Chris Sutton

Lab 8

EN.605.646.

```
from google.colab import drive
drive.mount('/content/drive')
import re
```

```
     Mounted at /content/drive
```

```
labdir = '/content/drive/My Drive/NaturalLangaugeProcessing/lab08'
%cd "$labdir"
!ls
```

```
     /content/drive/My Drive/NaturalLangaugeProcessing/lab08
     obits.test.txt  obits.train.txt
```

```
data= open('obits.test.txt',encoding='utf-8').read()
```

```
# (.*), .*(\d{3}|\d{2})(,|) of (\w+)
# (.*), .*(\d{3}|\d{2})(,|) of (\w+).* (he |his |she |her )
# (.*),( | age )(\d{3}|\d{2})(,|) of (\w+).* (he |she |her | his)
# re.sub(pattern, repl, string, count=0, flags=0)¶
# re.findall(r"(\w+) is made of (\w+)",data2, re.MULTILINE)
# (.*),( | age )(\d{3}|\d{2})(,|) of (\w+).* (he |she |her | his)
# (.*),( | age )(\d{3}|\d{2})(,|) (of|formerly) (\w+).* (he |she |her | his).* (husband|wife)
# .* ((husband|wife)(,|)|married) (\w+)
# .* (?:survived by (?:his|her)|Survivors include (?:his|her)) (\w+)
                                         G>
```

Saved successfully! ×

```
# print(data)
```

```
d= re.sub(r"(\.\n|\. \n|\!\n)",". ", data)
```

```
print(d)
```

```
     <P ID=100>
     Bernard Pugh, 90, of Centerville, passed away on April 21, 2019 at his residence. Bernar

     <P ID=101>
     Sherri Oden, 62, of Cincinnati, passed away on April 11, 2019 at Mercy Medical Center ir

     <P ID=102>
```

```
    Dennis Strickler, 62 of Moravia passed away Friday, April 5, 2019 at Mercy Medical Cente

    <P ID=103>
    Candita Marie Furlin, age 42 of Des Moines and formerly of Seymour passed away on March

    <P ID=104>
    Virginia Elizabeth Koestner, 97 of Centerville passed away Wednesday, March 13, 2019 at

    <P ID=105>
    Florence R. Schau, 91, formerly of South Wheeling, died Saturday, April 27, 2019 at Gooc

    <P ID=106>
    Linda Kay Yee, 68, of Wheeling, passed peacefully to the house of the Lord Tuesday, Apri

    <P ID=107>
    Clyde William “Casey” Caseman, 90, of Wheeling, WV, passed away peacefully Wednesday, Ap

    <P ID=108>
    John Patrick “Pud” Moyle, 97, of Wheeling, WV passed away on Tuesday, April 2, 2019 and

    <P ID=109>
    Haines, William "Spencer", 103 of Weirton, WV formerly of Bethlehem, WV died on Saturday
```

```
result = re.findall(r"(.*),(?: | age )(\d{3}|\d{2})(?:,|) (?:of|formerly of) (\w+).* (he|she|
```

```
print(result)
```

```
    [('Bernard Pugh', '90', 'Centerville', 'his'), ('Sherri Oden', '62', 'Cincinnati', 'her
```

```
obsplit = d.split('</P>\n')
```

Saved successfully! ×

```
band|wife)(?:,|)|married) (\w+)", rx) for rx in obsplit]
all(r".* (?:(?:husband|wife)(?:,|)|married) (\w+)", x), o
```

```
print(spouse)
```

```
    [['Jewell'], ['Paul'], ['Vicky'], [], ['John'], ['Tony'], ['Paul'], ['of'], ['Mary'], [
```

```
final=[]
for s,r in zip(spouse,result):
  if len(s) ==0:
    final.append((r[0],r[1],r[2],r[3],''))
  else:
    final.append((r[0],r[1],r[2],r[3],s[0]))
```

```
print(final)
```

```
[('Bernard Pugh', '90', 'Centerville', 'his', 'Jewell'), ('Sherri Oden', '62', 'Cincinna
```

```
for i in range(len(final)):
  if final[i][3] == 'his' or final[i][3] == 'he':
    if len(final[i][4]) ==0:
      final[i] = (final[i][0],final[i][1],final[i][2],'male',final[i][4])
    else:
      final[i] = (final[i][0],final[i][1],final[i][2],'male',final[i][4]+' '+final[i][0].spli
  else:
    if len(final[i][4]) ==0:
      final[i] = (final[i][0],final[i][1],final[i][2],'female',final[i][4])
    else:
      final[i] = (final[i][0],final[i][1],final[i][2],'female',final[i][4]+' '+final[i][0].sp
```

(12 points) Your task is to extract these relations about the deceased:

- name of deceased individual: usually the first noun in the first sentence, and may not be identified otherwise
- sex of the deceased
- age at death (in years)
- location(s) where the deceased was a resident of (may be more than one)
- spouse(s) of the deceased

Saved successfully! ×

```
[('Bernard Pugh', '90', 'Centerville', 'male', 'Jewell Pugh'), ('Sherri Oden', '62', 'Ci
```

Double-click (or enter) to edit

```
ner= re.sub(r", | ","\n", d)
```

```
# Change this path to point to your own directory:
labdir = '/content/drive/My Drive/NaturalLangaugeProcessing/lab06'
# !mkdir -p "$labdir"
# !mkdir -p "$labdir/data"
# !mkdir -p "$labdir/checkpoint"
%cd "$labdir"
!ls
```

```
/content/drive/My Drive/NaturalLangaugeProcessing/lab06
checkpoint  data  LM-LSTM-CRF
```

```
!pip install torch==1.2.0 torchvision==0.4.0
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/pub
Collecting torch==1.2.0
  Downloading torch-1.2.0-cp37-cp37m-manylinux1_x86_64.whl (748.9 MB)
     |████████████████████████████████| 748.9 MB 622 bytes/s
Collecting torchvision==0.4.0
  Downloading torchvision-0.4.0-cp37-cp37m-manylinux1_x86_64.whl (8.8 MB)
     |████████████████████████████████| 8.8 MB 49.1 MB/s
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from tor
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from torch
Requirement already satisfied: pillow>=4.1.1 in /usr/local/lib/python3.7/dist-packages (
Installing collected packages: torch, torchvision
  Attempting uninstall: torch
    Found existing installation: torch 1.12.1+cu113
    Uninstalling torch-1.12.1+cu113:
      Successfully uninstalled torch-1.12.1+cu113
  Attempting uninstall: torchvision
    Found existing installation: torchvision 0.13.1+cu113
    Uninstalling torchvision-0.13.1+cu113:
      Successfully uninstalled torchvision-0.13.1+cu113
ERROR: pip's dependency resolver does not currently take into account all the packages t
torchtext 0.13.1 requires torch==1.12.1, but you have torch 1.2.0 which is incompatible
torchaudio 0.12.1+cu113 requires torch==1.12.1, but you have torch 1.2.0 which is incomp
fastai 2.7.10 requires torch<1.14,>=1.7, but you have torch 1.2.0 which is incompatible
fastai 2.7.10 requires torchvision>=0.8.2, but you have torchvision 0.4.0 which is incon
Successfully installed torch-1.2.0 torchvision-0.4.0
```

Saved successfully!

```
eq_wc.py --load_arg ../checkpoint/ner_cwlm_lstm_crf.json\
 --load_check_point ../checkpoint/ner_cwlm_lstm_crf.model --gpu 0\
  --input_file ../data/testUpdatedNER.txt --output_file my-test-file-Suttonlab08.out
```

```
loading dictionary
loading corpus
loading model
annotating
```

```
dataNER= open('LM-LSTM-CRF/my-test-file-Suttonlab08.out',encoding='utf-8').read()
```

```
D_NER =dataNER.split('</P>')
```

```
print(D_NER)
```

```
['-DOCSTART- -DOCSTART- -DOCSTART-\n\n<P ID=100> <PER> Bernard Pugh </PER> 90 of <LOC> (
```

```
len(D_NER)
```

```
11
```

```
# (?:survived|Surviving).*<PER>(.*)</PER>
survived= list(map(lambda x: re.findall(r"(?:survived|Surviving|Survivors)(.*)</PER> (?:Funer
```

(12 points) Additionally, try to extract two or more of the following relations, though they may often

be missing and are more difficult:

```
# survived by-----------------
print(survived)
```

```
[[' by his daughter <PER> Cheryl “Cheri” Pugh; </PER> nieces: <PER> Janice Barnett Renee
```

```
# passed away on ---------------

# (?:passed away (?:on|)|died (?:on|))(.*\d+) (?:at|surrounded by|with)
passed = list(map(lambda x: re.findall(r"(?:passed away (?:on|)|died (?:on|))((?:\w+|) \w+ \d
```

```
print(passed)
```

Saved successfully!

```
1 2019'], ['Friday April 5 2019'], [' March 26 2019'], [
```

```
labdir = '/content/drive/My Drive/NaturalLangaugeProcessing/lab08'
%cd "$labdir"
!ls
```

```
/content/drive/My Drive/NaturalLangaugeProcessing/lab08
obits.test.txt  obits.train.txt
```

```
# in error I've been working on the test set from the begining. So to rectify this, I test on
data= open('obits.train.txt',encoding='utf-8').read()
```

```
d= re.sub(r"(\.\n|\. \n|\!\n)",". ", data)

result = re.findall(r"(.*),(?: | age )(\d{3}|\d{2})(?:,|) (?:of|formerly of) (\w+).* (he|she|


obsplit = d.split('</P>\n')


# spouse =[re.findall(r".* (?:(?:husband|wife)(?:,|)|married) (\w+)", rx) for rx in obsplit]
spouse = list(map(lambda x : re.findall(r".* (?:(?:husband|wife)(?:,|)|married) (\w+)", x), o


final=[]
for s,r in zip(spouse,result):
  if len(s) ==0:
    final.append((r[0],r[1],r[2],r[3],''))
  else:
    final.append((r[0],r[1],r[2],r[3],s[0]))


for i in range(len(final)):
  if final[i][3] == 'his' or final[i][3] == 'he':
    if len(final[i][4]) ==0:
      final[i] = (final[i][0],final[i][1],final[i][2],'male',final[i][4])
    else:
      final[i] = (final[i][0],final[i][1],final[i][2],'male',final[i][4]+' '+final[i][0].spli
  else:
    if len(final[i][4]) ==0:
      final[i] = (final[i][0],final[i][1],final[i][2],'female',final[i][4])
    else:
      final[i] = (final[i][0],final[i][1],final[i][2],'female',final[i][4]+' '+final[i][0].sp




print(final)
```

Saved successfully! ×

```
'70', 'Ft', 'male', 'of Milisich'), ('Robert (Bob) A. Ka
```

metric	Precision	recall	F1
Name	(6/6) 100.0%	(6/20) 30.0%	12/(26) 46.1%
Age	6/6 100.0%	(6/20) 30.0%	12/(26) 46.1%
Location	0 0.0%	0 0.0%	0%
sex	6/6 100.0%	(6/20) 30.0%	12/(26) 46.1%
Spouse	0 0.0%	0 0.0%	0%

(6 points) In addition to scores, describe what was easy and what was difficult, why some errors occur, and

whether errors observed are easily repairable or are more complicated to correct. Be sure to include

plentiful examples.

This lab I completely messed up, I worked on the test set accidentally. I hadn't noticed the name of the file I was working on as I simply was

refering to variables. The cover the questions asked here: It seemed that out of the five fields location was difficult because I could be stated in many forms (city;

city, state ; location in city). An example is: Des Moines or wheeling, south wheeling, wheeling, Wv

The addditonal fields chose to capture were "survived by" and "passed away on". These were extremely troubling to deal with. I passsed the text through conlleval to tag

people with the hope I could figure out a search for but the best I could come up with is to capture as much of the text as I could in bulk. Passed away on, went

easier as it was more straight forward to caputure dates, however, there was still trouble due to variablity in the text.

Double-click (or enter) to edit

Saved successfully! ×

✓ 0s completed at 10:06 PM ● ✕

Saved successfully! ✕