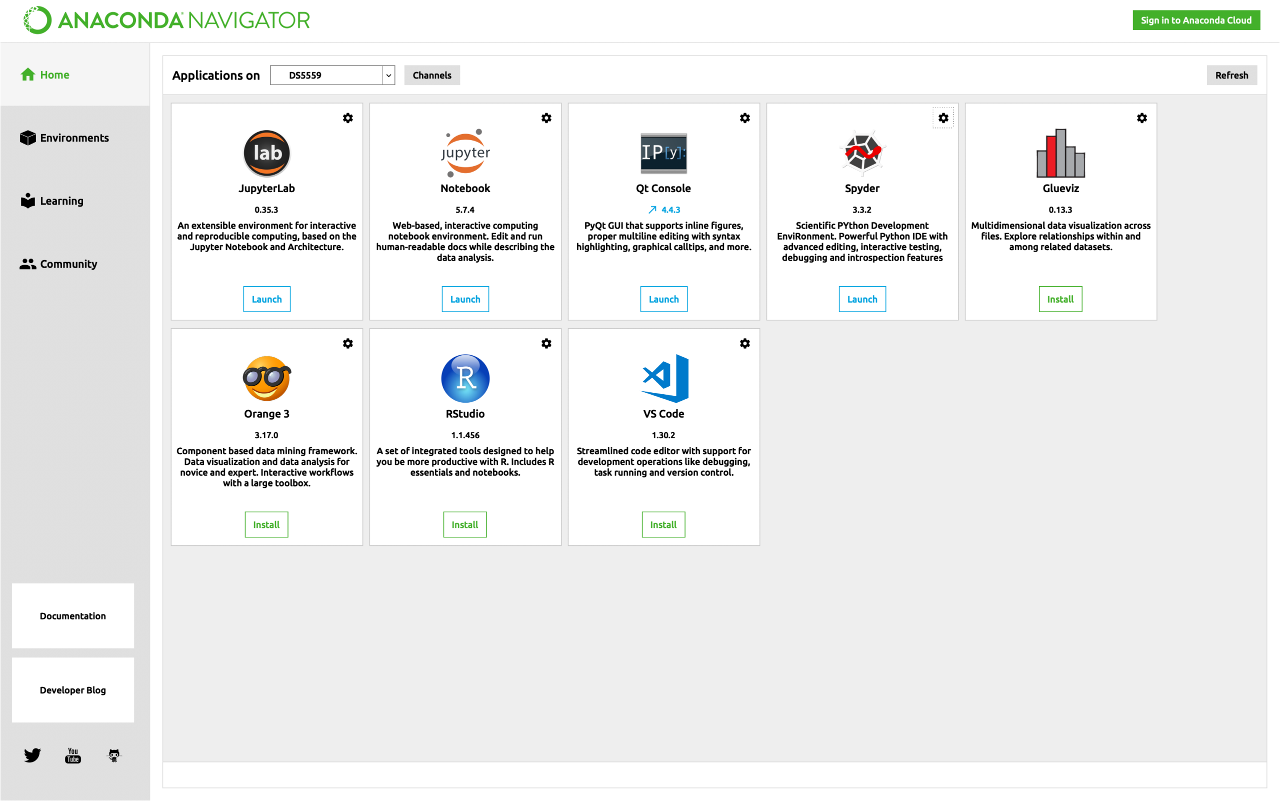
2019-01-17 Lab 1

Contents

Exercise 1: Install Anaconda

- ☐ Go to the Anaconda download page, find the distribution for your operating system, and download it.
- ☐ Install it once it is downloaded.
- ☐ Find the application icon in your toolbar or program manager and open it up – you should see something like this:



An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Streamlined code editor with support for development operations like debugging, task running and version control.

Exercise 2: Install Spyder and JupyterLabs

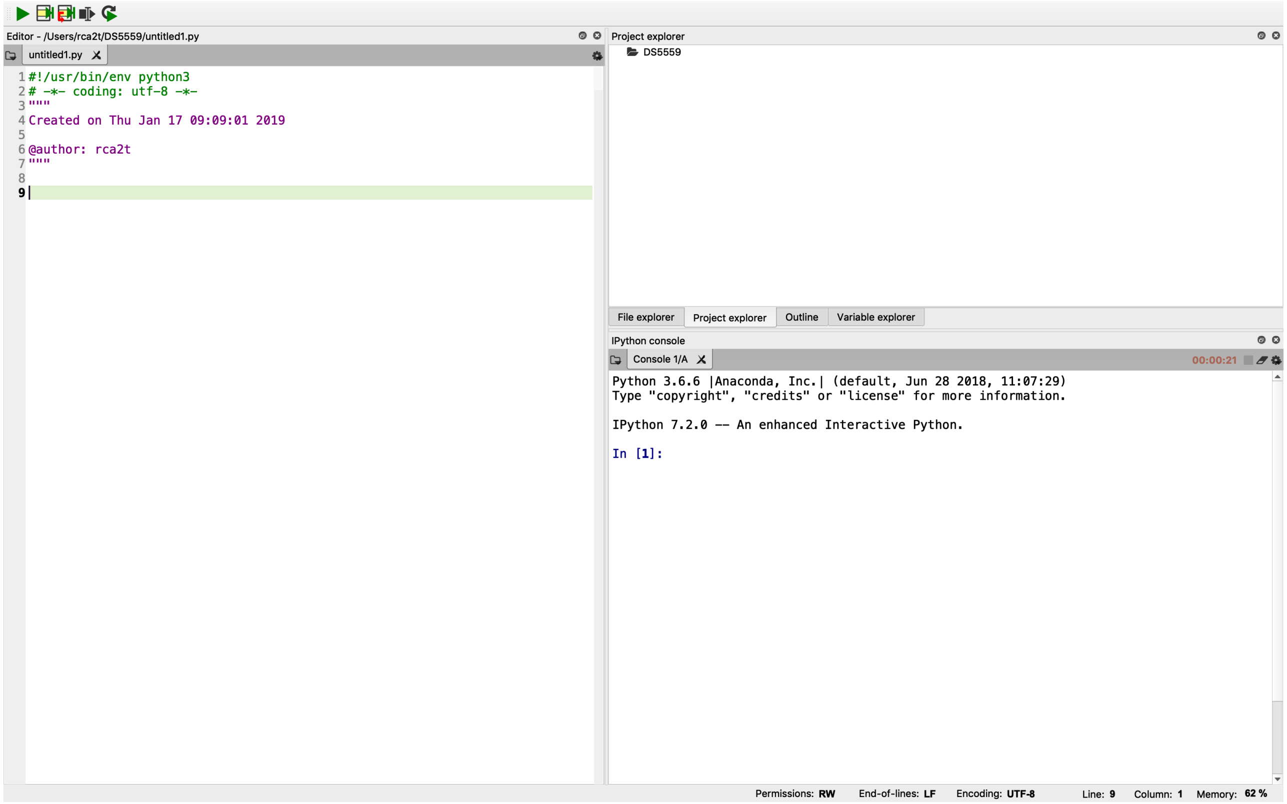
- ☐ In Anaconda Navigator, locate the cards for Jupyterlab and Spyder and click the Install button for each. You may also install Jupyter Notebook if you'd like.
- ☐ Once installed, click the Launch button to run it.
- ☐ Spyder should appear.
- ☐ Optionally configure the panels as instructed.

Exercise 3: Create a New Project in Spyder for this Course

□ In the menu, select **Projects New Project**.

□ Select **New Directory** and create a project called **DS5559**.

□ You should see something like this:



```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Thu Jan 17 09:09:01 2019

@author: rca2t
"""

```

```
Python 3.6.6 |Anaconda, Inc.| (default, Jun 28 2018, 11:07:29)
Type "copyright", "credits" or "license" for more information.

IPython 7.2.0 -- An enhanced Interactive Python.

In [1]:
```

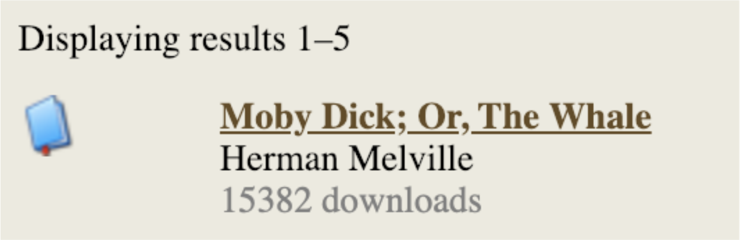
Permissions: **RW** End-of-lines: **LF** Encoding: **UTF-8** Line: **9** Column: **1** Memory: **62 %**

Exercise 4: Download and Open *Moby Dick* from Project Gutenberg

□ Go to Project Gutenberg.

□ Entry "Moby Dick" in the search box on the left.

□ Click on the first result –



Displaying results 1–5

Moby Dick; Or, The Whale
Herman Melville
15382 downloads

□ Click on the link for the Plaintext UTF-8 version and save the file to your computer. Save the file to your project directory or move the file there once it is downloaded.

□ From within Spyder, open the file.

□ Investigate the file along with the class.

- □ Is the file complete?
- □ How many lines does the file have?
- □ What kind of text encoding does it have?
- □ What kind of line endings does it have?
- □

- □ Does the file come with metadata?
- □ What content do we want to keep?
- □ Where does the cruft begin and end?

Exercise 5: Open File in Python and Convert to Text From 1

- □ Create a new Python file called `moby.py`.
- □ Follow instructor for specific commands.

Code

```
# -*- coding: utf-8 -*-

import re
import pandas as pd


# Identify the source text (F0)
src_file = '2701-0.txt'

# Import the text as list of lines
lines = open(src_file, 'r', encoding='utf-8').readlines()

# Trim the cruft we identified
lines = lines[340:21964]

# Convert the lines into one big line, preserving line breaks
bigline = ''.join(lines)

# Split the bigline into paragraphs
paras = re.split(r'\n\n+', bigline)

# Break line into paragraphs

# Split by non-character
paras2 = []
for para in paras:
    tokens = re.split(r'\W+', para)
    paras2.append(tokens)


# Split by non-character but keep them
paras3 = []
for para in paras:
    tokens = re.split(r'(\W+)', para)
    paras3.append(tokens)

# Split by non-character using list comprehension
```

```
paras4 = [re.split(r'\W+', para) for para in paras]

# Try in Pandas

# Import paragraphs into a data frame
df = pd.DataFrame(paras, columns=['line'])
df.index.name = 'line_id'
df.line = df.line.str.strip()

# Tokenize using this one trick
df2 = df.line.str.split(r'\W+', expand=True)\
    .stack()\
    .to_frame()\
    .rename(columns={0: 'token'})
df2.index.names = ['line_id', 'token_id']

# Do a simple normalization
df2['norm'] = df2.token.str.lower()

# Get top N tokens
N = 30
df2['norm'].value_counts().head(N).sort_values().plot(kind='barh')
```

```
# Visualize dispersion plots of 'ahab' and 'whale'
(df2['norm'] == 'whale').astype('int').plot(figsize=(10, 1))
(df2['norm'] == 'ahab').astype('int').plot(figsize=(10, 1))
```