Lab Assignment 1: How to Get Yourself Unstuck

DS 6001: Practice and Application of Data Science

Instructions

Please answer the following questions as completely as possible using text, code, and the results of code as needed. Format your answers in a Jupyter notebook. To receive full credit, make sure you address every part of the problem, and make sure your document is formatted in a clean and professional way.

Problem 0

Import the following libraries:

In [62]:
```
import numpy as np
import pandas as pd
import os
import math
```

Problem 1

Python is open-source, and that's beautiful: it means that Python is maintained by a world-wide community of volunteers, that Python develops at the same rate as advancements in science, and that Python is completely free of charge. But one downside of being open-source is that different people design many alternative ways to perform the same task in Python.

Read the following Stack Overflow post: https://stackoverflow.com/questions/11346283/renaming-columns-in-pandas/46912050 (https://stackoverflow.com/questions/11346283/renaming-columns-in-pandas/46912050). The question is simply how to rename the columns of a dataframe using Pandas. Count how many unique different solutions were proposed, and write this number in your lab report. (Hint: the number of solutions is not the number of answers to the posted question.)

Remember: your goal as a data scientist needs to be to process/clean/wrangle/manage data as quickly as possible while still doing it correctly. A big part of that job is knowing how to seek help to find the right answer quickly. Given the number of proposed solutions on this Stack Overflow page, what's the problem with developing a habit of using Google and Stack Overflow as your first source for seeking help? (2 points)

Answer 1

32 proposed solutions (although there is no clear definition of what to count as a unique solution)

Problem 2

There are several functions implemented in Python to calculate a logarithm. Both the `numpy` and `math` libraries have a `log()` function. Your task in this problem is to calculate $\log_3(7)$ directly (without using the change-of-base formula). Note that this particular log has a base of 3, which is unusual. For this problem:

- Write code to display the docstrings for each function.
- Read the docstrings and explain, in words in your lab report, whether it is possible to use each function to calculate $\log_3(7)$ or not. Why did you come to this conclusion?

If possible, use one or both functions to calculate $\log_3(7)$ and display the output. (2 points)

Answer 2

In [63]:
```
1 help(math.log)
```

```
Help on built-in function log in module math:

log(...)
    log(x, [base=math.e])
    Return the logarithm of x to the given base.

    If the base not specified, returns the natural logarithm (base e) of
x.
```

In [64]:
```
1 import numpy as np
2 np.info(np.log)
```

```
log(x, /, out=None, *, where=True, casting='same_kind', order='K', dtype=
None, subok=True[, signature, extobj])

Natural logarithm, element-wise.

The natural logarithm `log` is the inverse of the exponential function,
so that `log(exp(x)) = x`. The natural logarithm is logarithm in base
`e`.

Parameters
----------
x : array_like
    Input value.
out : ndarray, None, or tuple of ndarray and None, optional
    A location into which the result is stored. If provided, it must have
    a shape that the inputs broadcast to. If not provided or None,
    a freshly-allocated array is returned. A tuple (possible only as a
    keyword argument) must have length equal to the number of outputs.
where : array_like, optional
    This condition is broadcast over the input. At locations where the
    condition is True, the `out` array will be set to the ufunc result.
    Elsewhere, the `out` array will retain its original value.
    Note that if an uninitialized `out` array is created via the default
    ``out=None``, locations within it where the condition is False will
    remain uninitialized.
**kwargs
    For other keyword-only arguments, see the
    :ref:`ufunc docs <ufuncs.kwargs>`.

Returns
-------
y : ndarray
    The natural logarithm of `x`, element-wise.
    This is a scalar if `x` is a scalar.

See Also
--------
log10, log2, log1p, emath.log

Notes
-----
Logarithm is a multivalued function: for each `x` there is an infinite
number of `z` such that `exp(z) = x`. The convention is to return the
`z` whose imaginary part lies in `[-pi, pi]`.

For real-valued input data types, `log` always returns real output. For
each value that cannot be expressed as a real number or infinity, it
yields ``nan`` and sets the `invalid` floating point error flag.

For complex-valued input, `log` is a complex analytical function that
has a branch cut `[-inf, 0]` and is continuous from above on it. `log`
handles the floating-point negative zero as an infinitesimal negative
number, conforming to the C99 standard.
```

```
References
----------
.. [1] M. Abramowitz and I.A. Stegun, "Handbook of Mathematical Function
s",
       10th printing, 1964, pp. 67. http://www.math.sfu.ca/~cbm/aands/ (h
ttp://www.math.sfu.ca/~cbm/aands/)
.. [2] Wikipedia, "Logarithm". https://en.wikipedia.org/wiki/Logarithm (h
ttps://en.wikipedia.org/wiki/Logarithm)

Examples
--------
>>> np.log([1, np.e, np.e**2, 0])
array([  0.,   1.,   2., -Inf])
```

I think it is easy to specify the base of the log function when using built in `math.log()` method, not very straightforward with `numpy.log()` though as it does not take base as a parameter. Same problem with `numpy.log2()` and `numpy.log10()` .

In [65]:
```
# Below is `math.log()` demonstration:
math.log(7, 3)
```

Out[65]:
```
1.7712437491614221
```

In [66]:
```
# Check, expected 7:
3**math.log(7, 3)
```

Out[66]:
```
6.999999999999999
```

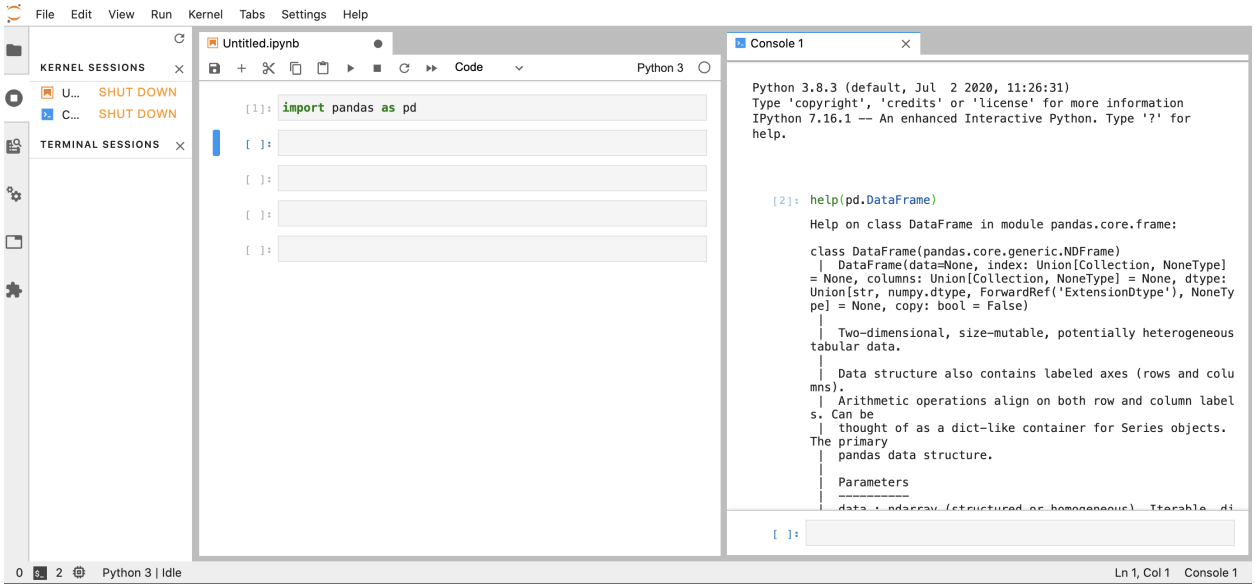
Problem 3

Open a console window and place it next to your notebook in Jupyter labs. Load the kernel from the notebook into the console, then call up the docstring for the `pd.DataFrame` function. Take a screenshot and include it in your lab report. (To include a locally saved image named `screenshot.jpg` , for example, create a Markdown cell and paste

```
<img src="screenshot.jpg" width=600>
```

(2 points)

Answer 3



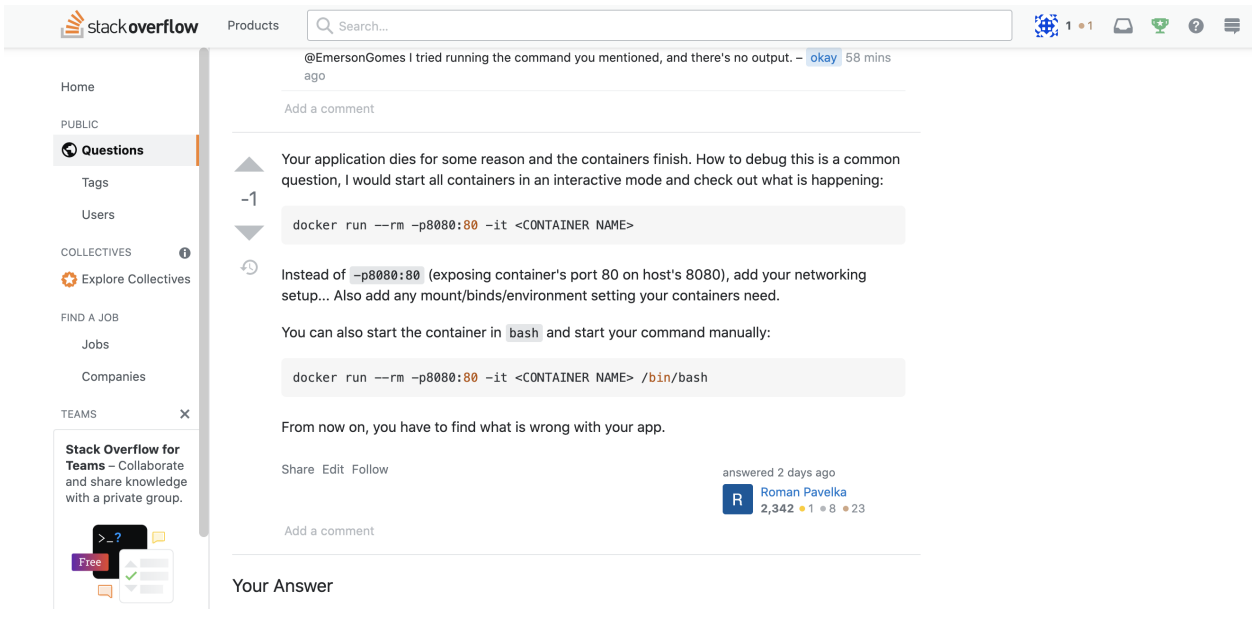
Problem 4

Search through the questions on Stack Overflow tagged as Python questions: https://stackoverflow.com/questions/tagged/python (https://stackoverflow.com/questions/tagged/python). Find a question in which an answerer exhibits passive toxic behavior as defined in this module's notebook. Provide a link, and describe what specific behavior leads you to identify this answer as toxic. (2 points)

Answer 4

link: https://stackoverflow.com/questions/70635266/docker-containers-randomly-shutting-down (https://stackoverflow.com/questions/70635266/docker-containers-randomly-shutting-down)

Somewhat rude comment is shown below



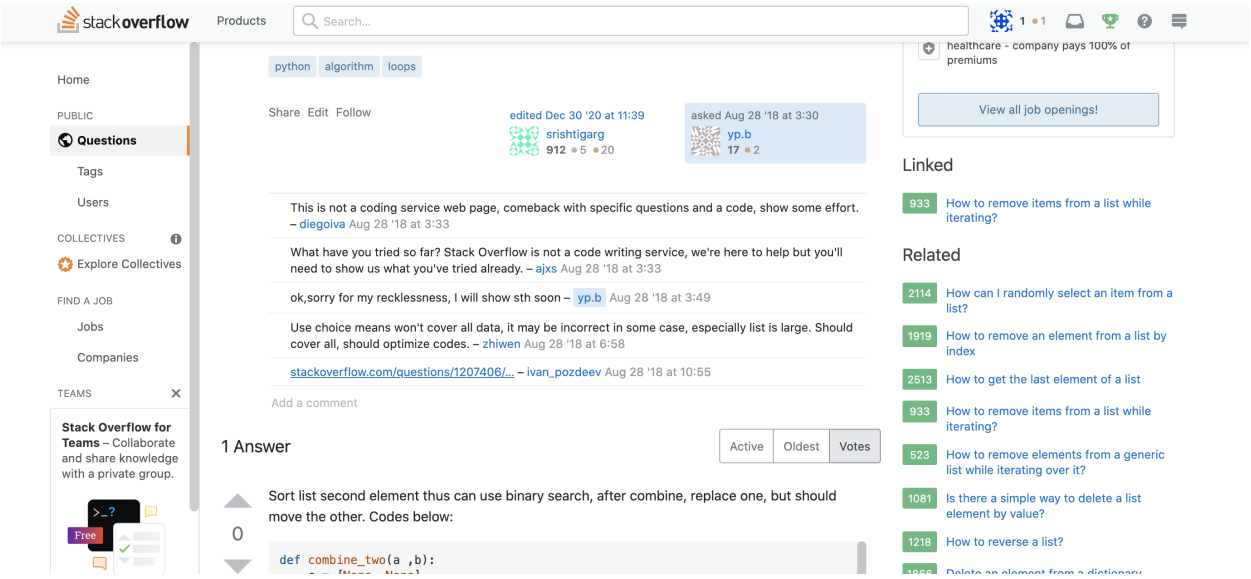
Problem 5

Search through the questions on Stack Overflow tagged as Python questions: https://stackoverflow.com/questions/tagged/python (https://stackoverflow.com/questions/tagged/python). Find a question in which a questioner self-sabotages by asking the question in a way that the community does not appreciate. Provide a link, and describe what the questioner did specifically to annoy the community of answerers. (2 points)

Answer 5

link: https://stackoverflow.com/questions/52049684/how-to-handle-2-element-includ-delete-from-list-in-python-while-iterating-over (https://stackoverflow.com/questions/52049684/how-to-handle-2-element-includ-delete-from-list-in-python-while-iterating-over)

Community complaining that no effort was made by the questioner:



Problem 6

These days there are so many Marvel superheros, but only six superheros count as original Avengers: Hulk, Captain America, Iron Man, Black Widow, Hawkeye, and Thor. I wrote a function, `is_avenger()` , that takes a string as an input. The function looks to see if this string is the name of one of the original six Avengers. If so, it prints that the string is an original Avenger, and if not, it prints that the string is not an original Avenger. Here's the code for the function:

In [67]:
```
def is_avenger(name):
    if name=="Hulk" or "Captain America" or "Iron Man" or "Black Widow"
        print(name  + "'s an original Avenger!")
    else:
        print(name + " is NOT an original Avenger.")
```

To test whether this function is working, I pass the names of some original Avengers to the function:

In [68]:
```
1 is_avenger("Black Widow")
```

```
Black Widow's an original Avenger!
```

In [69]:
```
1 is_avenger("Iron Man")
```

```
Iron Man's an original Avenger!
```

In [70]:
```
1 is_avenger("Hulk")
```

```
Hulk's an original Avenger!
```

Looks good! But next, I pass some other strings to the function

In [71]:
```
1 is_avenger("Spiderman")
```

```
Spiderman's an original Avenger!
```

In [72]:
```
1 is_avenger("Beyonce")
```

```
Beyonce's an original Avenger!
```

Beyonce is a hero, but she was too busy going on tour to be in the Avengers movie. Also, Spiderman definitely was NOT an original Avenger. It turns out that this function will display that any string we write here is an original Avenger, which is incorrect. To fix this function, let's turn to Stack Overflow.

Part a

The first step to solving a problem using Stack Overflow is to do a comprehensive search of available resources to try to solve the problem. There is a post on Stack Overflow that very specifically solves our problem. Do a Google search and find this post. In your lab report, write the link to this Stack Overflow page, and the search terms you entered into Google to find this page.

Then apply the solution on this Stack Overflow page to fix the `is_avenger()` function, and test the function to confirm that it works as we expect. (2 points)

Answer 6

Part a

Google search words: "python if else statement string comparison"

First suggested Stack Overflow page: https://stackoverflow.com/questions/6762959/if-statement-for-strings-in-python (https://stackoverflow.com/questions/6762959/if-statement-for-strings-in-python)

In [73]:
```
def is_avenger_check(name):
    if name in ["Hulk","Captain America","Iron Man","Black Widow","Hawk
        print(name  + "'s an original Avenger!")
    else:
        print(name + " is NOT an original Avenger.")
```

In [74]:
```
is_avenger_check('Thor')
```

```
Thor's an original Avenger!
```

In [75]:
```
is_avenger_check('Beyonce')
```

```
Beyonce is NOT an original Avenger.
```

Part b

Suppose that no Stack Overflow posts yet existed to help us solve this problem. It would be time to consider writing a post ourselves. In your lab report, write a good title for this post. Do NOT copy the title to the posts you found for part a. (Hint: for details on how to write a good title see the slides or https://stackoverflow.com/help/how-to-ask (https://stackoverflow.com/help/how-to-ask)) (3 points)

Answer 6

Part b

Title for a possible post: "Spark TF_IDF how to convert dafarame column to param map or a list/tuple of param maps"

Part c

One characteristic of a Stack Overflow post that is likely to get good responses is a minimal working example. A minimal working example is code with the following properties:

1. It can be executed on anyone's local machine without needing a data file or a hard-to-get package or module
2. It always produces the problematic output
3. It using as few lines of code as possible, and is written in the simplest way to write that code

Write a minimal working example for this problem. (3 points)

In [76]:
```
# I have 2 dictionaries and I want to join their content
d1 = {"a":2, "b":3, "c":5}
d2 = {1:10, 2:20}
```

In [77]:
```
d1d2 = {} # placeholder to join 2 dictionaries
for d in (d1, d2):
    for key, value in d.items():
        d1d2[key].append(value)
```

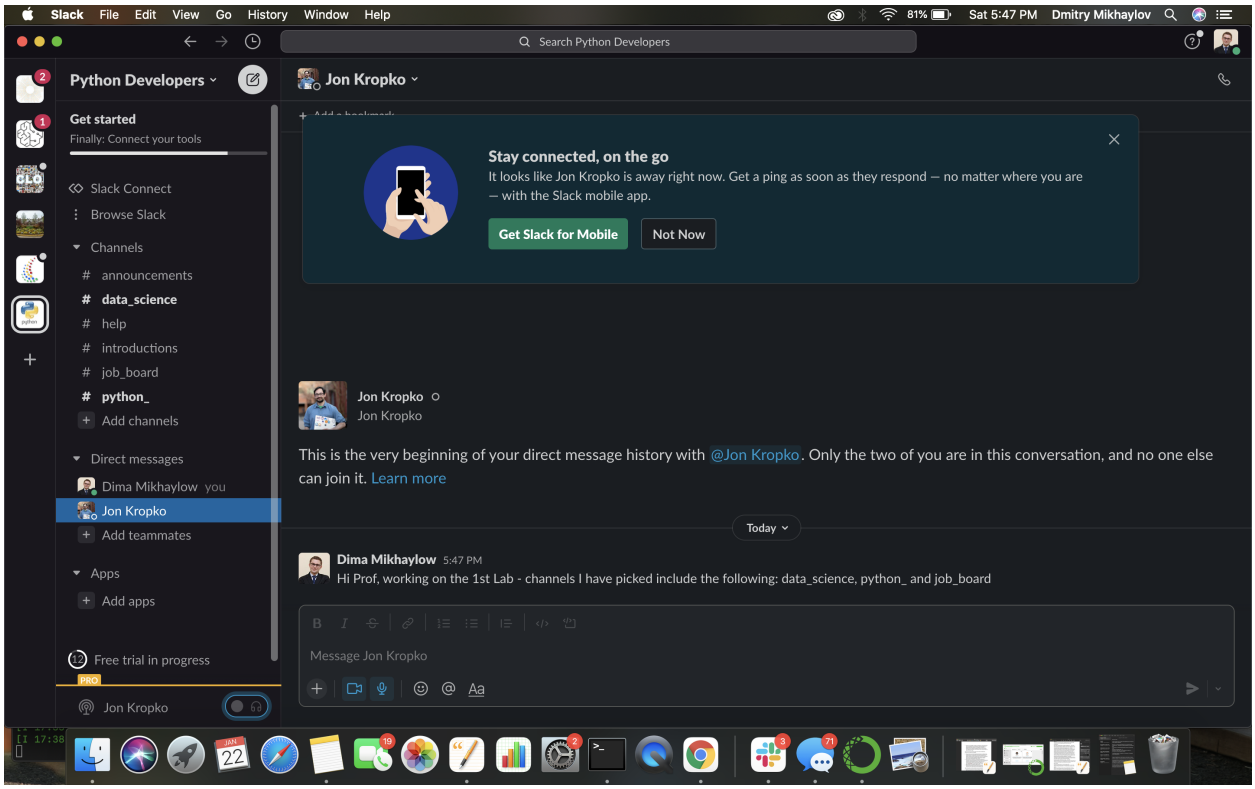
```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-77-aff40388a1ce> in <module>
      2 for d in (d1, d2):
      3     for key, value in d.items():
----> 4         d1d2[key].append(value)

KeyError: 'a'
```

Problem 7

Sign on to the PySlackers slack page and send me a private message in which you tell me which three channels on that Slack workspace look most interesting to you. (2 points)

Answer 7



In [ ]:
```
1
```