Homework 9

Akshay Kumar Varanasi (av32826)

This homework is due on April 16, 2019 at 4:00pm. Please submit as a PDF file on Canvas. Before submission, please re-run all cells by clicking "Kernel" and selecting "Restart & Run All."

Problem 1 (2 points): Using Biopython and the Pubmed database, calculate the average number of papers per year that Dr. Wilke has published from 2015-2019 (inclusive, so that's 5 years total).

Hints: Dr. Wilke will always appear as "Wilke CO" in the Pubmed database. Also, make sure to set the `retmax` argument to at least `50` in `Entrez.esearch()` so that you retrieve all of the papers.

In [1]:
```
# You will need Entrez and Medline to solve this problem
from Bio import Entrez, Medline

Entrez.email = "akshayvaranasi@utexas.edu"

handle = Entrez.esearch(db="pubmed",  # database to search
                        term="Wilke CO[Author] AND 2015[Date - Publication]:201
9[Date - Publication]",  # search term
                        retmax=50  # number of results that are returned
                        )
record = Entrez.read(handle)
handle.close()

# search returns PubMed IDs (pmids)
pmid_list = record["IdList"]
print("Average number of papers per year that Dr. Wilke has published from 2015
-2019 is "+str(round(len(pmid_list)/5)))
```

```
Average number of papers per year that Dr. Wilke has published from 2015-2019 i
s 9
```

Problem 2 (4 points): From the years 2015-2019 (inclusive), how many different co-authors did Dr. Wilke publish with and how many times did Dr. Wilke publish a paper with each co-author? Print out each co-author and the number of times Dr. Wilke published a paper with that co-author. Make sure you don't print the same co-author's name twice.

Hint: In class 21, we parsed the results of a literature search with `Medline.parse()`. This allows us to look at the references we found and to retrieve different parts of the reference with a key. For example, to retrieve the abstract, we would write `record['AB']`. You can find a list of possible keys here (https://www.nlm.nih.gov/bsd/mms/medlineelements.html).

In [2]:
```
handle = Entrez.efetch(db="pubmed", id=pmid_list, rettype="medline", retmode="t
ext")
records = Medline.parse(handle)
coauthors = {} # start with empty dict of coauthors
for record in records:
    au_list = record['AU']
    for author in au_list:
        if author != "Wilke CO" and author in coauthors:
            coauthors[author]+=1
        if author != "Wilke CO" and author not in coauthors:
            coauthors[author]=1

        #else:
         #   coauthors[author]+=1
handle.close()
print('No.of co-authors are:',len(coauthors))
print('Co-authors of "Wilke CO" in 2015:2019 :')
for author in coauthors:
    print(" ", author," published ",coauthors[author], " times with Dr.Wilke")
```

```
No.of co-authors are: 92
Co-authors of "Wilke CO" in 2015:2019 :
  Jack BR  published  6  times with Dr.Wilke
  Teufel AI  published  6  times with Dr.Wilke
  Johnson MM  published  2  times with Dr.Wilke
  Laurent JM  published  3  times with Dr.Wilke
  Kachroo AH  published  3  times with Dr.Wilke
  Marcotte EM  published  6  times with Dr.Wilke
  Caglar MU  published  4  times with Dr.Wilke
  Hockenberry AJ  published  2  times with Dr.Wilke
  Ritchie AM  published  1  times with Dr.Wilke
  Liberles DA  published  1  times with Dr.Wilke
  Jewett MC  published  1  times with Dr.Wilke
  Amaral LAN  published  1  times with Dr.Wilke
  Paff ML  published  2  times with Dr.Wilke
  Smith BL  published  5  times with Dr.Wilke
  Bull JJ  published  2  times with Dr.Wilke
  Chen G  published  1  times with Dr.Wilke
  Krug RM  published  1  times with Dr.Wilke
  Wu DC  published  1  times with Dr.Wilke
  Yao J  published  1  times with Dr.Wilke
  Ho KS  published  1  times with Dr.Wilke
  Lambowitz AM  published  1  times with Dr.Wilke
  Sydykova DK  published  4  times with Dr.Wilke
  Spielman SJ  published  10  times with Dr.Wilke
  Jiang Q  published  1  times with Dr.Wilke
  Jackson EL  published  5  times with Dr.Wilke
  Tucker AT  published  1  times with Dr.Wilke
  Leonard SP  published  1  times with Dr.Wilke
  DuBois CD  published  1  times with Dr.Wilke
  Knauf GA  published  1  times with Dr.Wilke
  Cunningham AL  published  1  times with Dr.Wilke
  Trent MS  published  2  times with Dr.Wilke
  Davies BW  published  1  times with Dr.Wilke
  Guo F  published  1  times with Dr.Wilke
  Li S  published  1  times with Dr.Wilke
  Mao Z  published  1  times with Dr.Wilke
  Liu W  published  1  times with Dr.Wilke
  Woodman A  published  1  times with Dr.Wilke
  Arnold JJ  published  1  times with Dr.Wilke
  Huang TJ  published  1  times with Dr.Wilke
  Cameron CE  published  1  times with Dr.Wilke
  Boutz DR  published  4  times with Dr.Wilke
  Chapman SD  published  1  times with Dr.Wilke
  Adami C  published  1  times with Dr.Wilke
  B Kc D  published  1  times with Dr.Wilke
  Houser JR  published  2  times with Dr.Wilke
  Barnhart CS  published  1  times with Dr.Wilke
  Carroll SM  published  2  times with Dr.Wilke
  Dasgupta A  published  2  times with Dr.Wilke
  Lenoir WF  published  1  times with Dr.Wilke
  Sridhara V  published  3  times with Dr.Wilke
  Vander Wood D  published  1  times with Dr.Wilke
  Marx CJ  published  2  times with Dr.Wilke
  Barrick JE  published  3  times with Dr.Wilke
  Brown CW  published  1  times with Dr.Wilke
  Person MD  published  1  times with Dr.Wilke
  Echave J  published  4  times with Dr.Wilke
  Lipsitch M  published  1  times with Dr.Wilke
  Barclay W  published  1  times with Dr.Wilke
  Raman R  published  1  times with Dr.Wilke
  Russell CJ  published  1  times with Dr.Wilke
  Belser JA  published  1  times with Dr.Wilke
```

Problem 3 (4 points): From 2015-2019 (inclusive), how many of Dr. Wilke's papers contain the terms "evolution" or "evolutionary" in the abstract? Use python and **regular expressions** to find an answer.

Hint: In a regular expression, you can match the same word with slightly different endings using the "|" (or) operator. For example, the regex "bacteri(a|um)" would match both "bacteria" and "bacterium".

In [3]:
```
# You'll need the module re for regular expressions
import re
handle = Entrez.efetch(db="pubmed", id=pmid_list, rettype="medline", retmode="t
ext")
records = Medline.parse(handle)
count=0

for record in records:
    if 'AB' in record:
        abstract = record['AB']
        #print(record)
        match = re.search(r"\bevolutio(n\b|nary\b)",abstract)
        if match:
            count+=1
            #print(abstract)
            #print()



print("No.of papers which contain the terms \"evolution\" or \"evolutionary\" i
n the abstract is",count)
```

```
No.of papers which contain the terms "evolution" or "evolutionary" in the abstr
act is 26
```