DSE 315 : Data Science In Practice
Course Instructor: Dr Parthiban Srinivasan

Assignment 1
Name : Rita Abani 19244
Department of Electrical Engineering and Computer Science
GitHub : https://github.com/DRA-chaos?tab=repositories

Using web scraping and Biopython to extract cancer data from NIH and ASCO to build a consolidated program for search and retrieval of Cancer Research centers and oncologists

Motivation

Cancer research is **crucial to improve the prevention, detection and treatment of these cancers, and ensure that survivors live longer, better quality lives**. Research also helps identify the causes of cancer and is pointing the way to improved methods of diagnosis and treatment. On these lines, statistical tools in Data analysis and biostatistics help us find avenues to understand trends in cancer technology , novel findings and reliability of research that can help make diagnosis more accessible and the process smoother altogether.

Packages/Libraries used :

1. Pandas
2. Numpy
3. Biopython - specifically the Entrez database
4. BeautifulSoup

Introduction

In this assignment, Web Scraping has been done on Multiple research websites under the NIH so as to form a consolidated consortium to access data relating to oncologists, medical practitioners and cancer research centres. The data pertaining to the National Cancer Institute (NCI) afilliated research centres has been downloaded and converted to a CSV file from the website : https://www.cancer.gov/research/nci-role/cancer-centers.

We then proceed to scrape data from the Cancer.Net portal's The 'Find an Oncologist Database' is made available by ASCO as an informational resource for patients and caregivers. The database includes the names of physicians and other health professionals from certain ASCO membership categories who have given their permission to be identified publicly. We match the oncologists obtained from the ASCO portal to the hospitals from the NCI portal in the last step (we add the names of additional hospitals not present in the previous list and save both the data in the onco_df frame.

We then move on to using the Biopython's entrez functionality that helps us access the Pubmed database. From this we obtain the Pubmed ID of the corresponding authors in the onto_df data frame. This gives us useful information about the research works carried out by the oncologist on the specialization of cancer or any novel method they have been working on.

EDA is then performed on the consolidated data frame to understand the correlation between number of oncologists and performance of the cancer institute (We cross check if our Data Analysis matches with the trends shown in the news).

We further move on the explore the number of publications amongst the Oncologist's community and rank the researchers as well as plot the distribution.

The workflow is as follows :

1. Outsourcing information on the ambit of our sample space, i.e the designated cancer centers or hospitals on the website.The cancer centers have been downloaded from this link. I have uploaded the CSV file in this google colab notebook. There are 71 NCI associated Cancer centers saved in the center_df dataframe

	center_name	nci_link
0	Abramson Cancer Center	/research/nci-role/cancer-centers/find/upennab...
1	Albert Einstein Cancer Center	/research/nci-role/cancer-centers/find/alberte...
2	Alvin J. Siteman Cancer Center	/research/nci-role/cancer-centers/find/washing...
3	Arizona Cancer Center	/research/nci-role/cancer-centers/find/arizonacc
4	Case Comprehensive Cancer Center	/research/nci-role/cancer-centers/find/casewes...

center_df.head()

2. Outsourcing or obtaining data about the oncologists working in these centers.In this step, we would scrap data from this link, the 'Find an Oncologist Database' which is made available by ASCO as an informational resource for patients and caregivers.

- Information consists of Name, Phone Number, Specialization and address of the given doctor.
- We then merged the scraped data obtained from above with the Cancer center's center_df dataframe by performing an intersection operation. (We additionally add those cancer centers that weren't present in the center_df

frame but are present in the ASCO data). The merged data frame is labelled as ‘onco_df’

	name	degree	phone	center_name2	address	city_state	speciality	certificate	center_name
0	John H. Glick	MD, FASCO	Search for Phone Number	University of Pennsylvania-Abramson Cancer Center	3400 Civic Center Blvd Ste 3-300S	Philadelphia, PA 19104-5127, US	[Breast Cancer, Cancer Prevention]	[Internal Medicine, Medical Oncology]	Abramson Cancer Center
1	Arthur M. Feldman	MD	(215) 696-3540	University of Pennsylvania-Abramson Cancer Center	51 N 39th St Ste 103APenn Presbyterian Medcl Ctr	Philadelphia, PA 19104-2640, US	[Breast Cancer, Geriatrics Oncology]	[Internal Medicine, Medical Oncology]	Abramson Cancer Center
2	David M. Mintzer	MD	Search for Phone Number	Abramson Cancer Center at Pennsylvania Hospital	230 W Washington Sq Fl 2	Philadelphia, PA 19106-3500, US	[Breast Cancer, Lung Cancer, Palliative Care/E...	[Hematology, Hospice and Palliative Medicine, ...	Abramson Cancer Center
3	Kristina Lynne Maletz Novick	MD	Search for Phone Number	Abramson Cancer Center	1425 Portland Ave	Rochester, NY 14621-3011, US	[]	[Radiation Oncology]	Abramson Cancer Center
4	Charles John Schneider	MD, FACP	Search for Phone Number	Hospital of the University of Pennsylvania, Ab...	3400 Civic Center BlvdPereleman Center for Adv...	Philadelphia, PA 19104-5127, US	[Clinical Research, Developmental Therapeutics...	[Medical Oncology]	Abramson Cancer Center

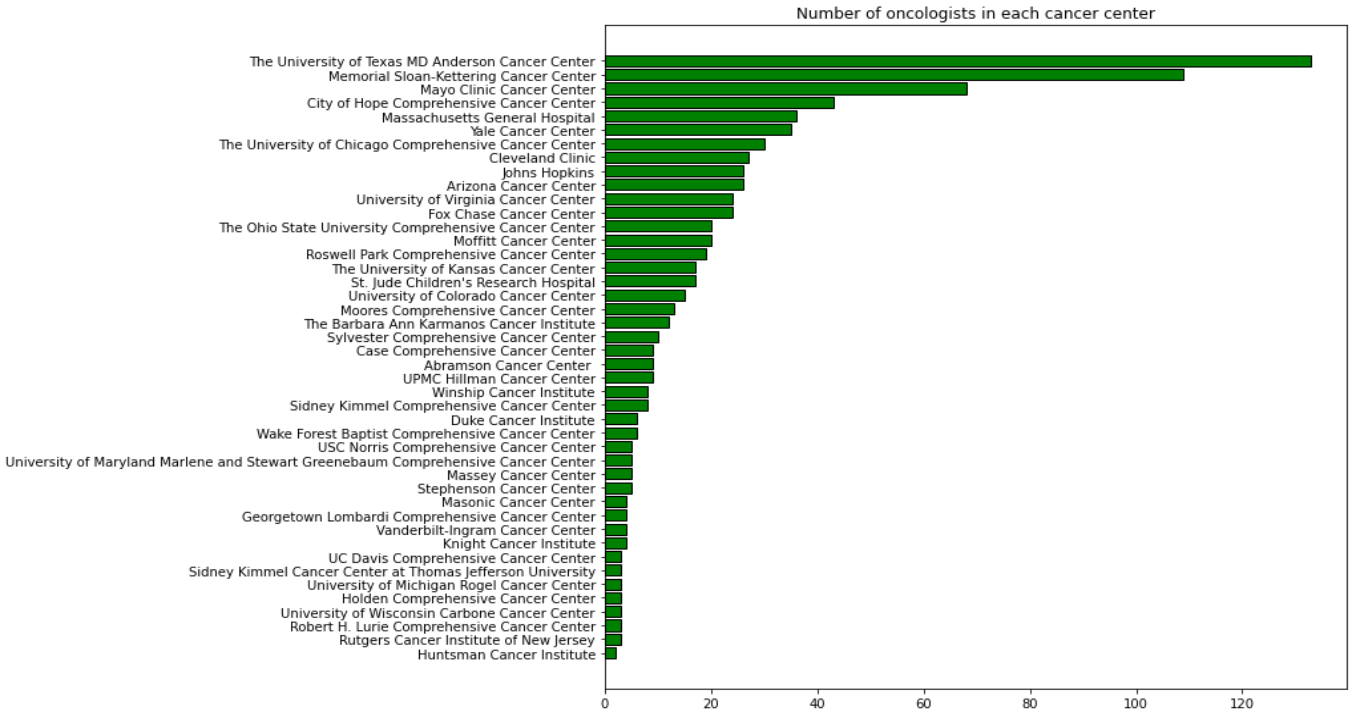
3. I have then used Biopython, a python based Bioinformatics package to access the NCBI database through Entrez (documentation). Entrez is an online search system provided by NCBI. It provides access to nearly all known molecular biology databases with an integrated global query supporting Boolean operators and field search. It returns results from all the databases with information like the number of hits from each databases, records with links to the originating database, etc. In our case we would be using Entrez as a tool to obtain the PubMed ID (which can give helpful information about the research track record of an oncologist).

	name	degree	phone	center_name2	address	city_state	speciality	certificate	center_name	article_num
0	John H. Glick	MD, FASCO	Search for Phone Number	University of Pennsylvania-Abramson Cancer Center	3400 Civic Center Blvd Ste 3-300S	Philadelphia, PA 19104-5127, US	[Breast Cancer, Cancer Prevention]	[Internal Medicine, Medical Oncology]	Abramson Cancer Center	7
1	Arthur M. Feldman	MD	(215) 696-3540	University of Pennsylvania-Abramson Cancer Center	51 N 39th St Ste 103APenn Presbyterian Medcl Ctr	Philadelphia, PA 19104-2640, US	[Breast Cancer, Geriatrics Oncology]	[Internal Medicine, Medical Oncology]	Abramson Cancer Center	1
2	David M. Mintzer	MD	Search for Phone Number	Abramson Cancer Center at Pennsylvania Hospital	230 W Washington Sq Fl 2	Philadelphia, PA 19106-3500, US	[Breast Cancer, Lung Cancer, Palliative Care/E...	[Hematology, Hospice and Palliative Medicine, ...	Abramson Cancer Center	13
3	Kristina Lynne Maletz Novick	MD	Search for Phone Number	Abramson Cancer Center	1425 Portland Ave	Rochester, NY 14621-3011, US	[]	[Radiation Oncology]	Abramson Cancer Center	0
4	Charles John Schneider	MD, FACP	Search for Phone Number	Hospital of the University of Pennsylvania, Ab...	3400 Civic Center BlvdPereleman Center for Adv...	Philadelphia, PA 19104-5127, US	[Clinical Research, Developmental Therapeutics...	[Medical Oncology]	Abramson Cancer Center	0

4. We then do some elementary EDA on the information that was compiled through scraping and the use of Entrez to make the onco_df data frame containing information about Oncologists.

- Research question : Is there a correlation between the number of oncologists in a cancer facility and the rankings or quality of the cancer institutes (as observed from the trends in research and rankings)

	center_name	counts
0	The University of Texas MD Anderson Cancer Center	133
1	Memorial Sloan-Kettering Cancer Center	109
2	Mayo Clinic Cancer Center	68
3	City of Hope Comprehensive Cancer Center	43
4	Massachusetts General Hospital	36



- University of Texas MD Anderson Cancer Center. Houston, TX 77030-4000. ...
- Memorial Sloan Kettering Cancer Center. 1-520-263-8939. ...
- Mayo Clinic. ...
- Dana-Farber/Brigham and Women's Cancer Center. ...
- UCLA Medical Center. ...
- Cleveland Clinic. ...
- City of Hope Comprehensive Cancer Center. ...
- Hospitals of the University of Pennsylvania-Penn Presbyterian.

More items...

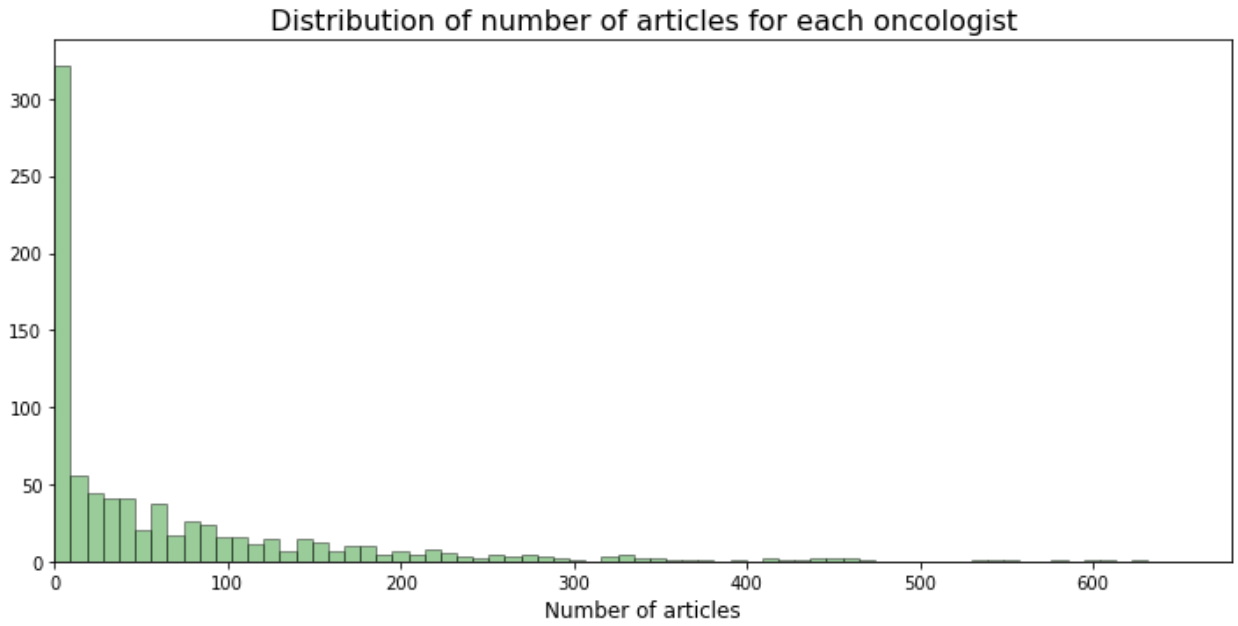
https://health.usnews.com › Best Hospitals › Rankings

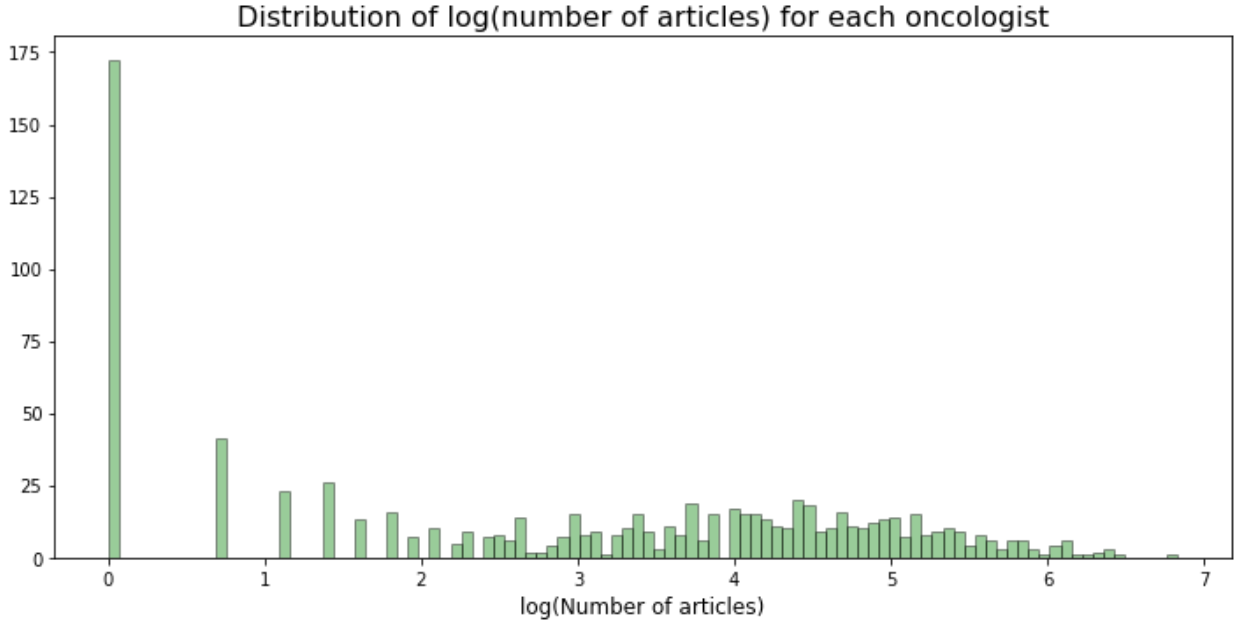
Best Hospitals for Cancer | Rankings & Ratings

```
print('Cancer centers with the most oncologists are:')
for i in range(5):
    print('Rank {}: {} has {} oncologists'.format(i+1, onco_counts_df.loc[i,'center_name'], onco_counts_df.loc[i,'counts']))
```

```
Cancer centers with the most oncologists are:
Rank 1: The University of Texas MD Anderson Cancer Center has 133 oncologists
Rank 2: Memorial Sloan-Kettering Cancer Center has 109 oncologists
Rank 3: Mayo Clinic Cancer Center has 68 oncologists
Rank 4: City of Hope Comprehensive Cancer Center has 43 oncologists
Rank 5: Massachusetts General Hospital has 36 oncologists
```

- Plotted the distribution tabulating the publications per doctor


Distribution of number of articles for each oncologist
300
250
200
150
100
50
0
0
100
200
300
400
500
600
Number of articles


Distribution of log(number of articles) for each oncologist
175
150
125
100
75
50
25
0
0
1
2
3
4
5
6
7
log(Number of articles)

```
onco_df=onco_df.sort_values(by=['article_num'],ascending=False)
onco_df.reset_index(inplace=True)
print('Oncologists with the most publications are:')
for i in range(5):
    print('Rank {}: {} has {} publications'.format(i+1, onco_df.loc[i,'name'], onco_df.loc[i,'article_num']))
```

```
Oncologists with the most publications are:
Rank 1: Jing Li has 2024 publications
Rank 2: Yu Chen has 929 publications
Rank 3: Farhad Ravandi has 627 publications
Rank 4: Gabriel N. Hortobagyi has 609 publications
Rank 5: Leslie L. Robison has 602 publications
```