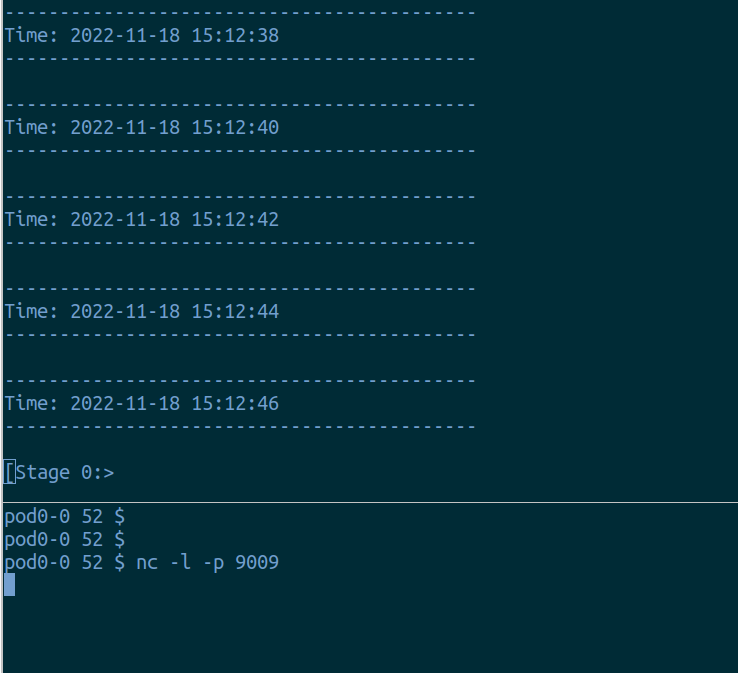
### 

### 

### Send data to spark program

Command: nc -l -p 9009

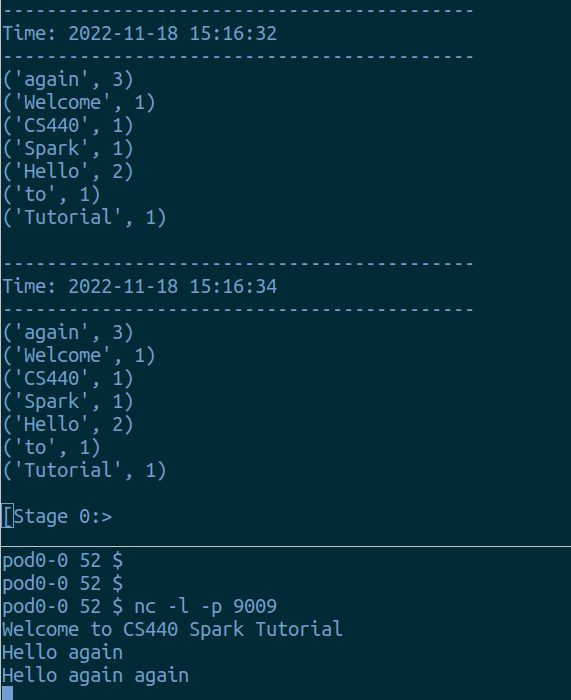
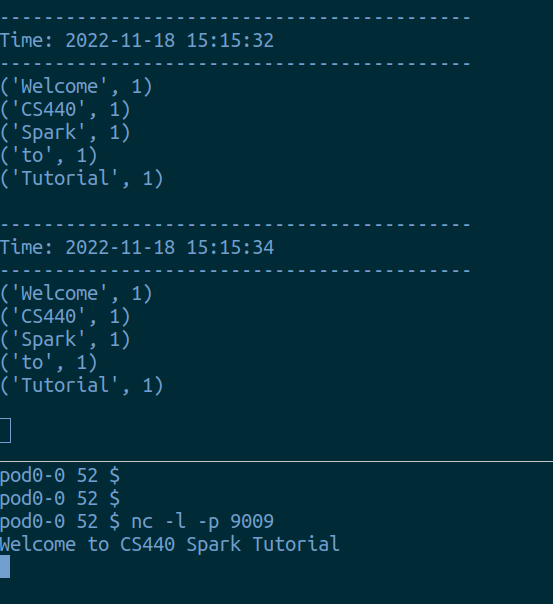
Open a new pop0-0 terminal, and use this command to connect with localhost:9009. One connected, you will find spark connection error message disappeared.



### 

### Counting Word

Type sentence in the second terminal (the nc window), and type ENTER to send this sentence to spark



## Other Sample

You can run another sample “tutorial\_word\_count\_window\_and\_global.py” with the same procedure.

One difference is “tutorial\_word\_count\_window\_and\_global.py” will output not only the global word count, but also the word count inside the current window.

And “tutorial\_word\_count\_window\_and\_global.py” will only count tag word(which include “#”).

