Masoud Nateghi – 10/16/2023

HW8 - BMI500

I selected the date category for removing protected health information (PHI) from the data. To accomplish this, we employed regular expressions (regex) targeting the following formats:

To identify dates in the format DD/MM/YYYY, with the stipulation that the year falls within the range of 1900 to 2023:

 $\label{eq:continuous} $$ \frac{1,2}{\langle 1,2 \rangle (19[0-9]\{2\}|200[0-9]|201[0-9]|202[0-2])|(0[1-9]|[12][0-9]|3[01])}{\langle 0[1-9]|1[1,2]\rangle (19[0-9]\{2\}|200[0-9]|201[0-9]|202[0-2])|^(3[01]|[12][0-9]|0?[1-9])(/|-)(1[0-2]|0?[1-9])(2(19[0-9]\{2\}|200[0-9]|201[0-9]|202[0-2])$$

To extract years:

\b(?:19[0-9][0-9]|20[0-1][0-9]|2022)\b

To extract months:

(January|Jan|February|Feb|March|Mar|April|Apr|May|June|Jun|July|Jul|August|Aug|September |Sep|October|Oct|November|Nov|December|Dec)|(january|jan|february|feb|march|mar|april|apr|may|june|jun|july|jul|august|aug|september|sep|october|oct|november|nov|december|dec)

The results obtained using these regular expressions were as follows:

Examining "DateYear" category.	Examining "Date" category.
Num of true positives = 2	Num of true positives = 51
Num of false positives = 1188	Num of false positives = 1156
Num of false negatives = 44	Num of false negatives = 431
Sensitivity/Recall = 0.043	Sensitivity/Recall = 0.106
PPV/Specificity = 0.453	PPV/Specificity = 0.468

Profiling:

As evident from the profiling results, a significant portion of the runtime was consumed by the check_for_date and deid_date functions, which extensively utilized re.finditer and re.findall, known for being time-consuming. Additionally, the print statements accounted for a noticeable fraction of the overall runtime.

```
Ordered by: internal time
   ncalls tottime percall
                                   cumtime percall filename:lineno(function)
               2.374
     2434
                           0.001
                                      3.970
                                                 0.002 deid_Nateghi.py:16(check_for_date)
     4344
               1.456
                           0.000
                                      1.456
                                                 0.000 {built-in method builtins.print}
                                                 4.794 deid_Nateghi.py:59(deid_date)
0.000 {method 'findall' of '_sre.SRE_Pattern' objects}
                          0.296
                                      4.794
               0.296
         1
                          0.000
    70358
               0.274
                                      0.274
                          0.000
                                      0.110
                                                 0.000 re.py:286(_compile)
    70358
               0.108
    70358
               0.103
                          0.000
                                      0.486
                                                 0.000 re.py:214(findall)
                                                 0.000 {method 'write' of ' io.TextIOWrapper' objects}
     4606
               0.067
                           0.000
                                      0.104
                                                    0.000 {built-in method builtins.len}
70415/70413
                           0.000
                                        0.021
                 0.020
                                                 0.000 cp1256.py:18(encode)
     4606
               0.019
                           0.000
                                      0.034
                                                 0.000 {built-in method _codecs.charmap_decode}
0.000 {built-in method _codecs.charmap_encode}
0.000 {method 'finditer' of '_sre.SRE_Pattern' objects}
0.000 {method 'format' of 'str' objects}
               0.015
                           0.000
                                      0.015
       269
     4606
               0.015
                           0.000
                                      0.015
     2434
               0.010
                           0.000
                                      0.010
     2434
               0.008
                          0.000
                                      0.008
                                                         {method 'start' of ' sre.SRE Match' objects}
     6516
               0.007
                           0.000
                                      0.007
                                                 0.000
                                                        {method 'start of _sre.sht_hatch object
{method 'lower' of 'str' objects}
{method 'end' of '_sre.SRE_Match' objects}
                          0.000
                                     0.006
                                                 0.000
               0.006
     2434
     2434
               0.004
                          0.000
                                      0.004
                                                 0.000
     4344
               0.003
                           0.000
                                      0.003
                                                 0.000
                                                 0.000 codecs.py:275(reset)
0.000 {method 'group' of '_sre.SRE_Match' objects}
     4606
               0.003
                          0.000
                                      0.003
                                      0.002
               0.002
                           0.000
     2172
               0.001
                                      0.016
                                                 0.000 cp1256.py:22(decode)
       269
                          0.000
                                                 0.000 {built-in method io.open}
               0.001
                           0.000
                                      0.001
                                                 0.000 enum.py:803(__and_
0.000 enum.py:517(__new_
       103
               0.000
                           0.000
                                      0.001
       222
               0.000
                          0.000
                                      0.000
       6/2
               0.000
                           0.000
                                      0.001
                                                 0.001 sre_compile.py:64(_compile)
                          0.000
                                      0.001
                                                 0.000 enum.py:267(__call__)
       222
               0.000
       4/2
               0.000
                           0.000
                                      0.001
                                                 0.000 sre_parse.py:470(_parse)
                                                 0.000 sre_parse.py:173(getwidth)
0.000 {built-in method builtins.isinstance}
     10/6
               0.000
                           0.000
                                      0.000
       138
               0.000
                           0.000
                                      0.000
        54
               0.000
                           0.000
                                      0.000
                                                 0.000 sre_parse.py:232(__next)
               0.000
                                                 0.000 sre_parse.py:342(_escape)
        14
                           0.000
                                      0.000
                                                 0.000 sre_parse.py:171(append)
0.000 {method 'append' of 'list' objects}
        48
               0.000
                           0.000
                                      0.000
               0.000
                           0.000
       191
                                      0.000
        50
               0.000
                           0.000
                                      0.000
                                                 0.000 sre_parse.py:253(get)
         1
               0.000
                           0.000
                                      0.000
                                                 0.000 enum.py:767( create pseudo member )
                                                 0.000 {built-in method _sre.getlower}
               0.000
                          0.000
                                      0.000
        41
                                                 0.001 sre_compile.py:557(compile)
0.000 enum.py:797(_or__)
0.000 sre_parse.py:844(parse)
                           0.000
               0.000
                                      0.002
         8
               0.000
                           0.000
                                      0.000
               0.000
                           0.000
                                      0.001
         2
               0.000
                           0.000
                                      0.000
                                                 0.000 enum.py:839(_decompose)
               0.000
                           0.000
       4/2
                                      0.001
                                                 0.000 sre_parse.py:407(_parse_sub)
                                                 0.000 _bootlocale.py:11(getpreferredencoding)
0.000 sre_compile.py:482(_compile_info)
               0.000
                                      0.000
         2
                           0.000
               0.000
                           0.000
                                      0.000
```