Assignment 2

CS 5154 – Information Retrieval

# Results of running IR2A initially

## Code (after adding utf-8 as encoding):

# IR2A.py CS5154/6054 cheng 2022  
# read lines from a text file  
# tokenize and count words  
# use WordCloud to show words with top frequencies  
# Usage: python IR2A.py bible.txt  
  
import re  
import sys  
from collections import Counter  
from wordcloud import WordCloud  
from matplotlib import pyplot as plt  
  
f = open(sys.argv[1], 'r', encoding="utf8")  
   
counter = Counter()  
for t in f:  
 counter.update(re.findall('\w+', t))  
  
wc = WordCloud()  
wc.generate\_from\_frequencies(counter)  
plt.imshow(wc)  
plt.axis("off")  
plt.show()

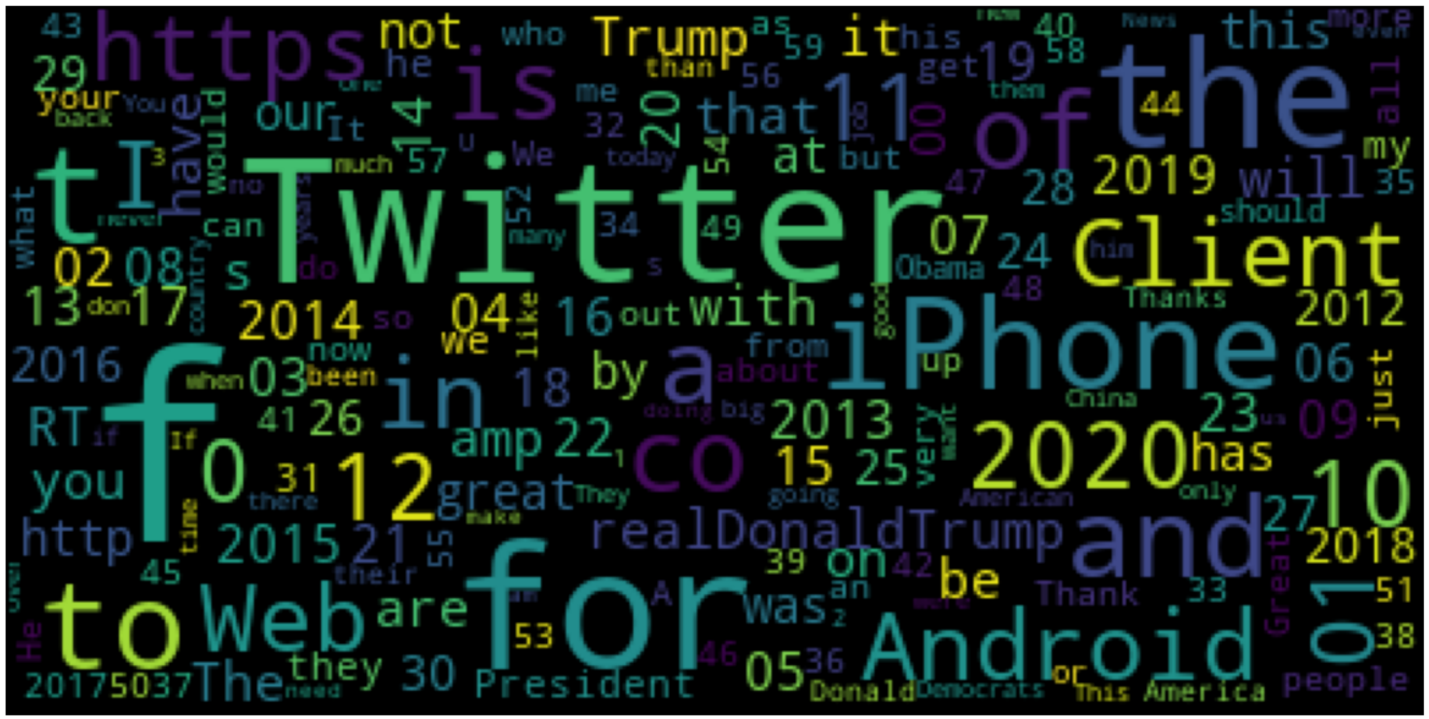
## WordCloud Plots:

### Gutprotocol.txt

Text

Description automatically generated

### tweet\_01-08-2021.py



### Hair-bearing\_skin\_generation.txt



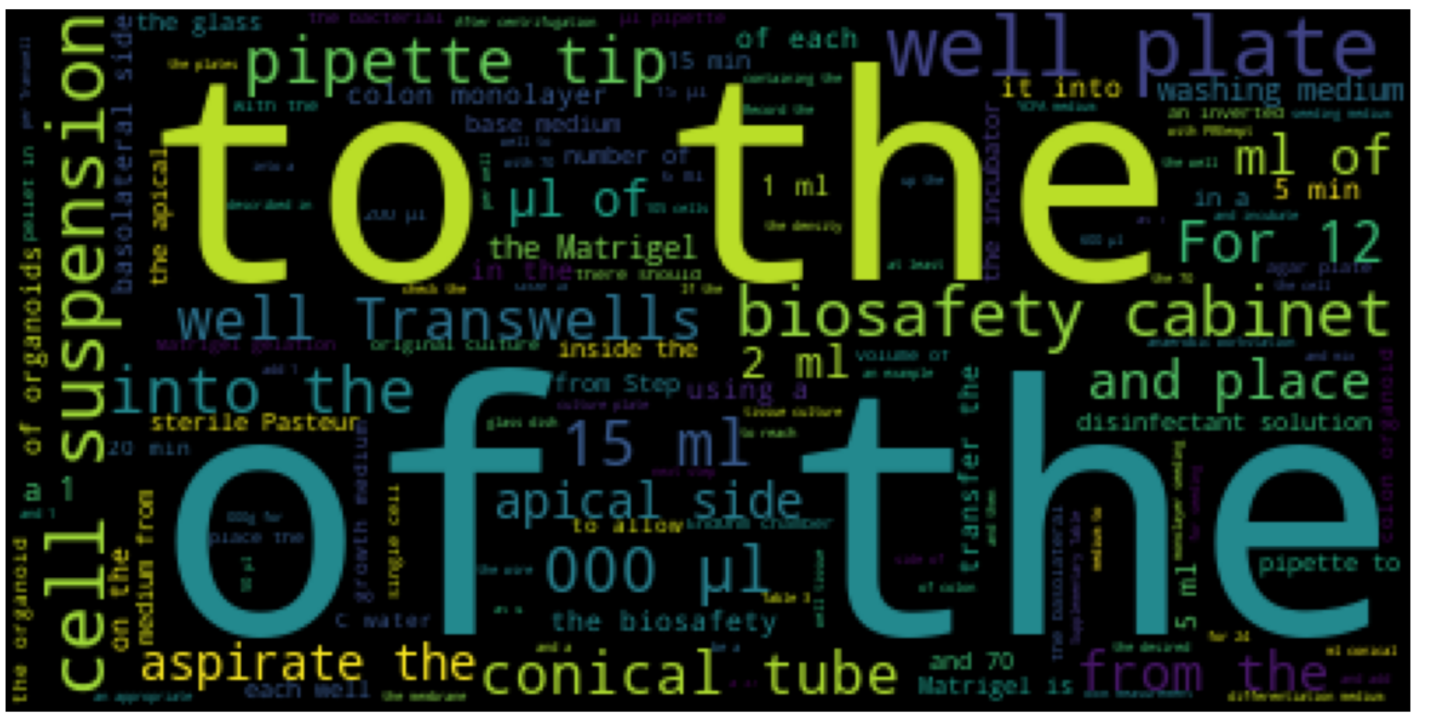
# Results for bigrams

## Code (after updating regex for bigrams)

# IR2A.py CS5154/6054 cheng 2022  
# read lines from a text file  
# tokenize and count words  
# use WordCloud to show words with top frequencies  
# Usage: python IR2A.py bible.txt  
  
import re  
import sys  
from collections import Counter  
from wordcloud import WordCloud  
from matplotlib import pyplot as plt  
  
f = open(sys.argv[1], 'r', encoding="utf8")  
   
counter = Counter()  
for t in f:  
 counter.update(re.findall('\w+\s\w+', t))  
  
wc = WordCloud()  
wc.generate\_from\_frequencies(counter)  
plt.imshow(wc)  
plt.axis("off")  
plt.show()

## Results:

### Gutprotocol.txt



### tweet\_01-08-2021.py

Text

Description automatically generated

### Hair-bearing\_skin\_generation.txt

