NATURAL LANGUAGE PROCESSING CSCE-5290

Assignment 1(b) Problem 3 Name: Zakir Shaik

Unt Id: 11145367

I have done this in Microsoft Visual studio 2017. I have created a separate project and copied date\_tokenizer.py code into it. Regs.txt and input2.txt files are present in date\_tokenizer/date\_tokenizer/ path. I took the help of following links for date formats and holidays.

SOURCES:

1. <https://en.wikipedia.org/wiki/Public_holidays_in_the_United_States>
2. <https://www.w3.org/QA/Tips/iso-date>
3. <https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Dates_and_numbers>
4. <https://www.redcort.com/us-federal-bank-holidays/>

In order to test the regular expressions I copied some of the data from source 3 into input2.txt and tested on it. I tried to make expressions as general as possible. While testing I had to make the regular expressions up and down in the regs.txt as more generalized regex expressions should be at bottom.

I have written regular expressions for date ranges as source 3 contains various formats of dates along with date ranges. Also some of the dates in input2.txt are not printed as the algorithm checks only one regular expression for line and breaks if anything matches. I modified some of the sentences which has two or three dates in a single line and made it as three lines to print them. I read some of the news paper articles but all of them match with the dates in source 3.

I am adding the list of examples and sources to regs.txt file for each regular expression. I have used some of the data in SOURCE 3 as my test cases. I have copied it in input2.txt.

Also the regular expressions are written with case insensitivity using (?i).

The following is the final output generated by the regular expressions.





