

Course: INT 93S
Quarter: Summer 2017
Assigned: 27 June 2017
Due: 9:00 29 June 2017

Regular Expressions Lab

The goal of this lab is to learn how to import a file, use regular expressions to extract information and output a file.

Using test code to write a regular expression that collects all the dates, extracting the day, month and year. If needed, you can use multiple regular expressions. Each part is broken down into multiple steps. Make sure your code works for the previous steps and back it up before moving onto the next. There are also strings at the end of the file that you should avoid detecting.

Part 1:

1. Copy and rename the test code to "lastname_firstname_lab1part1.py"
2. Use test code to import file input1_step1.txt
3. For each date append to the output list with the dates reformatted to the applicable format below dependent on the information in the date provided:
 - a. dd/mm/yyyy
 - b. dd/mm
 - c. mm/yyyy
 - d. yyyy
4. Use test code to write each item on the list on a separate line in a file titled "lastname_firstname_lab1_part1.txt"
5. Update the code to handle steps 2-4

Part 2:

1. Copy and rename the test code to "lastname_firstname_lab1part2.py"
2. Use test code to import file input2_step1.txt
3. For each date append to the output list with the dates reformatted to:
(use CE for Common Era and BCE for Before Common Era)
 - a. If day, month and year is provided
 - i. 10 July 1992 CE
 - ii. 10 July 1992 to 21 July 1992 CE
 - b. If month and year is provided
 - i. July 1992 CE
 - ii. July 1992 to August 1992 CE
 - c. If only year is provided
 - i. 1992 CE
 - ii. 1992 to 1993 CE

4. Use test code to write each item on the list on a separate line in a file titled "lastname_firstname_lab1_part2.txt"
5. Update the code to handle steps 2 and 3

Part 3: (Challenge)

1. Find a wikipedia page of a country or sport's history
2. Get the content from the page and copy it into a txt file
You can get the simplified content by replacing page title in the query below:
https://en.wikipedia.org/w/api.php?format=xml&action=query&titles=History_of_France&prop=revisions&rvprop=content
3. In a new python script titled "lastname_firstname_history.py" import the txt file
4. Use regular expressions to extract sentences with dates (ignore references) and the dates
5. Order dates in chronological order
6. Write to "lastname_firstname_timeline.txt" file that for each date is re-formatted to the format described in Part 2, followed by the sentence in quotes, per line

Turn-in

1. Create a readme file titled "lastname_firstname_readme.txt" which includes:
 - a. First name and last name of you and your partner
 - b. How much of the lab you finished in class
 - c. Any references you used
 - d. Anything you would like the professor and TA to know.
2. Put the input file, code and output file for Part 1, 2 and 3 in a folder named "lastname_firstname_lab1"
3. Zip the folder and upload it to Gauchospace