
Assignment1 (Score: 5.0 / 5.0)

1. Test cell (Score: 4.0 / 4.0)
 2. Test cell (Score: 1.0 / 1.0)
-

Assignment 1¶

In this assignment, you'll be working with messy medical data and using regex to extract relevant information 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 the 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 according 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. **This Series should be sorted by a tie-break sort in the format of ("extracted date", "original row number").**

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 (https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient), a correlation measure for ordinal data.

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

In []:

In [1]:

```
import pandas as pd

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

df = pd.Series(doc)
df.head(10)
```

Out[1]:

```
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...\n
3              7 on 9/27/75 Audit C Score Current:\n
4      2/6/96 sleep studyPain Treatment Pain Level (N...\n
5              .Per 7/06/79 Movement D/O note:\n
6      4, 5/18/78 Patient's thoughts about current su...\n
7      10/24/89 CPT Code: 90801 - Psychiatric Diagnos...\n
8              3/7/86 S0S-10 Total Score:\n
9      (4/10/71)Score-1Audit C Score Current:\n
dtype: object
```

In [2]:

Student's answer

```
def date_sorter():

    import numpy as np
    import re

    # Your code here
    # Testing Data
    # df = ["•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","2June, 1999",
    #      "•Mar 20th, 2009;", "Mar 21st, 2009;", "Mar 22nd, 2009",
    #      "•Feb 2009;", "Sep 2009;", "Oct 2010",
    #      "•6/2008;", "12/2009",
    #      "•2009;", "2010"]

    # df = ["•04/20/2009;", "04/20/09;", "4/20/09;", "4/3/09;",
    pattern1 = r'(0?[1-9]|1[0-2])([\\/-])(0?[1-9]|12|\\d|30|31)([\\/-])(\\d{4}|\\d{2}) '
    df1 = df.str.extractall(pattern1)
    df1.columns = ["month", "day", "year"]
    df1 = df1.reset_index()
    #df1

    # "•Mar-20-2009;", "Mar 20, 2009;", "March 20, 2009;", "Mar. 20, 2009;", "Mar 20 2009;",
    #October 14 1974
    # "•Mar 20th, 2009;", "Mar 21st, 2009;", "Mar 22nd, 2009",
    #pattern2 = r'(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)[a-z\\.]*[ -](\\d{1,2})[a-
```

```

pattern2 = r'(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)[a-z\.]*[-](\d{1,2})[a-z]
df2=df.str.extractall(pattern2)
df2.columns = ["month", "day", "year"]
df2 = df2.reset_index()
#df2

# "•20 Mar 2009;","20 March 2009;","20 Mar. 2009;","20 March, 2009","2June, 1999",
# "•Feb 2009;","Sep 2009;","Oct 2010",
pattern3 = r'(\d{1,2})?[-]?(\d{1,2})?(Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)[a-z\.\,]*
df3=df.str.extractall(pattern3)
df3.columns = ["day", "month", "year"]
df3 = df3.reset_index()
#df3

# "•6/2008;","12/2009",
pattern4 = r'(\d{1,2})[/](\d{4})'
df4 = df.str.extractall(pattern4)
df4.insert(0, column='day', value=np.nan)
df4.columns = ["day", "month", "year"]
df4 = df4.reset_index()
#df4

## "•2009;","2010"
pattern5 = r'(\d{4})'
df5 = df.str.extractall(pattern5)
df5.insert(0, column='day', value=np.nan)
df5.insert(1, column='month', value=np.nan)
df5.columns = ["month", "day", "year"]
df5 = df5.reset_index()
#df5

output = df1.append(df2[~df2.level_0.isin(df1.level_0)])
#output.shape
output = output.append(df3[~df3.level_0.isin(output.level_0)])
#output.shape
output = output.append(df4[~df4.level_0.isin(output.level_0)])
#output.shape
output = output.append(df5[~df5.level_0.isin(output.level_0)])
#output.shape

output = pd.DataFrame(output,columns = ["level_0", "match", "day", "month","year"])
output.year = np.where(output.year.apply(len)==2, "19"+output.year, output.year)
output = output.fillna("1")

month_replace ={
    'Jan' : 1,
    'Feb' : 2,
    'Mar' : 3,
    'Apr' : 4,
    'May' : 5,
    'Jun' : 6,
    'Jul' : 7,
    'Aug' : 8,
    'Sep' : 9,
    'Oct' : 10,
    'Nov' : 11,
    'Dec' : 12
}

```

```

output.month = output.month.replace(month_replace)

output.day = output.day.astype(int)
output.month = output.month.astype(int)
output.year = output.year.astype(int)

output["date"] = pd.to_datetime(output.loc[:,["year", "month", "day"]])

output = output.sort_values(["date", "level_0"]).reset_index(drop=True)
#output.info()
#return_value = pd.Series(output.level_0, name="index")

return output.level_0#return_value # Your answer here

date_sorter()

```

Out[2]:

```

0      9
1     84
2      2
3     53
4     28
...
495   427
496   141
497   186
498   161
499   413
Name: level_0, Length: 500, dtype: int64

```

In [3]:

Grade cell: cell-373f878879c00996

Score: 4.0 / 4.0

Kendall Tau: 0.9999999999999999

In [4]:

Grade cell: cell-0ebae76e6cd794be

Score: 1.0 / 1.0

In [5]:

```
type(date_sorter())
```

Out[5]:

```
pandas.core.series.Series
```

In []:

This assignment was graded by mooc_adswpy:5a1483384bca, v1.47.103123