Tutorial_1_Farjad_Ahmed

April 25, 2022

Farjad Ahmed - 1747371

This is the jupyter notebook for tutorial 1.

Please run the first cell to load libraries and generate the fileslist, then all other cells should work just fine.

[21]:
```
from collections import Counter
import glob
import codecs
import re

filelist = glob.glob("infect/*.txt")
```

[22]:
```
# HEARST 1
n=0
myList=[]
for files in filelist:
    file = codecs.open(files,'r','utf8')
    for line in file:
        result = re.search('(\w+)\ssuch\sas\s(\w+)\s((and|or)\s\w+)?',line)
        if result:
            myList.append(result.group(0))
            n+=1
print("Samples found for HEARST_1 are total: ", n, '\n')
for items in myList:
    print(items)
file.close()
```

```
Samples found for HEARST_1 are total:  176

phagocytes such as macrophages and neutrophils
profiles such as the
cells such as natural
diseases such as Type
CD4 such as macrophages and dendritic
groups such as IgG and IgA
diseases such as Chagas
Projects such as these
```

programs such as this
pathogens such as intestinal
tasks such as carrying
articles such as medical
measures such as personal
products such as hot
infections such as respiratory
tests such as determining
methods such as polymerase
Organizations such as the
resistance such as the
management such as prescribing
vectors such as urban
mosquito such as Haemogogus
mosquito such as Aedes
pathogens such as Staphylococcus
organisms such as Staphylococcus
factors such as total
patients such as those
means such as DNA
diseases such as smallpox or SARS
diseases such as typhoid and cholera
bacteria such as rickettsia and chlamydia
membranes such as a
structures such as protein
advantages such as resistance
diseases such as multiple
vectors such as when
factors such as host
Cells such as the
proteins such as vaccine
foods such as lactose
hosts such as humans
strains such as the
Methods such as gas
animals such as the
risks such as workplace
equipment such as harnesses and guardrails
problems such as the
countries such as the
projects such as Healthy
Programs such as these
agents such as viroids
yeasts such as Candida
nematodes such as parasitic
contact such as by
procedures such as injection or transplantation
plates such as these

dye such as Giemsa
contact such as airborne
prevention such as avoiding
membranes such as the
symptoms such as a
bacteria such as Streptococcus and Pseudomonas
issues such as shorter
viruses such as the
vectors such as mosquitoes or other
mitigation such as face
conditions such as cystic
measures such as finding and isolating
celebrities such as Taylor
conditions such as anxiety
products such as hand
sites such as the
vehicles such as the
landmarks such as the
germicide such as a
Insects such as mosquitoes and flies
criteria such as the
occurrences such as these
terms such as the
Mucolytics such as acetylcysteine and carbocystine
factors such as epidermal
bacteria such as Staphylococcus
pets such as dogs or cats
irritations such as ulcers or inflammation
diseases such as influenza
density such as train
causes such as benign
tumor such as pancreatic
factors such as hunger and poverty
countries such as the
diseases such as heart
Protozoa such as Giardia
Protozoans such as Giardia
antibiotics such as the
drugs such as loperamide and diphenoxylate
preservatives such as glycerol or kept
conditions such as cancer or diabetes
people such as the
syndromes such as hemophagocytic
factors such as nuclear
Cytokines such as tumor
sites such as respiratory
others such as those
arrhythmia such as torsades

elements such as the
phase such as the
authorities such as Landcom and the
wards such as Ward
coinfections such as those
conditions such as cardiovascular and respiratory
malaria such as cerebral
diseases such as Lyme
animals such as bats
changes such as deforestation
diseases such as Ebola
animals such as bats and rats
animals such as bats and birds
diseases such as African
practices such as improved
object such as a
Diseases such as emphysema and habits
bacteria such as Mycoplasma
antibiotics such as cefpodoxime
problems such as emphysema or heart
aminoglycoside such as gentamicin or tobramycin
rules such as the
antibiotics such as nitrofurantoin or trimethoprim
bacteria such as Enterococcus and Staphylococcus
others such as the
Complications such as ureteral
devices such as catheters and artificial
responses such as the
procedures such as those
rooms such as walls
diseases such as varicella and tuberculosis
flora such as staph
agents such as Ciprofloxacin
nonspecific such as fever and headache
arboviruses such as West
techniques such as polymerase
diseases such as AIDS and genital
disease such as diphtheria or measles
infections such as cellulitis
measures such as hand
precautions such as hand
sterilants such as glutaraldehydes or formaldehyde
materials such as Hepatitis
actions such as isolation
diseases such as the
viruses such as herpes
indicators such as length
agents such as pathogenic

```
factors such as contaminated
symptoms such as a
infections such as canine
air such as chemical
antibiotics such as methicillin and penicillin
safety such as avoiding
symptoms such as B
organisms such as the
drugs such as ciclosporin and azathioprine
models such as those
bacteria such as Mycobacterium
viruses such as measles and herpes
Complications such as pleural
medications such as amantadine or rimantadine
lactam such as cephazolin
considered such as a
diseases such as rheumatoid
colors such as bright
diseases such as hepatitis and poliomyelitis
methods such as travel
cohort such as a
system such as summarizing
groups such as women
organs such as the
```

[23]:
```
#HEARST 2
n=0
myList=[]
for files in filelist:
    file = codecs.open(files,'r','utf8')
    for line in file:
        result = re.search('(\w+)(,?)\sespecially\s(\w+)\s((and|or)\s\w+)?
 ↪',line)
        if result:
            myList.append(result.group(0))
            n+=1
print("Samples found for HEARST_2 are total: ", n, '\n')
for items in myList:
    print(items)
file.close()
```

```
Samples found for HEARST_2 are total:  67

is especially good
infections, especially in
more, especially with
or especially severe
administration, especially with
```

```
is especially true
eradication, especially for
partnerships, especially if
disease, especially in
infections, especially when
be especially beneficial
countries, especially in
but especially in
are especially associated
viruses, especially hepatitis
pylori, especially if
Africa, especially when
of especially plant
important, especially in
Drinks especially high
people, especially AIDS

enforced, especially in
fear especially if
fatalities, especially in
were especially high
health, especially when
OSHA, especially for
used, especially in
is especially important
animals, especially those
was especially necessary
is especially useful
prove especially useful
is especially infective or easily
panic, especially for
rights, especially in
is especially important
diseases, especially viral
risk, especially in
circumstances, especially in
are especially vulnerable
administered, especially if
are especially damaged
is especially common
physicians, especially after
furious, especially when
relationships, especially within
unique, especially the
vector, especially in
is especially susceptible
products, especially pork
obstruction, especially in
```

```
pathogens, especially those
is especially important
tests, especially those
months, especially in
resolve, especially in
be especially troublesome
disease, especially prevalent
stray, especially if
bat, especially in
sex, especially sexual
be especially troublesome
diseases, especially viruses
routinely, especially during
disease, especially in
people, especially those
```

[24]:
```
#HEARST 3
n=0
myList=[]
for files in filelist:
    file = codecs.open(files,'r','utf8')
    for line in file:
        result = re.search('(\w+)(,?)\sincluding\s(\w+)\s((and|or)\s\w+)?',line)
        if result:
            myList.append(result.group(0))
            n+=1
print("Samples found for HEARST_3 are total: ", n, '\n')
for items in myList:
    print(items)
file.close()
```

```
Samples found for HEARST_3 are total:  122

viruses, including HIV
association including studies
symptoms, including abdominal
packages including four or five
agents, including certain
agents, including epidemiologically
transmission, including proper
bacteria, including Klebsiella and Proteus
HIV, including previous
testing, including failure
antibiotics, including the
reasons including cost and regulation
wildlife, including the
mosquitoes including Haemagogus
infection including lung
```

fluid, including biological and environmental
body, including the
throat, including the
microorganisms, including bacteria and archaea
viruses, including influenza
families, including both
species including tomatoes and peppers
phytoplankton including harmful
syndrome including parasites and fungus
organisms, including Clostridium
science, including epidemiology and medicine
lifestyle, including their
environments, including the
infections, including environmental
biohazards, including animal
pathologies, including impetigo and strep
infections, including tuberculosis and meningitis
Ascomycota, including yeasts
Basidiomycota, including the
organisms, including microscopic
humans including Candida
bacteria including both
resuscitation including the
factors including a
locals, including 19
used, including the
all, including essential
and including complete
measures, including closing
services including special
activity, including motor and motor
patients, including those
Health, including the
days, including the
orders, including a
then, including for
voyage, including 36
antibiotics, including penicillin and methicillin
Toxins including Tropical
postulates, including viruses
microbes, including novel
tract, including the
individuals, including the
instruments, including balloons and baskets
peens, including half
groups, including Roman
symptoms including cough
cities, including 30

General including the
processes, including excessive
sepsis, including people
War, including the
them, including Matron
Avenue, including retaining
Avenue, including ornamental
setting, including adjacent
setting, including associated
setting, including sandstone
Group, including former
Coastline, including coastal
Cemetery including its
Site including Critical
elements, including rock
plans including retaining
time including separate
buildings, including Heffron
hospital, including Ward
personnel, including Dr
buildings, including the
features, including Pine
community, including the
evidence, including oral
institutions, including the
regions, including the
countries, including countries
populations including children and the
change, including the
ways including by
roundworm, including species
yeast, including those
Headaches, including migraines

causes including Streptococcus
infection, including bacterial
UTIs including acute
fluoroquinolones, including a
agents, including certain
agents, including epidemiologically
microorganisms, including bacteria and fungi
lyssaviruses, including the
lyssaviruses including the
extremities, including the
subspecies, including yaws
testing, including email and text
figures, including Franz
literature including John

```
rate, including specific
surfaces including medical
of including all
ailments including rheumatism and psoriasis
symptoms including reduced or alteration
organisms, including CDV and CAV
administration, including parenteral and intranasal
hazards, including needles
occasionally, including hantaviruses and coronaviruses
lungs, including Toxoplasma
pneumonias including SARS
factors, including the
```

[25]:
```
#HEARST 4
n=0
myList=[]
for f in filelist:
    file = codecs.open(f,'r','utf8')
    for line in file:
        result = re.search('(((\w+\s){1,2})|(\w+(,?
 ↪|\s)))and\sother\s(\w+\s){2,2}',line)
        if result:
            myList.append(result.group(0))
            n+=1
print("Samples found for HEARST_4 are total: ", n, '\n')
for items in myList:
    print(items)
file.close()
```

```
Samples found for HEARST_4 are total:  58

scabies and other ectoparasites and
Chromoblastomycosis and other deep mycoses
how NTDs and other diseases interact
pharmaceutical companies and other private and
monitors and other general hospital
to food and other necessities for
gonorrhea infection and other sexually transmitted
of membranes and other obstetrical complications
skin and other human microbiomes
AIDS and other forms of
feral dogs and other wild canine
bullets and other foreign bodies
in 1989 and other areas in
complex capsids and other structures on
of humans and other animals have
comprises cells and other mechanisms that
infect humans and other animals because
```

```
hot surfaces and other hazards with
Pesticides and other chemicals used
with foundries and other harmful types
customers and other stakeholders can
exchanging ideas and other different approaches
leaving home and other consequences of
of skin and other superficial structures
infected individuals and other interactions within
encounter insects and other animals harboring
Ships and other cargo carriers
spacecraft and other property returning
leaf browning and other issues in
He and other three people
of leprosy and other contagious diseases
school closures and other social distancing
of wood and other natural materials
at Marseille and other places in
on Wuhan and other major cities
air travel and other tourism to
in tropical and other communicable diseases

emerging diseases and other infectious and
to these and other diseases from
to this and other effects of
the lettuce and other uncooked ingredients
street vendors and other establishments where
campers and other outdoor recreationalists
fossil pollens and other microflora in
infects humans and other animals caused
the lung and other affected organs
noroviruses and other intestinal tract
the tonsils and other parts of
bloodstream and other parts of
prescribed antibiotics and other antimicrobial drugs
sickness behavior and other signs of
with syphilis and other sexually transmitted
monitors and other general hospital
adding tabs and other modifications to
segmental atelectasis and other severe side
borne pathogens and other diseases whenever
United Kingdom and other parts of
in syphilis and other various sexually
```

[26]:
```
#HEARST 5
n=0
myList=[]
for f in filelist:
```

```
    file = codecs.open(f,'r','utf8')
    for line in file:
        result = re.search('(\w+(,?))\sor\sother\s(\w+\s)',line)
        if result:
            myList.append(result.group(0))
            n+=1
print("Samples found for HEARST_5 are total: ", n, '\n')
for items in myList:
    print(items)
file.close()
```

```
Samples found for HEARST_5 are total:  26

cells or other allergic
medicines or other shopping

neurologic, or other disease
feces or other bodily
animal or other form
milk, or other body
fever or other highly
blood or other bodily
feces or other bodily
insects or other creatures
heroin or other opioid
water or other clear
pressure or other evidence
vaccine, or other means
failure or other types
HBV, or other blood
incontinence, or other discharges
hospital or other health
cats or other pets

blood or other potentially
saliva, or other bodily
people or other occupied
blood or other bodily
criteria or other diagnostic
drugs, or other medical
person or other organism
```