MODULE 4.2 Performing Text Analysis: Basic Approaches with Python

KEY TOOLS/PLATFORMS

Python
A programming language that is good for working with data.

pip
A package manager for Python

pyplot
Visualization function in the Python data science package, Pandas

HTRC Extracted Features
A downloadable dataset of text data and metadata extracted and abstracted from volumes in the HathiTrust Digital Library.

HTRC Feature Reader
A Python library for working with HTRC Extracted Features.

ACTIVITY: Identify the method

☞ ***Slide M4.2 - 6***

What are the broad areas and methods used for the research examples we read earlier?

Project summaries: http://go.illinois.edu/ddrf-research-examples

	Broad area	Specific method
Rowling and "Galbraith": an authorial analysis	Natural Language Processing	Stylometry
Significant Themes in 19th Century Literature	Machine Learning	Topic modeling
The Emergence of Literary Diction	Machine Learning	Naive Bayes Classification

ACTIVITY: View adjectives in Extracted Features files

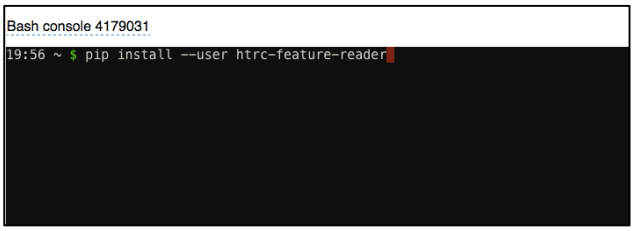
☞ ***Slide M4.2 - 26***

1. Go to your PythonAnywhere dashboard and click on the "Browse files" button to check if all files and directories are in place 

   - On your "Files" page, you will see your directories on the left and your files on the right.
   - You should have two directories "1930/" and "1970/" listed on the left.

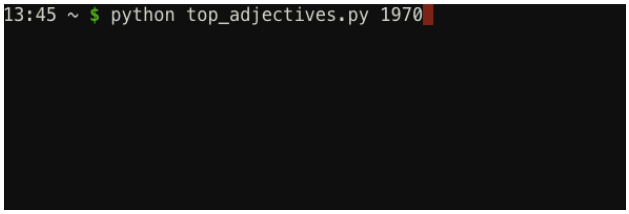
- Click on each directory.
- There should be 16 json.bz2 files in the “1970/” directory that correspond to the volumes in the poli_science_DDRF workset atn 5 json.bz2 files in the “1930/” directory of Presidential Papers volumes from the 1930s.

2. Open the Bash console.
3. Install the Feature Reader Library by typing:
   `pip install --user htrc-feature-reader`
4. Hit enter.

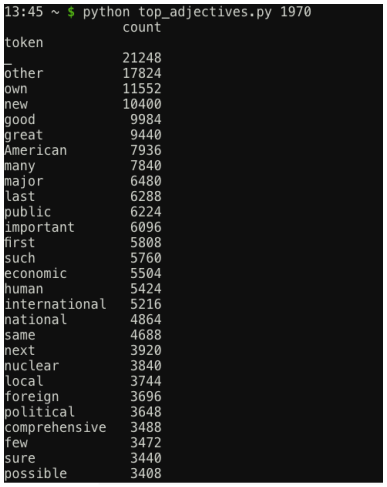


5. From the Bash shell, run the script. Enter command (remember to use the tab key to help you automatically complete the file names):

   `python top_adjectives.py 1970`



6. Hit Enter. The results will be printed out directly in the window.



ACTIVITY: Now you try!

☞ ***Slide M4.2 - 32***

- Compare the adjectives used by presidents in the 1970s with those used in the 1930s.
- Work and discuss with your neighbor.
  - *How do you need to change your command?* **python top_adjectives.py 1930**
  - *What differences do you see? Similarities?* **1930s: industrial 1970s: nuclear, international much higher on the list**

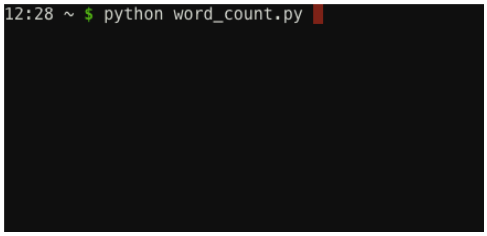
Challenge	Try modifying the script to search verbs or another part of speech. (Hint: Don’t forget the Penn Tree Bank – you can find it here: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

Open the file, and replace JJ with another code from the Penn Tree Bank, such as VB for verb.

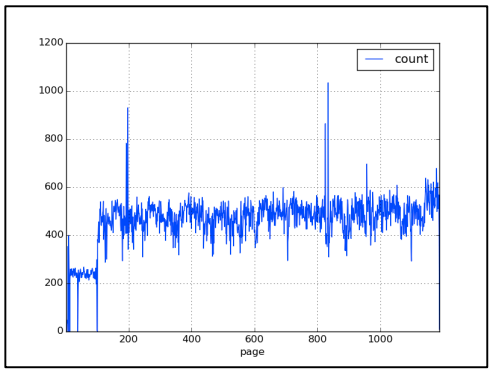
ACTIVITY: Visualize word trends

☞ *Slide M4.2 - 36*

1. Open the Bash console
2. Enter the command “python word_count.py”
3. Click the “PythonAnywhere” logo to go back to the home screen and go to your Files.



4. The “words.png” file should be in your Files. Click on the download icon to the right of the file name to open the file.



ACTIVITY: Now you try!

☞ *Slide M4.2 - 40*

- *Can you modify the script to look at another volume?*

  In line 10 on the script, change the path to another volume:
  Example 1: path = glob.glob('1970/ mdp.49015002221761.json.bz2')
  Example 2: path =glob.glob('1930/miua.4925052,1928,001.json.bz2')