EDA - Exploratory Data Analysis

In this notebook, we will import data and do some simple data analysis like word-clouds

A high-level verbal description of your data:

The "scientific_lay_summarisation" dataset targets simplifying biomedical research articles into summaries understandable to non-specialists. It's part of a project to democratize access to scientific findings, featuring articles from PLOS and eLife journals. The dataset includes the full text of articles, section titles, keywords, article titles, publication years, and non-technical summaries. With 850.44 MB of dataset files, expanding to 1.32 GB upon generation.

In [31]:
```
!tlmgr install tcolorbox
```

```
texlive-scripts package not found (?!), skipping version consistency check
tlmgr: package repository https://ctan.mirror.rafal.ca/systems/texlive/tlnet
(not verified: gpg unavailable)
[1/1, ??:??/??:??] install: tcolorbox [229k]
running mktexlsr ...
done running mktexlsr.
tlmgr: package log updated: /Users/chloe/Library/TinyTeX/texmf-var/web2c/tlm
gr.log
tlmgr: command log updated: /Users/chloe/Library/TinyTeX/texmf-var/web2c/tlm
gr-commands.log
```

In [1]:
```
import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import numpy as np
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
```

In [2]:
```
# high level inspection
file_path = "./data/biolaysumm2024_data/"
file_names = ["eLife_train.jsonl", "eLife_val.jsonl", "eLife_test.jsonl",
              "PLOS_train.jsonl", "PLOS_val.jsonl", "PLOS_test.jsonl"
             ]

print("High level data inspection:")
print("=============================")
for filename in file_names:
    print("Processing file =", filename)
    df = pd.read_json(file_path+filename,
                      orient="records",
                      lines=True)
    print("Number of records =", len(df))

    # split by space for simple word count
    print("Counting words...")
```

```
    df["article_n_word"] = df["article"].apply(lambda text: len(text.split("
    if "lay_summary" in df.columns:
        df["summary_n_word"] = df["lay_summary"].apply(lambda text: len(text
        print("Overall description =\n", df[["article_n_word", "summary_n_wo
    else:
        print("Overall description =\n", df[["article_n_word"]].describe())

    # print a sample row
    k = 100
    item = df.iloc[k]
    print(f"Item {k}:")
    print(item)
    print("-----------------------------------------")

print("======= completed ========")
```

```
High level data inspection:
=============================
Processing file = eLife_train.jsonl
Number of records = 4346
Counting words...
Overall description =
        article_n_word  summary_n_word
count     4346.000000     4346.000000
mean     10159.277957      382.266222
std       3462.903717       64.334356
min        322.000000      177.000000
25%       7791.000000      338.000000
50%       9837.500000      379.000000
75%      12227.250000      423.000000
max      28308.000000      686.000000
Item 100:
lay_summary       Between birth and puberty , the bones of mamma...
article           Activating mutations in fibroblast growth fact...
headings          [Abstract, Introduction, Results, Discussion, ...
keywords                                 [developmental biology]
id                                                elife-31343-v2
article_n_word                                              6109
summary_n_word                                               406
Name: 100, dtype: object
-----------------------------------------
Processing file = eLife_val.jsonl
Number of records = 241
Counting words...
Overall description =
        article_n_word  summary_n_word
count      241.000000      241.000000
mean      9989.273859      389.875519
std       3275.141885       69.910844
min       3393.000000      234.000000
25%       7776.000000      338.000000
50%       9646.000000      384.000000
75%      11910.000000      441.000000
max      23050.000000      672.000000
Item 100:
lay_summary       Diseases of the heart and blood vessels are li...
article           Systemic vascular pressure in vertebrates is r...
headings          [Abstract, Introduction, Results, Discussion, ...
keywords                                              [medicine]
id                                                elife-28755-v1
article_n_word                                              6734
summary_n_word                                               379
Name: 100, dtype: object
-----------------------------------------
Processing file = eLife_test.jsonl
Number of records = 142
Counting words...
Overall description =
        article_n_word
count      142.000000
mean      8911.373239
std       2566.833437
```

```
min       2496.000000
25%       7355.750000
50%       8486.000000
75%      10537.000000
max      16884.000000
Item 100:
article           Replay , the sequential reactivation within a ...
headings          [Abstract, Introduction, Results, Discussion, ...
keywords                                             [neuroscience]
id                                                    elife-79031-v3
article_n_word                                                  7926
Name: 100, dtype: object
-----------------------------------------
Processing file = PLOS_train.jsonl
Number of records = 24773
Counting words...
Overall description =
        article_n_word  summary_n_word
count    24773.000000    24773.000000
mean      6750.888911      194.895935
std       2259.685682       36.820113
min        750.000000        4.000000
25%       5157.000000      174.000000
50%       6577.000000      202.000000
75%       8085.000000      218.000000
max      26647.000000      511.000000
Item 100:
lay_summary       CTCF is a transcriptional regulator acting as ...
article           Within the genomes of metazoans , nucleosomes ...
headings          [Abstract, Introduction, Results, Discussion, ...
keywords          [gene regulation, regulatory proteins, dna-bin...
id                                              journal.pgen.1005940
article_n_word                                                  7432
summary_n_word                                                   140
Name: 100, dtype: object
-----------------------------------------
Processing file = PLOS_val.jsonl
Number of records = 1376
Counting words...
Overall description =
        article_n_word  summary_n_word
count     1376.000000     1376.000000
mean      6738.800145      194.499273
std       2334.563171       36.594346
min        755.000000       55.000000
25%       5216.250000      173.000000
50%       6564.500000      202.000000
75%       8072.750000      217.000000
max      20394.000000      384.000000
Item 100:
lay_summary       Some genes perform necessary organismal functi...
article           In a classic example of the invasion of a spec...
headings          [Abstract, Introduction, Results, Discussion, ...
keywords          [united states, invertebrates, medicine and he...
id                                              journal.pgen.1005920
article_n_word                                                  3933
```

```
summary_n_word                                                    110
Name: 100, dtype: object
-----------------------------------------
Processing file = PLOS_test.jsonl
Number of records = 142
Counting words...
Overall description =
        article_n_word
count      142.000000
mean      6943.197183
std       2592.016390
min       1590.000000
25%       5250.750000
50%       6335.000000
75%       8316.000000
max      18481.000000
Item 100:
article           Bat-pollinated flowers have to attract their p...
headings          [Abstract, Introduction, Results, Discussion, ...
keywords          [amniotes, bats, bioacoustics, plant anatomy, ...
id                                          journal.pcbi.1009706
article_n_word                                                5837
Name: 100, dtype: object
-----------------------------------------
======= completed ========
```

EDA - Exploratory Data Analysis - eLife

In [3]:

```
filename = "./data/biolaysumm2024_data/eLife_train.jsonl"
train_df = pd.read_json(filename,
                       orient="records",
                       lines=True)
train_df.head()
```

Out[3]:

	lay_summary	article	headings	keywords	id
0	In the USA , more deaths happen in the winter ...	In temperate climates , winter deaths exceed s...	[Abstract, Introduction, Results, Discussion, ...	[epidemiology and global health]	elife-35500-v1
1	Most people have likely experienced the discom...	Whether complement dysregulation directly cont...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-48378-v2
2	The immune system protects an individual from ...	Variation in the presentation of hereditary im...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-04494-v1
3	The brain adapts to control our behavior in di...	Rapid and flexible interpretation of conflicti...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-12352-v2
4	Cells use motor proteins that to move organell...	Myosin 5a is a dual-headed molecular motor tha...	[Abstract, Introduction, Results, Discussion, ...	[structural biology and molecular biophysics]	elife-05413-v2

In [4]:
```
train_df.describe()
```

Out[4]:

	lay_summary	article	headings	keywords	id
count	4346	4346	4346	4346	4346
unique	4346	4346	105	296	4346
top	In the USA , more deaths happen in the winter ...	In temperate climates , winter deaths exceed s...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-35500-v1
freq	1	1	3484	753	1

In [5]:
```
item = train_df.iloc[0]
```

In [6]:
```
item
```

Out[6]:
```
lay_summary    In the USA , more deaths happen in the winter ...
article        In temperate climates , winter deaths exceed s...
headings       [Abstract, Introduction, Results, Discussion, ...
keywords                         [epidemiology and global health]
id                                                 elife-35500-v1
Name: 0, dtype: object
```

In [7]:
```
# take a look at the article (full text, domain-specific language)
print(f"Article len = {len(word_tokenize(item.article)):,} words")
item.article[:500]
```

```
Article len = 3,093 words
```

Out[7]:
```
'In temperate climates , winter deaths exceed summer ones . However , there
is limited information on the timing and the relative magnitudes of maximum
and minimum mortality , by local climate , age group , sex and medical caus
e of death . We used geo–coded mortality data and wavelets to analyse the s
easonality of mortality by age group and sex from 1980 to 2016 in the USA a
nd its subnational climatic regions . Death rates in men and women ≥ 45 yea
rs peaked in December to February and were lowest'
```

In [8]:
```
# take a look at the lay summary (shorter, normal language)
print(f"Lay summary len = {len(word_tokenize(item.lay_summary)):,} words")
item.lay_summary[:500]
```

```
Lay summary len = 357 words
```

Out[8]:
```
'In the USA , more deaths happen in the winter than the summer . But when d
eaths occur varies greatly by sex , age , cause of death , and possibly reg
ion . Seasonal differences in death rates can change over time due to chang
es in factors that cause disease or affect treatment . Analyzing the season
ality of deaths can help scientists determine whether interventions to mini
mize deaths during a certain time of year are needed , or whether existing
ones are effective . Scrutinizing seasonal patterns'
```

In [9]:
```
n = len(train_df)
print("Number of rows =", n)
max_row = 1000 # for dev debug, set to n for full set

train_df_sample = train_df.sample(n=max_row) # take a random sample of the c
print("New number of rows =", len(train_df_sample))
```

```
Number of rows = 4346
New number of rows = 1000
```

In [10]:
```
# make word counter for article and lay_summary

train_df_sample["article_token_count"] = train_df_sample.article.apply(lambd
train_df_sample["summary_token_count"] = train_df_sample.lay_summary.apply(l

train_df_sample.head()
```

Out[10]:

	lay_summary	article	headings	keywords	id	article_token_
3726	Neurons constantly talk to each other by sendi...	Synaptic membrane-remodeling events such as en...	[Abstract, Introduction, Results, Discussion, ...	[cell biology, neuroscience]	elife-69597-v1	
2910	Our understanding of the living world has been...	Among various advantages , their small size ma...	[Abstract, Introduction, Results, Discussion, ...	[plant biology]	elife-01567-v1	
3607	The cerebellum is a region of the brain that i...	Rapid firing of cerebellar Purkinje neurons is...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-04193-v2	
2855	Getting older increases our risk of experienci...	Most age-related human diseases are accompanie...	[Abstract, Introduction, Results, Discussion, ...	[genetics and genomics]	elife-68610-v3	
4181	Healthy human cells employ many tricks to avoi...	Tumor suppressor p53 prevents cell transformat...	[Abstract, Introduction, Results, Discussion, ...	[cancer biology]	elife-26129-v3	

In [11]:

```
train_df_sample[['article_token_count', 'summary_token_count']].describe()
```

Out[11]:

	article_token_count	summary_token_count
count	1000.000000	1000.000000
mean	10132.453000	382.901000
std	3495.217593	61.516205
min	1806.000000	204.000000
25%	7727.000000	343.000000
50%	9819.500000	381.000000
75%	12210.250000	424.000000
max	27488.000000	556.000000

In [12]:

```
# confirm these numbers
print("Article avg word count =", np.average(train_df_sample.article_token_c
print("Summary avg word count =", np.average(train_df_sample.summary_token_c
```

```
Article avg word count = 10132.453
Summary avg word count = 382.901
```

Word-clouds

We are making some word clouds to see the overall frequent words

In [13]:
```
stopwords_en = stopwords.words("english")
```

In [14]:
```
def make_word_cloud(df=train_df, text_col="article", bigrams = True, bigram_
    """
        Create wordcloud to see which dish names appear frequently
        Parameters:
            restaurant_type: "ch" or "en". None to create wordcloud in gener
    """
    text = df[text_col] # build wordcloud for entire dataset

    print("Text counts = ", len(text))
    # print(text[0])
    text = " ".join(text)
    wcld = WordCloud(stopwords=stopwords_en,
                     collocations=bigrams,
                     collocation_threshold=bigram_threshold
                    )
    wordcloud = wcld.generate(text)


    # show wordcloud
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.show()

    # process text, sort by word count and get top words
    w_counter = wcld.process_text(text)
    result = sorted(w_counter.items(),
                    key=lambda item: item[1],
                    reverse=True)

    # print(result)
    return result
```

In [15]:
```
# flatten keywords list -> string, replace " " -> "_" for compound words
train_df_sample["keywords_flat"] = train_df_sample.keywords.apply(lambda wor


train_df_sample.head()
```

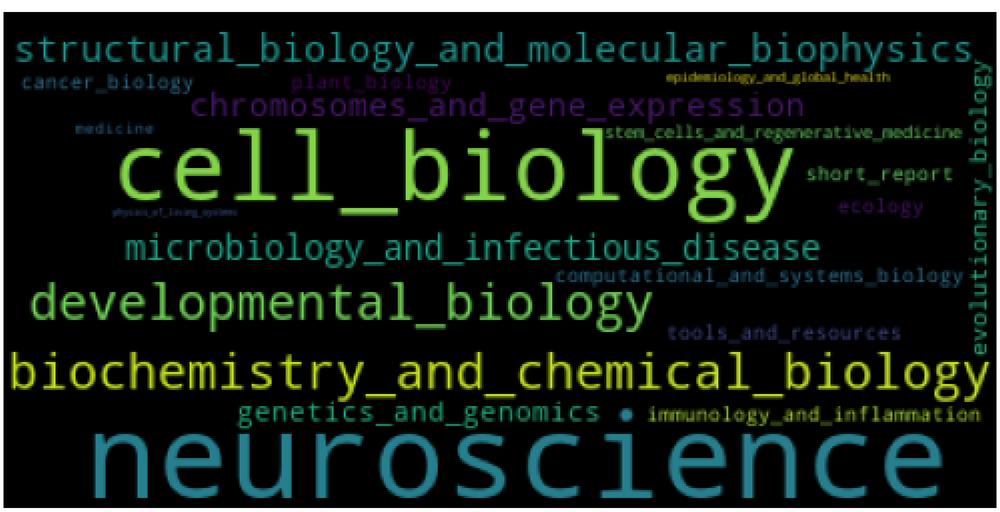
Out[15]:

	lay_summary	article	headings	keywords	id	article_token_
3726	Neurons constantly talk to each other by sendi...	Synaptic membrane-remodeling events such as en...	[Abstract, Introduction, Results, Discussion, ...	[cell biology, neuroscience]	elife-69597-v1	
2910	Our understanding of the living world has been...	Among various advantages , their small size ma...	[Abstract, Introduction, Results, Discussion, ...	[plant biology]	elife-01567-v1	
3607	The cerebellum is a region of the brain that i...	Rapid firing of cerebellar Purkinje neurons is...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-04193-v2	
2855	Getting older increases our risk of experienci...	Most age-related human diseases are accompanie...	[Abstract, Introduction, Results, Discussion, ...	[genetics and genomics]	elife-68610-v3	
4181	Healthy human cells employ many tricks to avoi...	Tumor suppressor p53 prevents cell transformat...	[Abstract, Introduction, Results, Discussion, ...	[cancer biology]	elife-26129-v3	

In [16]:

```
top_words_keyword = make_word_cloud(df=train_df_sample, text_col="keywords_f
# convert "_" back to " " for better readability
top_words_keyword = [(w.replace("_", " "), count)
                     for w, count in top_words_keyword
                    ]
top_words_keyword[:50]
```

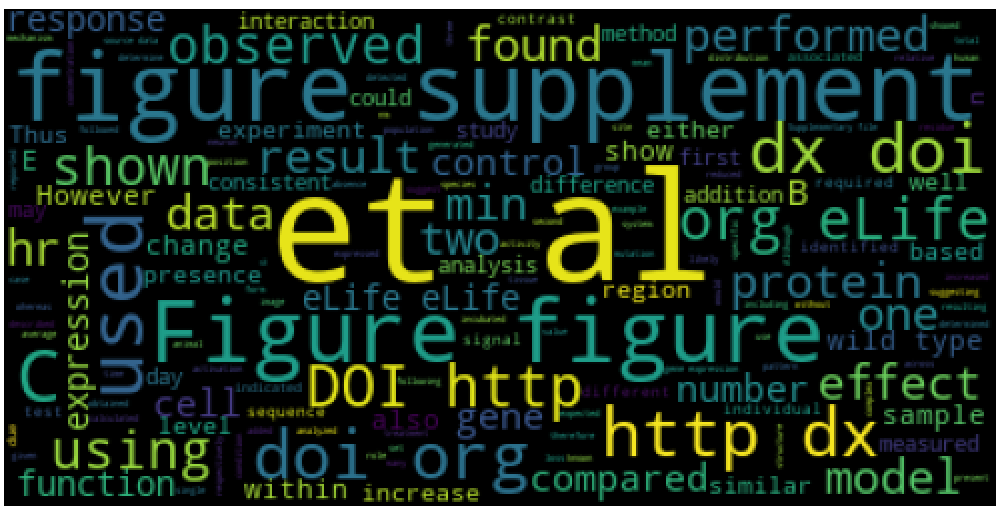
```
Text counts =  1000
```



Out[16]:
```
[('neuroscience', 307),
 ('cell biology', 188),
 ('biochemistry and chemical biology', 127),
 ('developmental biology', 121),
 ('structural biology and molecular biophysics', 105),
 ('microbiology and infectious disease', 97),
 ('chromosomes and gene expression', 83),
 ('genetics and genomics', 70),
 ('evolutionary biology', 61),
 ('computational and systems biology', 59),
 ('immunology and inflammation', 54),
 ('short report', 50),
 ('tools and resources', 45),
 ('ecology', 45),
 ('cancer biology', 40),
 ('plant biology', 37),
 ('stem cells and regenerative medicine', 35),
 ('medicine', 23),
 ('epidemiology and global health', 17),
 ('physics of living systems', 12),
 ('research communication', 4)]
```

In [17]:
```python
# WARNING: LONG PROCESS (a few minutes)
top_words_atc = make_word_cloud(df=train_df_sample, text_col="article", bigr
top_words_atc[:50]
```

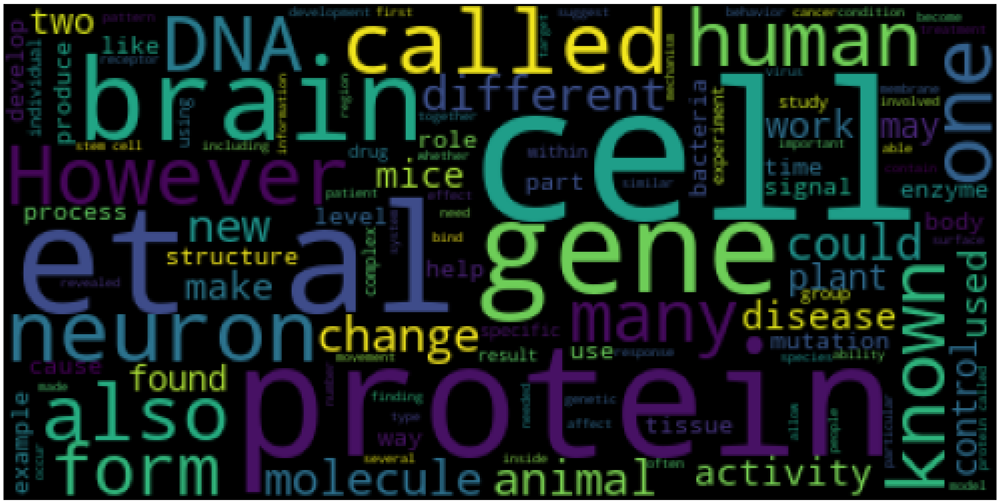
```
Text counts =  1000
```



```
Out[17]: [('et al', 80969),
 ('figure supplement', 20396),
 ('Figure figure', 16771),
 ('used', 12631),
 ('C', 10827),
 ('doi org', 9354),
 ('dx doi', 9337),
 ('http dx', 9336),
 ('org eLife', 9320),
 ('DOI http', 9055),
 ('using', 7828),
 ('observed', 7773),
 ('shown', 7766),
 ('protein', 6266),
 ('model', 6156),
 ('found', 6150),
 ('min', 6082),
 ('one', 6024),
 ('effect', 5933),
 ('result', 5929),
 ('hr', 5873),
 ('two', 5862),
 ('performed', 5610),
 ('data', 5545),
 ('compared', 5514),
 ('cell', 5476),
 ('control', 5438),
 ('number', 5425),
 ('gene', 5406),
 ('eLife eLife', 5334),
 ('expression', 5241),
 ('function', 5207),
 ('B', 5124),
 ('response', 5050),
 ('wild type', 4992),
 ('also', 4937),
 ('change', 4929),
 ('sample', 4754),
 ('However', 4750),
 ('show', 4638),
 ('E', 4622),
 ('within', 4597),
 ('experiment', 4552),
 ('n', 4353),
 ('level', 4328),
 ('presence', 4283),
 ('increase', 4255),
 ('similar', 4213),
 ('based', 3993),
 ('Thus', 3956)]
```

```
In [18]: top_words_summ = make_word_cloud(df=train_df_sample, text_col="lay_summary",
         top_words_summ[:50]
```

```
Text counts =  1000
```


cell
protein
et al
gene
brain
DNA
called
human
one
However
neuron
known
many
also
form
molecule
animal
different
change
disease
activity
could
used
control
new
mice
make
found
plant
two
work
may
mutation
tissue
structure
process
bacteria
level
signal
enzyme
body
role
part
time
like
drug
help
use
way
example
produce
develop
cause
individual
complex
study
experiment
result

```
Out[18]: [('cell', 3361),
 ('protein', 2158),
 ('et al', 1920),
 ('gene', 1322),
 ('brain', 1052),
 ('called', 961),
 ('neuron', 900),
 ('one', 831),
 ('However', 791),
 ('also', 764),
 ('known', 758),
 ('human', 736),
 ('DNA', 695),
 ('many', 683),
 ('form', 676),
 ('change', 654),
 ('animal', 651),
 ('different', 625),
 ('new', 612),
 ('could', 593),
 ('molecule', 588),
 ('activity', 576),
 ('make', 560),
 ('disease', 544),
 ('mice', 542),
 ('used', 538),
 ('work', 534),
 ('two', 526),
 ('may', 488),
 ('control', 466),
 ('found', 465),
 ('plant', 448),
 ('structure', 435),
 ('cause', 425),
 ('use', 423),
 ('time', 422),
 ('mutation', 408),
 ('process', 406),
 ('bacteria', 406),
 ('like', 401),
 ('role', 398),
 ('signal', 386),
 ('body', 385),
 ('help', 378),
 ('level', 375),
 ('produce', 371),
 ('enzyme', 370),
 ('tissue', 362),
 ('develop', 357),
 ('example', 357)]
```

EDA - Exploratory Data Analysis - PLOS

In [19]:
```
filename = "./data/biolaysumm2024_data/PLOS_train.jsonl"
print("Reading from file =", filename)
train_df = pd.read_json(filename,
                        orient="records",
                        lines=True)
train_df.head()
```

```
Reading from file = ./data/biolaysumm2024_data/PLOS_train.jsonl
```

Out[19]:

	lay_summary	article	headings	keywords	i
0	In the kidney , structures known as nephrons a...	Kidney function depends on the nephron , which...	[Abstract, Introduction, Results, Discussion, ...	[developmental biology, danio (zebrafish), ver...	journal.pgen.003018
1	Many species of bats in North America have bee...	White-nose syndrome is one of the most lethal ...	[Abstract, Introduction, Results, Discussion, ...	[sequencing techniques, fungal spores, vertebr...	journal.ppat.100607
2	The burden of dengue has been increasing over ...	Sustainable dengue intervention requires the p...	[Abstract, Introduction, Methods, Results, Dis...	[invertebrates, medicine and health sciences, ...	journal.pntd.000749
3	Estrogen exposure is the most important risk f...	Despite the central role of estrogen exposure ...	[Abstract, Introduction, Results, Discussion, ...	[oncology/breast cancer, oncology/gynecologica...	journal.pgen.100101
4	Melioidosis is a severe tropical infection cau...	Macrophage migration inhibitory factor (MIF)...	[Abstract, Introduction, Methods, Results, Dis...	[immunology/cellular microbiology and pathogen...	journal.pntd.000060

In [20]:
```
train_df.describe()
```

Out[20]:

	lay_summary	article	headings	keywords	id
count	24773	24773	24773	24773	24773
unique	24771	24771	517	19674	24773
top	The collective movement of animals in a group ...	Inference of interaction rules of animals movi...	[Abstract, Introduction, Results, Discussion, ...	[]	journal.pgen.0030189
freq	2	2	9345	3471	1

In [21]:
```
item = train_df.iloc[0]# take a look at the article (full text, domain-speci
print(f"Article len = {len(word_tokenize(item.article)):,} words")
item.article[:500]
```

```
Article len = 10,085 words
```

Out[21]:
```
'Kidney function depends on the nephron , which comprises a blood filter ,
a tubule that is subdivided into functionally distinct segments , and a col
lecting duct . How these regions arise during development is poorly underst
ood . The zebrafish pronephros consists of two linear nephrons that develop
from the intermediate mesoderm along the length of the trunk . Here we show
that , contrary to current dogma , these nephrons possess multiple proximal
and distal tubule domains that resemble the orga'
```

In [22]:
```
# take a look at the lay summary (shorter, normal language)
print(f"Lay summary len = {len(word_tokenize(item.lay_summary)):,} words")
item.lay_summary[:500]
```

```
Lay summary len = 233 words
```

Out[22]:
```
"In the kidney , structures known as nephrons are responsible for collectin
g metabolic waste . Nephrons are composed of a blood filter ( glomerulus )
followed by a series of specialized tubule regions , or segments , which re
cover solutes such as salts , and finally terminate with a collecting duct
. The genetic mechanisms that establish nephron segmentation in mammals hav
e been a challenge to study because of the kidney's complex organogenesis .
The zebrafish embryonic kidney ( pronephros ) cont"
```

In [23]:
```
n = len(train_df)
print("Number of rows =", n)
max_row = 5000 # for dev debug, set to n for full set
# train_df = train_df.iloc[:max_row]

train_df_sample = train_df.sample(n=max_row) # take a random sample of the d
print("New number of rows =", len(train_df_sample))
```

```
Number of rows = 24773
New number of rows = 5000
```

In [24]:
```
# make token counter for article and lay_summary
# WARNING : LONG PROCESS (a few minutes)
train_df_sample["article_token_count"] = train_df_sample.article.apply(lambd
train_df_sample["summary_token_count"] = train_df_sample.lay_summary.apply(l

train_df_sample.head()
```

Out[24]:

	lay_summary	article	headings	keywords	
16404	Selective attention can enhance processing of ...	Selective attention supports the prioritized p...	[Abstract, Introduction, Results, Discussion, ...	[medicine and health sciences, engineering and...	journal.pb
5913	Many fungal plant pathogens undergo a series o...	Phytopathogens secrete effector proteins to ma...	[Abstract, Introduction, Results, Discussion, ...	[plant science, plant biology, plant pathology...	journal.pp
7624	This paper describes how the use of three drug...	Public health interventions based on distribut...	[Abstract, Introduction, Methods, Results, Res...	[]	journal.pr
13177	Plasmodium falciparum is responsible for the m...	The process of erythrocyte invasion by merozoi...	[Abstract, Introduction, Results, Discussion, ...	[biochemistry, infectious diseases, cell biolo...	journal.pp
850	Protein pyrabactin resistance 1 (PYR1) belon...	The pyrabactin resistance 1 (PYR1) /PYR1-lik...	[Abstract, Introduction, Results/Discussion, M...	[biomacromolecule-ligand interactions, physics...	journal.p

In [25]:

```
train_df_sample[['article_token_count', 'summary_token_count']].describe()
```

Out[25]:

	article_token_count	summary_token_count
count	5000.000000	5000.000000
mean	7051.242600	194.917200
std	2359.226263	37.610345
min	1110.000000	4.000000
25%	5400.750000	174.000000
50%	6852.500000	203.000000
75%	8486.750000	219.000000
max	20586.000000	453.000000

In [26]:

```
# confirm these numbers
print("Article avg word count =", np.average(train_df_sample.article_token_c
print("Summary avg word count =", np.average(train_df_sample.summary_token_c
```

```
Article avg word count = 7051.2426
Summary avg word count = 194.9172
```

Word-clouds

We are making some word clouds to see the overall frequent words

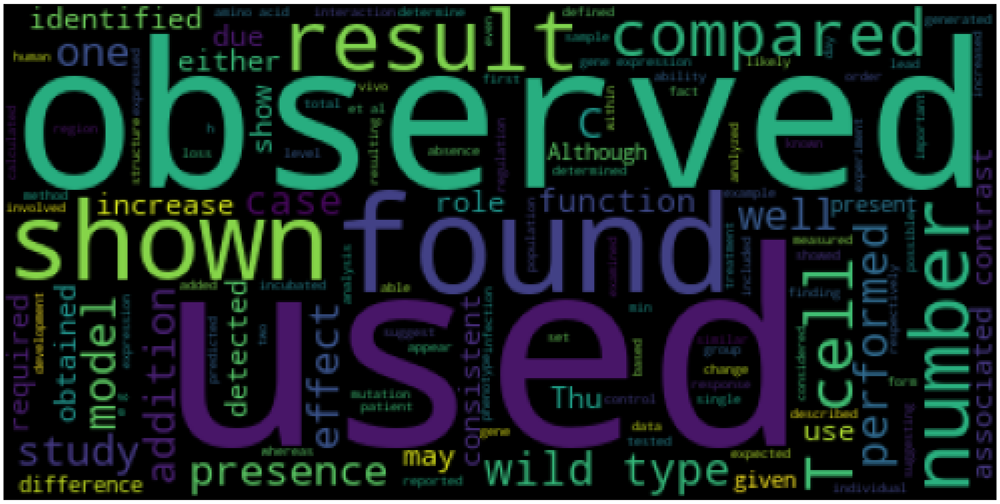
In [27]:

```
def make_word_cloud(df=train_df, text_col="article", bigrams = True, bigram_
    """
        Create wordcloud to see which dish names appear frequently
        Parameters:
            restaurant_type: "ch" or "en". None to create wordcloud in gener
    """
    text = df[text_col] # build wordcloud for entire dataset

    print("Text counts = ", len(text))
    # print(text[0])
    text = " ".join(text)
    wcld = WordCloud(stopwords=STOPWORDS,
                     collocations=bigrams,
                     collocation_threshold=bigram_threshold
                    )
    wordcloud = wcld.generate(text)


    # show wordcloud
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.show()

    # process text, sort by word count and get top words
    w_counter = wcld.process_text(text)
    result = sorted(w_counter.items(),
                    key=lambda item: item[1],
                    reverse=True)

    # print(result)
    return result
```

In [28]:

```
# WARNING: LONG PROCESS (a few minutes)
top_words_atc = make_word_cloud(df=train_df, text_col="article")
top_words_atc[:50]
```

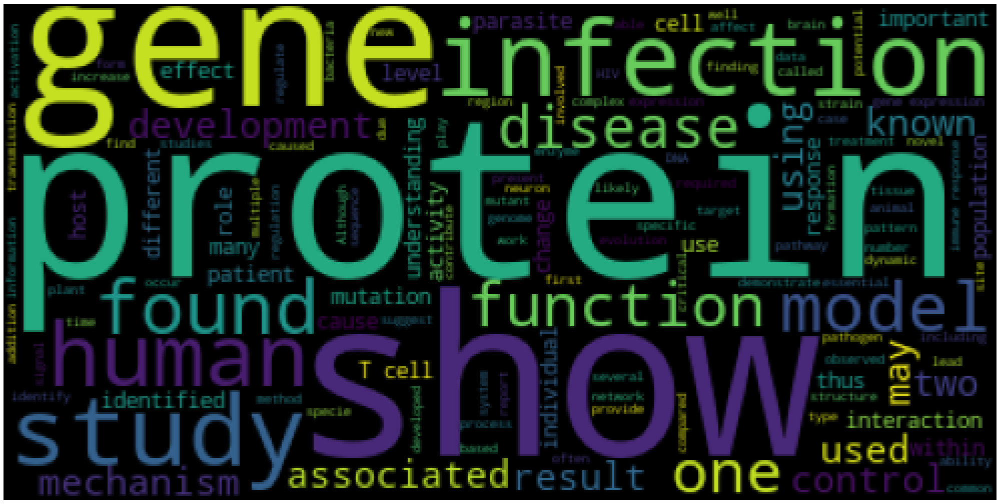
```
Text counts =  24773
```


observed
used
found
shown
result
compared
number
T cell
wild type
presence
performed
addition
effect
model
study
well
function
identified
one
either
due
case
increase
role
associated
contrast
obtained
required
detected
show
consistent
Although
Thu
may
use
given
difference
present
c

Out[28]:
```
[('used', 192971),
 ('observed', 125548),
 ('found', 123781),
 ('shown', 110827),
 ('result', 101601),
 ('number', 92995),
 ('compared', 86908),
 ('T cell', 86129),
 ('wild type', 84695),
 ('performed', 84163),
 ('model', 84156),
 ('study', 78772),
 ('C', 76800),
 ('presence', 76389),
 ('effect', 75909),
 ('well', 73170),
 ('one', 71378),
 ('addition', 71333),
 ('case', 65302),
 ('function', 65040),
 ('due', 59725),
 ('identified', 58945),
 ('required', 58496),
 ('role', 57679),
 ('increase', 56075),
 ('contrast', 55396),
 ('Thu', 55136),
 ('associated', 54487),
 ('consistent', 53923),
 ('obtained', 53388),
 ('use', 52516),
 ('show', 52502),
 ('may', 52253),
 ('either', 51641),
 ('detected', 51510),
 ('Although', 51467),
 ('difference', 50604),
 ('present', 50560),
 ('given', 49750),
 ('gene expression', 49413),
 ('absence', 48466),
 ('known', 47883),
 ('described', 47555),
 ('similar', 47432),
 ('change', 46718),
 ('interaction', 46433),
 ('individual', 46257),
 ('min', 46219),
 ('within', 45443),
 ('example', 45059)]
```

In [29]:
```
top_words_summ = make_word_cloud(df=train_df, text_col="lay_summary")
top_words_summ[:50]
```

```
Text counts =  24773
```


gene
protein
show
infection
study
human
found
function
model
disease
development
known
mechanism
associated
result
one
control
used
two
may
using
response
population
identified
interaction
mutation
patient
many
role
host
different
individual
understanding
activity
change
cause
T cell
thus
within
important
parasite
cell
effect
level
use

```
Out[29]: [('protein', 8642),
 ('show', 8141),
 ('gene', 8063),
 ('infection', 7485),
 ('study', 6996),
 ('human', 6819),
 ('found', 6709),
 ('model', 5878),
 ('function', 5734),
 ('one', 5711),
 ('disease', 4933),
 ('development', 4809),
 ('used', 4687),
 ('result', 4648),
 ('may', 4591),
 ('associated', 4585),
 ('using', 4386),
 ('known', 4367),
 ('mechanism', 4348),
 ('two', 4086),
 ('control', 3919),
 ('important', 3886),
 ('response', 3759),
 ('use', 3690),
 ('cell', 3652),
 ('role', 3631),
 ('identified', 3628),
 ('cause', 3515),
 ('many', 3506),
 ('interaction', 3476),
 ('host', 3445),
 ('mutation', 3384),
 ('change', 3379),
 ('patient', 3376),
 ('effect', 3367),
 ('parasite', 3348),
 ('different', 3291),
 ('population', 3287),
 ('understanding', 3246),
 ('activity', 3159),
 ('within', 3153),
 ('T cell', 3099),
 ('thus', 3086),
 ('level', 3071),
 ('individual', 3046),
 ('gene expression', 3040),
 ('specific', 3012),
 ('demonstrate', 2901),
 ('expression', 2880),
 ('region', 2870)]
```