1. For PA, the submission says: the second page (and onwards) should contain the corresponding list of python scripts you used to calculate the answers. Does this mean including screenshots of all our python scripts we wrote to calculate the answers for Iris dataset and Harry Potter?

Solution:

Yes, you need to add your written python scripts.

2. I'm just making sure that words with punctuation are counted as the same word: like "late", "late,", "late.\" would correspond to a word count of 3 for the word "late"?

Solution:

Yes, that’s right.

3. Is it okay that our output for the Harry Potter questions have the key as the word count and the value as the word? Like 3 "late" instead of "late" 3

Solution:

Yes, as long as you can show the correct order of most frequent words.

4. With AWS Educate, how do I find my key\_ID and secret\_access\_key? If I click "Account Details", it shows me an access key Id and secret access key, but when I plug this into mrjob.conf, it throws this error:

An error occurred (InvalidAccessKeyId) when calling the ListBuckets operation: The AWS Access Key Id you provided does not exist in our records.

Solution:

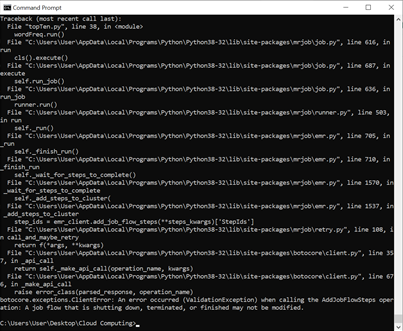
You need to access the key from your created AWS account (not starter account) as mentioned in the PDF.

5. Just wanted to make sure I had the correct idea for the part where it says to find the the difference in average sepal and petal length for all non-“Iris Setosa”.

Did you want us to compute the average for sepal and petal lengths for all non-“Iris Setosa” combined (regardless of their classification) and then take the difference of the two? Or should we compute the average sepal and petal lengths for each non-“Iris Setosa” classification and take the difference for each?

Solution: First case.

6. Getting the following error when trying to complete step 5 of the assignment. Did anyone else get this or have any ideas on how to resolve this? I can't really find much online for this particular error and I've tried running multiple times just to get the same thing.



Solution: Run the command given in the tutorial with an extra parameter as follows:

python file.py -r emr harry.txt --add-steps-in-batch

7. When I run step 5, I made the mrjob.conf file in the same directory as my .py files and then I try to run it like in the example, but then I get this error that says it couldn't find the credentials. Do I have to do anything special to tell it where to locate them even if they're already in the same folder?

Solution:

Try running using the following commands and see whether the credentials are located:  
If you store this file at a different directory, to run your job with -r emr, you need to pass it via --conf-path  
using the command below.  
python your\_mr\_job\_sub\_class.py -r emr --conf-path < input >  
E.g, python step4.py -r emr --conf-path /af1/hs6ms/mrjob.conf harry.txt

Also, check whether the indentations in the mrjob.conf file is correct.

8. I've spent a few hours trying to get MrJob to work and debugging with breakpoints/console output doesn't work either.

I followed the code in the slides and am getting the error "ValueError: could not convert string to float: '' when I try to "yield classification, float(sep\_width).

I also tried line.split(',') as in the slides and it says the index is out of range beyond 0.

Solution:

CHeck whether there is an extra emoty line at the end of iris.data file, if yes , remove it.