Parsing_Tutorial

May 30, 2021

1 Understanding the Social Networks of Emma B. Andrews

In this tutorial, we want to use Python to parse the Emma B. Andrews diaries, so as to explore the network of relationships. To parse the TEI documents, we will use several modules. These are as follows:

- csv
- Beautiful Soup 4
- lxml
- matplotlib
- nltk

The Beautiful Soup, lxml, and Matplotlib modules need to be installed. If you are running Jupyter Lab with the virtual environment, then these modules are already installed. The `csv` module comes preinstalled in the Python virtual environment.

1.1 Import Modules (Dependencies)

[2]:

```
import csv # Python's Comma Separate Values Parser
from bs4 import BeautifulSoup # Beautiful Soup is for parsing HTML and XML files
import lxml # lxml is a secondary parser for beautiful soup
import nltk # Natural Langauge Toolkit
import re # Python's Regular Expression Module
```

1.2 Read Volume into Python to Parse with Beautiful Soup

[3]:

```
# Since the journal volume we want exists in the same directory as our Jupyter␣
 ↪Notebook, we can use the document name with extension.
journal = 'volume17.xml'

# Now we want to create a Beautiful Soup object with our file. We will unpack␣
 ↪what this means in more detail below.
with open(journal) as xml:
    soup = BeautifulSoup(xml, 'lxml-xml')
```

1.3 Parse Diary for Network Analysis

The Diaries are encoded according to the `TEI` standards. Thus, the `<text>…</text>` element encloses the contents of the dairy. We want to parse every day of the dairy and then further manipulate the data for Graph Analysis. Each child within the `<text>` root is an entry according to the day.

1.3.1 Extract Daily Entries from Volume 17

[4]:
```
# To extract the daily entries, we need to traverse the text root and gather␣
 ↪together all <div> elements with a type of entry
entries = soup.find_all("div", {"type": "entry"})
num_entries = len(entries)
f'Volume 17 contains {num_entries} entries'
```

[4]:
```
'Volume 17 contains 43 entries'
```

1.3.2 Extract the Dates of the Entries and Create a Timeline

[5]:
```
# Iterate over the entries and get the Date for each
dates = soup.find_all(attrs={"xml:id": re.compile("EBA-[0-9--]+")})
total_dates = len(dates)
f'There are a total of {total_dates} entries'
```

[5]:
```
'There are a total of 43 entries'
```

1.3.3 Extract the persName Entities from Each Entry

[37]:
```
# Create a List of All the People
network = []
for entry in entries:
    peoples = entry.find_all('persName')
    for person in peoples:
#         person = person.text
        network.append(person['ref'])
```

[38]:
```
network
```

[38]:
```
['#Troyon_Constant',
 '#Corot_Jean_Baptiste_Camille',
 '#Rathbone_Mr',
 '#Rathbone_Mr #Rathbone_Mrs',
 '#Rathbone_Elena',
 '#Parsons_John #Parsons_Florence_Van_Corltandt',
 '#Draper_Mr',
 '#Gorst_Group',
 '#Trefusis_Walter',
 '#Carter_Bonham_Mr',
 '#Rathbone_Elena',
 '#Lovatt_Mr #Lovatt_Master',
 '#Lovatt_Mr #Lovatt_Master',
 '#Gay_Walter',
 '#Gay_Walter_Mr #Gay_Mrs',
 '#Gay_Mrs',
 '#Rathbone_Elena',
 '#Buckley_Mr #Buckley_Mrs',
 '#Carter_Howard',
 '#Weigall_Arthur',
 '#Nicol_Erskine',
 '#Davis_Theodore_M',
 '#Weigall_Arthur',
 '#Butler_Mrs',
 '#Nicol_Erskine',
 '#Butler_Mrs',
 '#Weigall_Arthur',
 '#Davis_Theodore_M',
 '#Burton_Harry',
 '#Maspero_Gaston',
 '#Davis_Theodore_M',
 '#Davis_Theodore_M',
 '#Jones_Harold',
 '#Jones_Cyril',
 '#Rathbone_Elena',
 '#Davis_Theodore_M',
 '#Carter_Howard',
 '#Nicol_Erskine',
 '#Davis_Theodore_M',
 '#Jones_Harold',
 '#Jones_Cyril',
 '#Jones_Harold',
 '#Mumm_von_Schwarzenstein_Alfons',
 '#Fahnestock_Gibson #Fahnestock_Mrs',
 '#Kelly_Miss',
 '#Whitaker_Mr',
 '#Whymper_Charles',
 '#Kelly_Mr #Kelly_Miss',
 '#Burton_Harry',
 '#Davis_Theodore_M',
 '#Crane_Lancelot',
 '#Horemheb',
 '#Trefusis_Walter',
 '#Davis_Theodore_M',
 '#Whitaker_Mr',
 '#Newberry_Percy #Newberry_Mrs',
 '#Maspero_Gaston #Maspero_Louise',
 '#Davis_Theodore_M',
 '#Maspero_Mme',
 '#McCormick_Mrs',
 '#Scott_Miss',
 '#Buckley_Mr #Buckley_Mrs',
 '#Carter_Howard',
 '#Burton_Harry',
 '#Davis_Theodore_M',
 '#Gorst_Lady',
 '#Gorst_Miss',
 '#Hunter_Mrs',
 '#Warner_Mrs #Warner_Miss',
 '#Webb_Miss',
 '#Guadalmina_Marquesa #Guadalmina_son_of',
 '#Davis_Theodore_M',
 '#Mohassib_Mohammed',
 '#Cassatt_Mary_S',
 '#Kelekian_Mr #Kelekian_Mrs',
 '#Nicol_Erskine',
 '#Peabody_Endicott',
 '#Fairfield_Osborn_Henry',
 '#Newberry_Percy #Newberry_Mrs',
 '#Fahnestock_Gibson #Fahnestock_Mrs',
 '#Davis_Theodore_M',
 '#Alexander_Charles #Alexander_Mrs',
 '#Alexander_girls',
 '#Hobhouse_Henry #Hobhouse_Mrs',
 '#Farrer_Gaspard',
 '#Williams_Mr #Williams_Mrs',
 '#Langley_Mr #Langley_Mrs',
 '#Foster_Giraud #Foster_Mrs',
 '#Ives_Miss',
 '#Alexander_Mrs',
 'Roosevelt_Theodore_Jr',
 '#Collander_Livingston_John #Collander_Livingston_Mrs',
 '#Davis_Theodore_M',
 '#Whymper_Charles',
 '#Foster_Giraud #Foster_Mrs',
 '#Graham_Mrs',
 '#Collander_Livingston_John #Collander_Livingston_Mrs',
 '#Hamilton_Fish_Webster_Mrs',
 '#Auchincloss_Mr #Auchincloss_Mrs',
 '#Jennings_Miss',
 '#Naville_Edouard #Naville_Marguerite',
 '#Rodier_M',
 '#Cherry_Mrs #Cherry_Miss',
 '#Cust_Mr',
 '#Maspero_Gaston #Maspero_Louise',
 '#Davis_Theodore_M',
 '#Carter_Howard',
 '#Carnarvon_Lord',
 '#Hobhouse_Henry #Hobhouse_Mrs',
 '#Farrer_Gaspard',
 '#Davis_Theodore_M',
 'Morgan_John_Pierpoint',
 '#Davis_Theodore_M',
 '#Layard_Lady',
 '#Nicol_Erskine',
 'Morgan_John_Pierpoint',
 '#Davis_Theodore_M',
 '#Auchincloss_Mrs',
 '#Jennings_Miss',
 '#Contardone_Contessa',
 '#Rathbone_Elena',
 '#Newberry_Percy #Newberry_Mrs',
 '#Johnson_Mr',
 '#Whymper_Charles',
 '#Nicol_Erskine',
 '#Burton_Harry',
 '#Rathbone_Elena',
 '#Rathbone_Elena',
 '#Naville_Edouard #Naville_Marguerite',
 '#Naville_Mme',
 '#Naville_Edouard #Naville_Marguerite',
 '#Whitmore_Mr',
 '#Dixon_Mr',
 '#Naville_Edouard #Naville_Marguerite',
 '#Rathbone_Elena',
 '#Akhenaten',
 '#Rathbone_Elena',
 '#Burton_Harry',
 '#Jones_Harold',
 '#Draper_Mr',
 '#Rathbone_Elena',
 '#Newberry_Percy_E',
 '#Deimer_Michael_Z',
 '#Kelekian_Dikran',
 '#Kassera',
 '#Davis_Theodore_M',
 '#Burton_Harry',
 '#Nachman',
 '#Rathbone_Elena',
 '#Burton_Harry',
 '#Davis_Theodore_M',
 '#Rathbone_Elena',
 '#Burton_Harry',
 '#Winlock_Herbert',
 '#Davis_Theodore_M',
 '#Kyticas_N_D',
 '#Davis_Theodore_M',
 '#Newberry_Percy #Newberry_Mrs',
 '#Johnson_Mr',
 '#Whittaker_Mr #Whittaker_Mrs',
 '#Layard_Lady',
 '#Johnson_Mr',
 '#Trefusis_Walter',
 '#Davis_Theodore_M',
 '#Pasha_Artin',
 '#Sayce_Archibald',
 '#Kyticas_N_D',
 '#Daressy_M',
 '#Tiye',
 '#Davis_Theodore_M',
 '#Duvar_Mrs',
 '#Rathbone_Elena',
 '#Bonham_Carter_Mr',
 '#Burton_Harry',
 '#Whymper_Charles',
 '#Davis_Theodore_M',
 '#Pasha_Artin',
 '#Sayce_Archibald',
 '#Bonham_Carter_Mr',
 '#Trefusis_Walter',
 '#Graham_Margery',
 '#Graham_John',
 '#Northampton_Lord',
 '#Duvar_Mrs',
 '#Coater_Miss']
```

```
[ ]:
```