Assignment 1

May 2, 2021

---

*You are currently looking at **version 1.1** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the Jupyter Notebook FAQ course resource.*

---

1 Assignment 1

In this assignment, you'll be working with messy medical data and using regex to extract relevant infromation from the data.

Each line of the `dates.txt` file corresponds to a medical note. Each note has a date that needs to be extracted, but each date is encoded in one of many formats.

The goal of this assignment is to correctly identify all of the different date variants encoded in this dataset and to properly normalize and sort the dates.

Here is a list of some of the variants you might encounter in this dataset: * 04/20/2009; 04/20/09; 4/20/09; 4/3/09 * Mar-20-2009; Mar 20, 2009; March 20, 2009; Mar. 20, 2009; Mar 20 2009; * 20 Mar 2009; 20 March 2009; 20 Mar. 2009; 20 March, 2009 * Mar 20th, 2009; Mar 21st, 2009; Mar 22nd, 2009 * Feb 2009; Sep 2009; Oct 2010 * 6/2008; 12/2009 * 2009; 2010

Once you have extracted these date patterns from the text, the next step is to sort them in ascending chronological order accoring to the following rules: * Assume all dates in xx/xx/xx format are mm/dd/yy * Assume all dates where year is encoded in only two digits are years from the 1900's (e.g. 1/5/89 is January 5th, 1989) * If the day is missing (e.g. 9/2009), assume it is the first day of the month (e.g. September 1, 2009). * If the month is missing (e.g. 2010), assume it is the first of January of that year (e.g. January 1, 2010). * Watch out for potential typos as this is a raw, real-life derived dataset.

With these rules in mind, find the correct date in each note and return a pandas Series in chronological order of the original Series' indices.

For example if the original series was this:

```
0    1999
1    2010
2    1978
3    2015
4    1985
```

Your function should return this:

```
0    2
1    4
2    0
3    1
4    3
```

Your score will be calculated using Kendall's tau, a correlation measure for ordinal data.

This function should return a Series of length 500 and dtype int.

```
In [2]: import pandas as pd

        doc = []
        with open('dates.txt') as file:
            for line in file:
                doc.append(line)

        df = pd.Series(doc)
        df.tail(100)
        df
```

```
Out[2]: 0            03/25/93 Total time of visit (in minutes):\n
        1                          6/18/85 Primary Care Doctor:\n
        2      sshe plans to move as of 7/8/71 In-Home Servic...
        3                   7 on 9/27/75 Audit C Score Current:\n
        4      2/6/96 sleep studyPain Treatment Pain Level (N...
        5                      .Per 7/06/79 Movement D/O note:\n
        6      4, 5/18/78 Patient's thoughts about current su...
        7      10/24/89 CPT Code: 90801 - Psychiatric Diagnos...
        8                           3/7/86 SOS-10 Total Score:\n
        9               (4/10/71)Score-1Audit C Score Current:\n
        10     (5/11/85) Crt-1.96, BUN-26; AST/ALT-16/22; WBC...
        11                         4/09/75 SOS-10 Total Score:\n
        12     8/01/98 Communication with referring physician...
        13     1/26/72 Communication with referring physician...
        14     5/24/1990 CPT Code: 90792: With medical servic...
        15     1/25/2011 CPT Code: 90792: With medical servic...
        16            4/12/82 Total time of visit (in minutes):\n
        17           1; 10/13/1976 Audit C Score, Highest/Date:\n
        18                    4, 4/24/98 Relevant Drug History:\n
        19     ) 59 yo unemployed w referred by Urgent Care f...
        20            7/21/98 Total time of visit (in minutes):\n
        21                        10/21/79 SOS-10 Total Score:\n
        22      3/03/90 CPT Code: 90792: With medical services\n
        23      2/11/76 CPT Code: 90792: With medical services\n
        24     07/25/1984 CPT Code: 90791: No medical services\n
        25     4-13-82 Other Child Mental Health Outcomes Sca...
        26      9/22/89 CPT Code: 90792: With medical services\n
        27        9/02/76 CPT Code: 90791: No medical services\n
```

```
28                                9/12/71 [report_end]\n
29     10/24/86 Communication with referring physicia...
                             ...
470    y1983 Clinic Hospital, first hospitalization, ...
471    tProblems Urinary incontinence : mild urge inc...
472    .2010 - wife; nightmares and angry outbursts; ...
473         shx of TBI (1975) ISO MVA.Medical History:\n
474    sPatient reported losing three friends that pa...
475                         TSH okay in 2015 Prior EKG:\n
476    1989 Family Psych History: Family History of S...
477    oEnjoys animals, had a dog x 14 yrs who died i...
478    eHistory of small right parietal subgaleal hem...
479    sIn KEP Psychiatryfor therapy and medications ...
480    1. Esophageal cancer, dx: 2013, on FOLFOX with...
481                                     y1974 (all)\n
482    h/o restraining order by sister/mother in 1990...
483    sTexas Medical Center; Oklahoma for 2 weeks; 1...
484    Death of former partner in 2004 by overdose as...
485    Was "average" student.  "I didn't have too man...
486    Contemplating jumping off building - 1973 - di...
487    appendectomy s/p delivery 1992 Prior relevant ...
488    tProblems renal cell cancer : s/p nephrectomy ...
489        ran own business for 35 years, sold in 1985\n
490                               Lab: B12 969 2007\n
491                                )and 8mo in 2009\n
492    .Moved to USA in 1986. Suffered from malnutrit...
493                                          r1978\n
494    . Went to Emerson, in Newfane Alaska. Started ...
495    1979 Family Psych History: Family History of S...
496    therapist and friend died in ~2006 Parental/Ca...
497                          2008 partial thyroidectomy\n
498    sPt describes a history of sexual abuse as a c...
499    . In 1980, patient was living in Naples and de...
Length: 500, dtype: object
```

```
In [23]: def date_sorter():
             words = df.str.extract(r'((?:\d{,2} )?(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|No
             numbers = df.str.extract(r'((?:\d{1,2})(?:(?:\/|-)\d{1,2})(?:(?:\/|-)\d{2,4}))')
             no_days = df.str.extract(r'((?:\d{1,2}(?:-|\/))?\d{4})')
             result = pd.to_datetime(words.fillna(numbers).fillna(no_days).str.replace('Decemebe
             sorted_df = pd.Series(result.sort_values().index)
             return sorted_df

          #date_sorter()
```

```
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:2: FutureWarning: currently extract

/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:3: FutureWarning: currently extract
```

```
  This is separate from the ipykernel package so we can avoid doing imports until
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:4: FutureWarning: currently extract
  after removing the cwd from sys.path.
```

```
Out[23]: 0        9
         1       84
         2        2
         3       53
         4       28
         5      474
         6      153
         7       13
         8      129
         9       98
         10     111
         11     225
         12      31
         13     171
         14     191
         15     486
         16     335
         17     415
         18      36
         19     405
         20     323
         21     422
         22     375
         23     380
         24     345
         25      57
         26     481
         27     436
         28     104
         29     299
               ...
         470    220
         471    208
         472    243
         473    139
         474    320
         475    383
         476    244
         477    286
         478    480
         479    431
         480    279
         481    198
```

```
482    381
483    463
484    366
485    439
486    255
487    401
488    475
489    257
490    152
491    235
492    464
493    253
494    427
495    231
496    141
497    186
498    161
499    413
Length: 500, dtype: int64
```

In [ ]:

In [ ]: