MSSP 607: Homework 2

Extra

1. What percentages of featured biographies describe people who are currently alive? What percentage of pages did your code fail to find data on, and why? What is the distribution of dates of birth and death on featured biographies, and what does that tell us about the site?

Among 1188 entries I was able to extra, 736 entries have value in the “Died” column. If I assume that missing data in “Died” column means that the person is still alive (it’s quite irresponsible of me to simply assume this but I’m also too tired), the percentage for people that are alive is about 38.05%.

Among 1399 entries, I was able to extract 1188 infoboxes (84.92%). Among the failed cases, 240 of which are caused by that lack of infoboxes on the biographies’ pages (such as the pages of Fanny Imlay (no info table) and Gabriel Pleydell (only a picture with no extra info)) and by weird box-formatting that my code failed to read (such as in Choe Bu’s and Bob Meusel’s bio page).

Coincidentally, I calculated the birth years of the people in the list in Part 2-4 of the assignment, and concluded that the site includes modern history the most. So for this part, I calculated the life spans of the figures passed (weird data already cleaned, so it may not cover all the deceased) to see if there’s any pattern. The result is as shown in the following table.

Life Span (years)	Count	50-59	78
0-9	2 (may be error)	60-69	105
10-19	8	70-79	112
20-29	41	80-89	105
30-39	51	90-99	58
40-49	57	100-110/110-120	5/1

Most of people made to more than 60 years old. So speaking, people included in the list relatively lived long lives. Also, a problem with my calculation this time is that I did not distinguish people from BC or AC as I did in part 2-4, which may cause some inaccuracy in the life span calculation.

2. Technical documentation

The “extra_infoboxes.csv” file contains information extracted from the infoboxes in feature article biographies’ pages in Wikipedia. In each row, information about a biography’s protagonist is stored and each column stands for a feature category in the infoboxes. In each column, information extracted from the infoboxes is stored as a string, which may contain more than one detail about the person. For example, the content of “Born” column for Jean Bellette is “1908-03-25, 25 March 1908, Hobart, Tasmania, Australia”, which contains the numeric and spell form of her birthday, and her birthplace. Due to the large amount of content, problems concerning category extraction occurred during the constructing of the dataframe. Among all columns, some of them are not actually representative as a rightful category which can be applied to the entire group and please ignore those. To access the dataset and test whether the download was successful, one can use following sample codes and compare the results to the quoted ones.

```
Codes:
# open and read the file; store it in a dataframe
df = pd.read_csv('extra_infoboxes.csv')

# get information about Bronwyn Bancroft
# by name searching
# using .dropna(axis=1, how='all') to clean columns with no value in it
bronwyn_bancroft_info1 = df.loc[df.Name == "Bronwyn
Bancroft"].dropna(axis=1, how='all')
# or by index searching
# using .dropna(axis=0, how='all') to clean rows with no value in it
bronwyn_bancroft_info2 = df.iloc[0,:].dropna(axis=0, how='all')

# to match the infobox picture included in the instructions
# check the page for Ursula K. Le Guin
ursula_le_guin_info = df.loc[df.Name == "Ursula K. Le Guin"].dropna(axis=1,
how='all')
```

```
Printer Functions:
# to print the result
print("###  Bronwyn Bancroft 1  ###")
print(bronwyn_bancroft_info1)

print("###   Bronwyn Bancroft 2  ###")
print(bronwyn_bancroft_info2)

print("###  Ursula K. Le Guin  ###")
print(ursula_le_guin_info)
```

And the result looks like:

```
###    Bronwyn Bancroft 1    ###
      Unnamed: 0    ...                                                    Notable work
0                    0    ...              Prevention of AIDS, 1992, Tempe Reserve sports...

[1 rows x 5 columns]

###      Bronwyn Bancroft 2    ###
Unnamed: 0                                                                            0
Name                                                                  Bronwyn Bancroft
Born                                 1958 (age&#160;60&#8211;61, Tenterfield, New S...
Nationality                                                                 Australian
Notable work                           Prevention of AIDS, 1992, Tempe Reserve sports...
Name: 0, dtype: object

###    Ursula K. Le Guin    ###
         Unnamed: 0    ...                                   Period
252                  252    ...    c, 8201;1959, 160;&#8211; 2018

[1 rows x 12 columns]
```