Question 3

In [17]:
```
# Setup and define function for parsing
from tika import parser
def convertPdf2TxtWithTika(in_pdf_file, out_text_file):
    # Load a file and extract information
    print ("INFO: - reading file = " + in_pdf_file)

    raw = parser.from_file(in_pdf_file)
    text = raw['content']

    ## Post-processing explained at:
    # https://medium.com/@justinboylantoomey/fast-text-extraction-with-python-and-tika-41ac34b0fe61
    # Convert to string
    text = str(text)
    # Ensure text is utf-8 formatted
    safe_text = text.encode('utf-8', errors='ignore')
    # Escape any \ issues
    safe_text = str(safe_text).replace('\\', '\\\\').replace('"', '\\"')

    # Write out extracted content
    text_pdf = open(out_text_file, 'w')
    print ("INFO: - writing file = " + out_text_file)
    text_pdf.write(text)
    text_pdf.close()
```

In [18]:
```
# Find pdf files in a directory and process its content
import os
import glob

txtfiles = []
inpath = 'resume/'
outpath = 'out/'
count = 0
for file in glob.glob(inpath + '*.pdf'):
    justfile = os.path.basename(file)
    justfile = justfile.replace(".pdf","")
    print('INFO: processing file = ' + os.path.basename(file))
    output_file = outpath + justfile + '.txt'
    print('INFO: - in = ' + file + ', out = ' + output_file)
    convertPdf2TxtWithTika(file, output_file)
    count = count + 1
print('INFO: processed total files = ' + str(count))
```

```
INFO: processing file = Thien_Le_Resume.pdf
INFO: - in = resume/Thien_Le_Resume.pdf, out = out/Thien_Le_Resume.txt
INFO: - reading file = resume/Thien_Le_Resume.pdf
INFO: - writing file = out/Thien_Le_Resume.txt
INFO: processing file = Rowen_Burney_Resume.pdf
INFO: - in = resume/Rowen_Burney_Resume.pdf, out = out/Rowen_Burney_Resume.txt
INFO: - reading file = resume/Rowen_Burney_Resume.pdf
INFO: - writing file = out/Rowen_Burney_Resume.txt
INFO: processing file = Resume - Duayne Wright Jr.pdf
INFO: - in = resume/Resume - Duayne Wright Jr.pdf, out = out/Resume - Duayne Wright Jr.txt
INFO: - reading file = resume/Resume - Duayne Wright Jr.pdf
INFO: - writing file = out/Resume - Duayne Wright Jr.txt
INFO: processing file = Jacob T. Frierson Resume.pdf
INFO: - in = resume/Jacob T. Frierson Resume.pdf, out = out/Jacob T. Frierson Resume.txt
INFO: - reading file = resume/Jacob T. Frierson Resume.pdf
INFO: - writing file = out/Jacob T. Frierson Resume.txt
INFO: processing file = Belcher__Taylor_Resume (3).pdf
INFO: - in = resume/Belcher__Taylor_Resume (3).pdf, out = out/Belcher__Taylor_Resume (3).txt
INFO: - reading file = resume/Belcher__Taylor_Resume (3).pdf
INFO: - writing file = out/Belcher__Taylor_Resume (3).txt
INFO: processing file = CS Resume.pdf
INFO: - in = resume/CS Resume.pdf, out = out/CS Resume.txt
INFO: - reading file = resume/CS Resume.pdf
INFO: - writing file = out/CS Resume.txt
INFO: processing file = Dudley Hallie MSBA Resume final June 2024.pdf
INFO: - in = resume/Dudley Hallie MSBA Resume final June 2024.pdf, out = out/Dudley Hallie MSBA Resume final Jun
e 2024.txt
INFO: - reading file = resume/Dudley Hallie MSBA Resume final June 2024.pdf
INFO: - writing file = out/Dudley Hallie MSBA Resume final June 2024.txt
INFO: processing file = Resume_Protik_Nag.pdf
INFO: - in = resume/Resume_Protik_Nag.pdf, out = out/Resume_Protik_Nag.txt
INFO: - reading file = resume/Resume_Protik_Nag.pdf
INFO: - writing file = out/Resume_Protik_Nag.txt
INFO: processing file = Jordan_Wood_Resume.pdf
INFO: - in = resume/Jordan_Wood_Resume.pdf, out = out/Jordan_Wood_Resume.txt
INFO: - reading file = resume/Jordan_Wood_Resume.pdf
```

```
INFO: - writing file = out/Jordan_Wood_Resume.txt
INFO: processing file = resume-nicholas-miklaucic.pdf
INFO: - in = resume/resume-nicholas-miklaucic.pdf, out = out/resume-nicholas-miklaucic.txt
INFO: - reading file = resume/resume-nicholas-miklaucic.pdf
INFO: - writing file = out/resume-nicholas-miklaucic.txt
INFO: processing file = BCS Resume.pdf
INFO: - in = resume/BCS Resume.pdf, out = out/BCS Resume.txt
INFO: - reading file = resume/BCS Resume.pdf
INFO: - writing file = out/BCS Resume.txt
INFO: processing file = MalikSiddResume2024.pdf
INFO: - in = resume/MalikSiddResume2024.pdf, out = out/MalikSiddResume2024.txt
INFO: - reading file = resume/MalikSiddResume2024.pdf
INFO: - writing file = out/MalikSiddResume2024.txt
INFO: processing file = Resume(4-2-2024).pdf
INFO: - in = resume/Resume(4-2-2024).pdf, out = out/Resume(4-2-2024).txt
INFO: - reading file = resume/Resume(4-2-2024).pdf
INFO: - writing file = out/Resume(4-2-2024).txt
INFO: processing file = Ritvik_G.pdf
INFO: - in = resume/Ritvik_G.pdf, out = out/Ritvik_G.txt
INFO: - reading file = resume/Ritvik_G.pdf
INFO: - writing file = out/Ritvik_G.txt
INFO: processing file = Resume_Burkholder,Eric.pdf
INFO: - in = resume/Resume_Burkholder,Eric.pdf, out = out/Resume_Burkholder,Eric.txt
INFO: - reading file = resume/Resume_Burkholder,Eric.pdf
INFO: - writing file = out/Resume_Burkholder,Eric.txt
INFO: processing file = JohnAydin-2024-Resume.pdf
INFO: - in = resume/JohnAydin-2024-Resume.pdf, out = out/JohnAydin-2024-Resume.txt
INFO: - reading file = resume/JohnAydin-2024-Resume.pdf
INFO: - writing file = out/JohnAydin-2024-Resume.txt
INFO: processing file = Resume_Nafisa_Mehtaj.pdf
INFO: - in = resume/Resume_Nafisa_Mehtaj.pdf, out = out/Resume_Nafisa_Mehtaj.txt
INFO: - reading file = resume/Resume_Nafisa_Mehtaj.pdf
INFO: - writing file = out/Resume_Nafisa_Mehtaj.txt
INFO: processing file = Francis_Resume_24.pdf
INFO: - in = resume/Francis_Resume_24.pdf, out = out/Francis_Resume_24.txt
INFO: - reading file = resume/Francis_Resume_24.pdf
INFO: - writing file = out/Francis_Resume_24.txt
INFO: processing file = AndyWaters-Resume2024 - 08.22.24.pdf
INFO: - in = resume/AndyWaters-Resume2024 - 08.22.24.pdf, out = out/AndyWaters-Resume2024 - 08.22.24.txt
INFO: - reading file = resume/AndyWaters-Resume2024 - 08.22.24.pdf
INFO: - writing file = out/AndyWaters-Resume2024 - 08.22.24.txt
INFO: processing file = August 2024 Ryan Karbowniczak Resume.pdf
INFO: - in = resume/August 2024 Ryan Karbowniczak Resume.pdf, out = out/August 2024 Ryan Karbowniczak Resume.txt
INFO: - reading file = resume/August 2024 Ryan Karbowniczak Resume.pdf
INFO: - writing file = out/August 2024 Ryan Karbowniczak Resume.txt
INFO: processing file = Resume - Eli Bryson.pdf
INFO: - in = resume/Resume - Eli Bryson.pdf, out = out/Resume - Eli Bryson.txt
INFO: - reading file = resume/Resume - Eli Bryson.pdf
INFO: - writing file = out/Resume - Eli Bryson.txt
INFO: processing file = Nayeem Mohammad.pdf
INFO: - in = resume/Nayeem Mohammad.pdf, out = out/Nayeem Mohammad.txt
INFO: - reading file = resume/Nayeem Mohammad.pdf
INFO: - writing file = out/Nayeem Mohammad.txt
INFO: processing file = Khan Waleed Resume.pdf
INFO: - in = resume/Khan Waleed Resume.pdf, out = out/Khan Waleed Resume.txt
INFO: - reading file = resume/Khan Waleed Resume.pdf
INFO: - writing file = out/Khan Waleed Resume.txt
INFO: processing file = JoshuaKolbuszResume.pdf
INFO: - in = resume/JoshuaKolbuszResume.pdf, out = out/JoshuaKolbuszResume.txt
INFO: - reading file = resume/JoshuaKolbuszResume.pdf
INFO: - writing file = out/JoshuaKolbuszResume.txt
INFO: processing file = tylerbeasley_resume.pdf
INFO: - in = resume/tylerbeasley_resume.pdf, out = out/tylerbeasley_resume.txt
INFO: - reading file = resume/tylerbeasley_resume.pdf
INFO: - writing file = out/tylerbeasley_resume.txt
INFO: processing file = Murphy_Keenan_resume_copy.pdf
INFO: - in = resume/Murphy_Keenan_resume_copy.pdf, out = out/Murphy_Keenan_resume_copy.txt
INFO: - reading file = resume/Murphy_Keenan_resume_copy.pdf
INFO: - writing file = out/Murphy_Keenan_resume_copy.txt
INFO: processing file = TrevorSeestedt_Resume.pdf
INFO: - in = resume/TrevorSeestedt_Resume.pdf, out = out/TrevorSeestedt_Resume.txt
INFO: - reading file = resume/TrevorSeestedt_Resume.pdf
INFO: - writing file = out/TrevorSeestedt_Resume.txt
INFO: processing file = Résumé Zak Elguindi.pdf
INFO: - in = resume/Résumé Zak Elguindi.pdf, out = out/Résumé Zak Elguindi.txt
INFO: - reading file = resume/Résumé Zak Elguindi.pdf
INFO: - writing file = out/Résumé Zak Elguindi.txt
INFO: processing file = Resume_Kolipaka,Pranavi.pdf
INFO: - in = resume/Resume_Kolipaka,Pranavi.pdf, out = out/Resume_Kolipaka,Pranavi.txt
INFO: - reading file = resume/Resume_Kolipaka,Pranavi.pdf
INFO: - writing file = out/Resume_Kolipaka,Pranavi.txt
INFO: processed total files = 29
```

In [19]:
```
# Now we define a function to do word cloud
from wordcloud import WordCloud,STOPWORDS
import matplotlib.pyplot as plt
%matplotlib inline

def wordcloud_draw(data, color = 'black'):
    words = ' '.join(data)
    cleaned_word = " ".join([word for word in words.split()
                             if 'http' not in word
                                 and not word.startswith('@')
                                 and not word.startswith('#')
                                 and word != 'RT'
                             ])
    wordcloud = WordCloud(stopwords=STOPWORDS,
                          background_color=color,
                          width=2500,
                          height=2000
                         ).generate(cleaned_word)
    plt.figure(1,figsize=(13, 13))
    plt.imshow(wordcloud)
    plt.axis('off')
    plt.show()
```

Resume Exercise - Programming: Word Processing

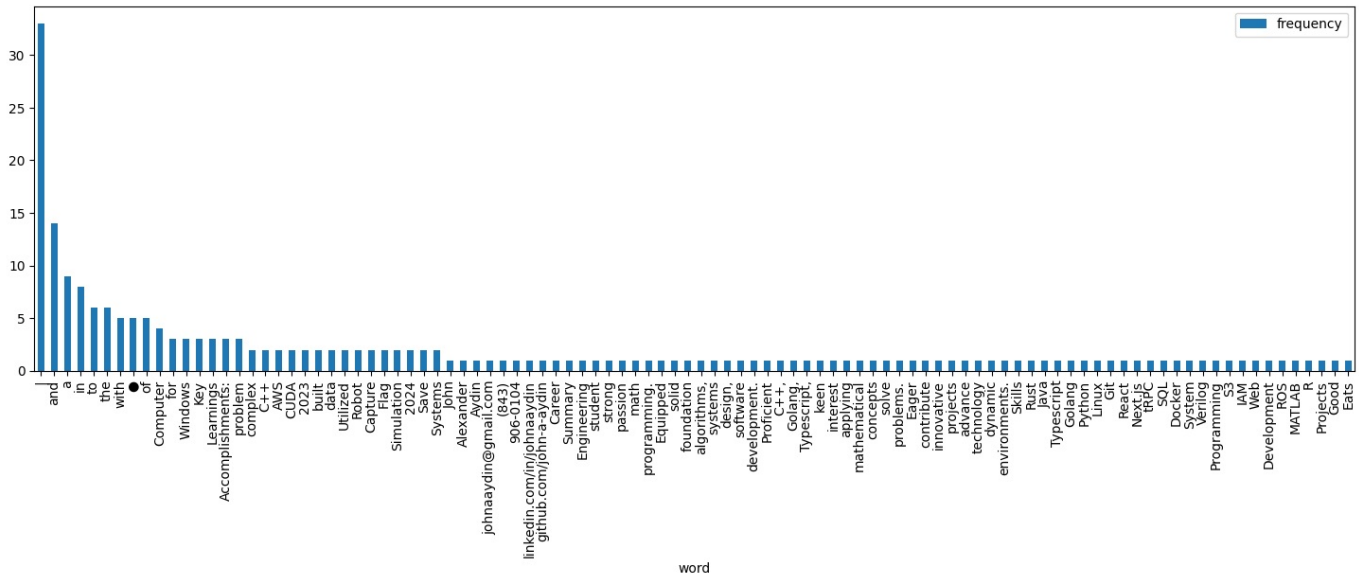
1

In [20]:
```
# Get content in all files into one string
file = 'out/JohnAydin-2024-Resume.txt'
file_handle = open(file, 'r')
content = str( file_handle.read()).split()
#content_as_str = " ".join(sorted(set(content), key=content.index))
# Has duplicates
my_content_as_str = " ".join(content)
```

2

In [21]:
```
import pandas as pd
from collections import Counter
cnt = Counter(my_content_as_str.split())
top100 = cnt.most_common(100)
df = pd.DataFrame(top100, columns=['word', 'frequency'])
df.plot(kind='bar', x='word', figsize=(18, 5))
```

Out[21]:
```
<Axes: xlabel='word'>
```

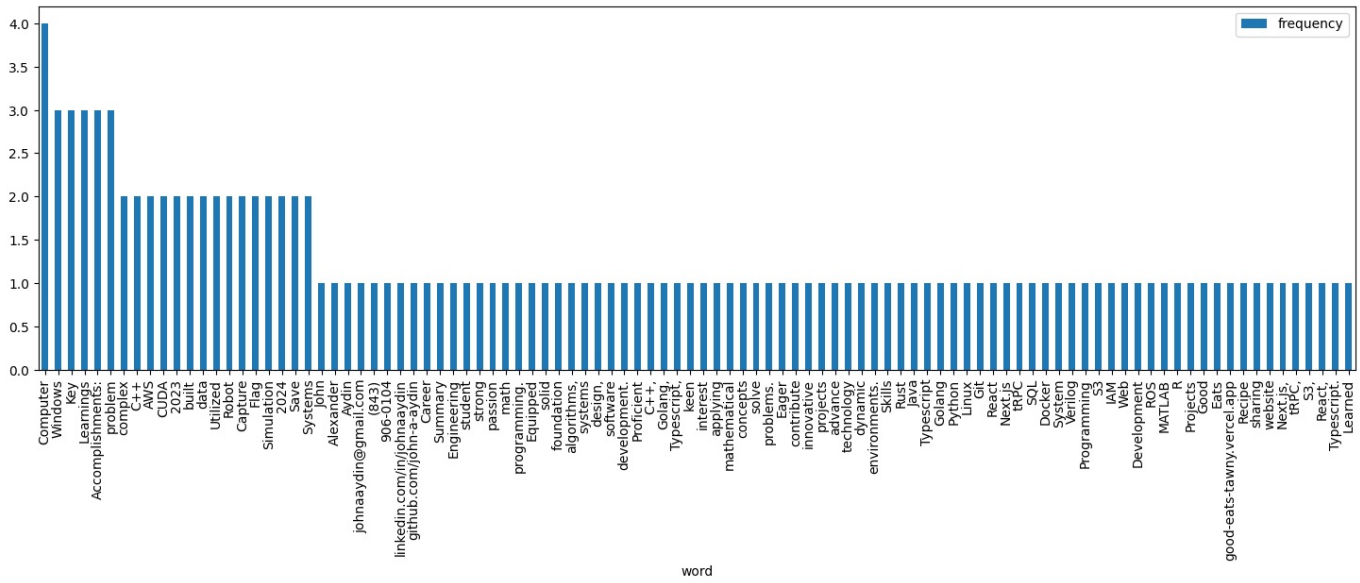


3-5

In [22]:
```
ignore = ['|', 'and', 'a', 'in', 'to', 'the', 'with', '●', 'of', 'for']
for word in list(cnt):
    if word in ignore:
        del cnt[word]
top100 = cnt.most_common(100)
df = pd.DataFrame(top100, columns=['word', 'frequency'])
```

```
df.plot(kind='bar', x='word', figsize=(18, 5))
```

Out[22]: <Axes: xlabel='word'>



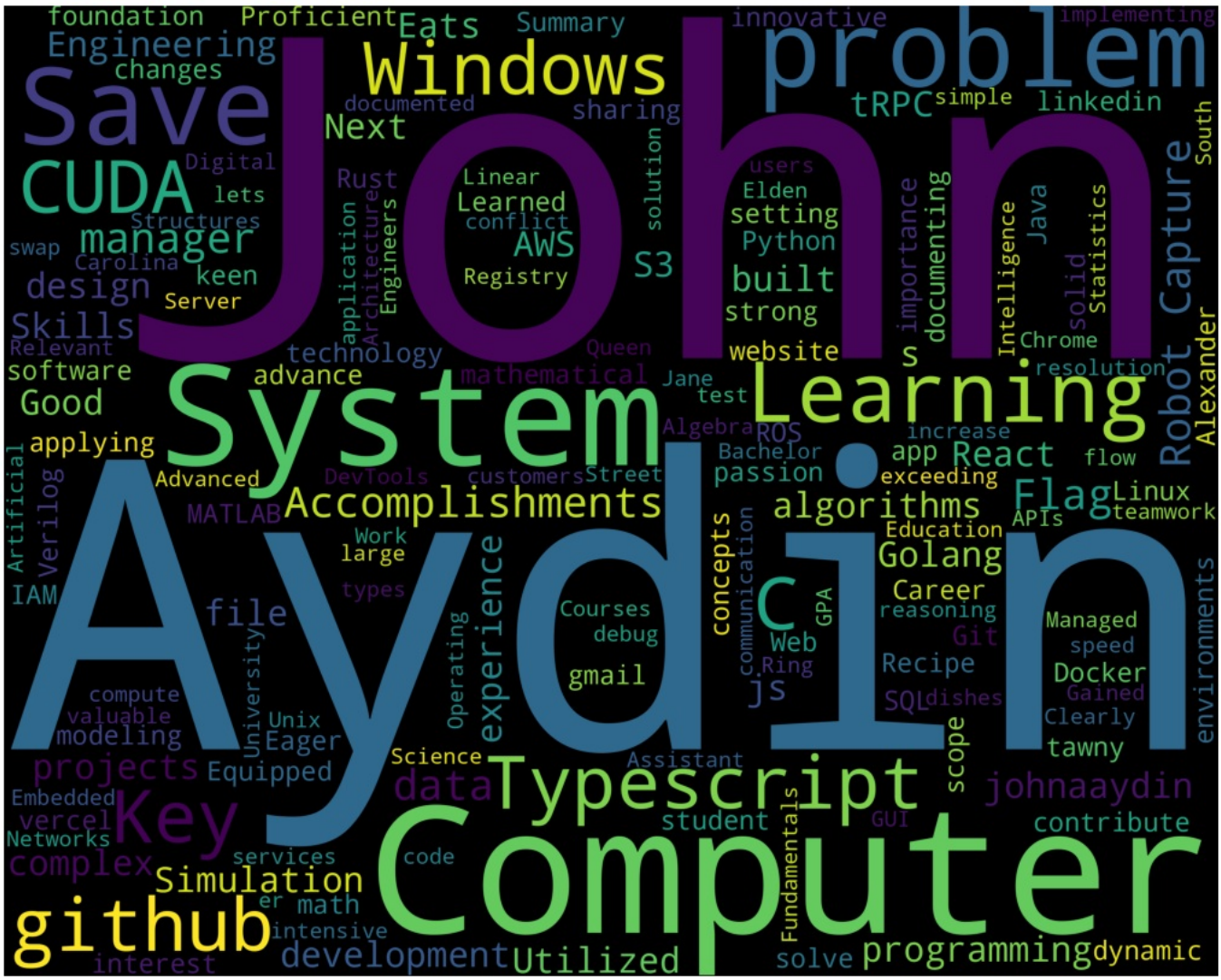
6

The frequencies are much lower and fairly evenly distributed with most only appering once. This is probalbly because I try fit as many key words as I can in my resume while not sounding repetative.

Resume Exercise - Programming: Word Tag Cloud

Task 1

In [23]:
```
# Now do word tag cloud
wordcloud_draw(my_content_as_str.split())
```



Task 2

In [24]:

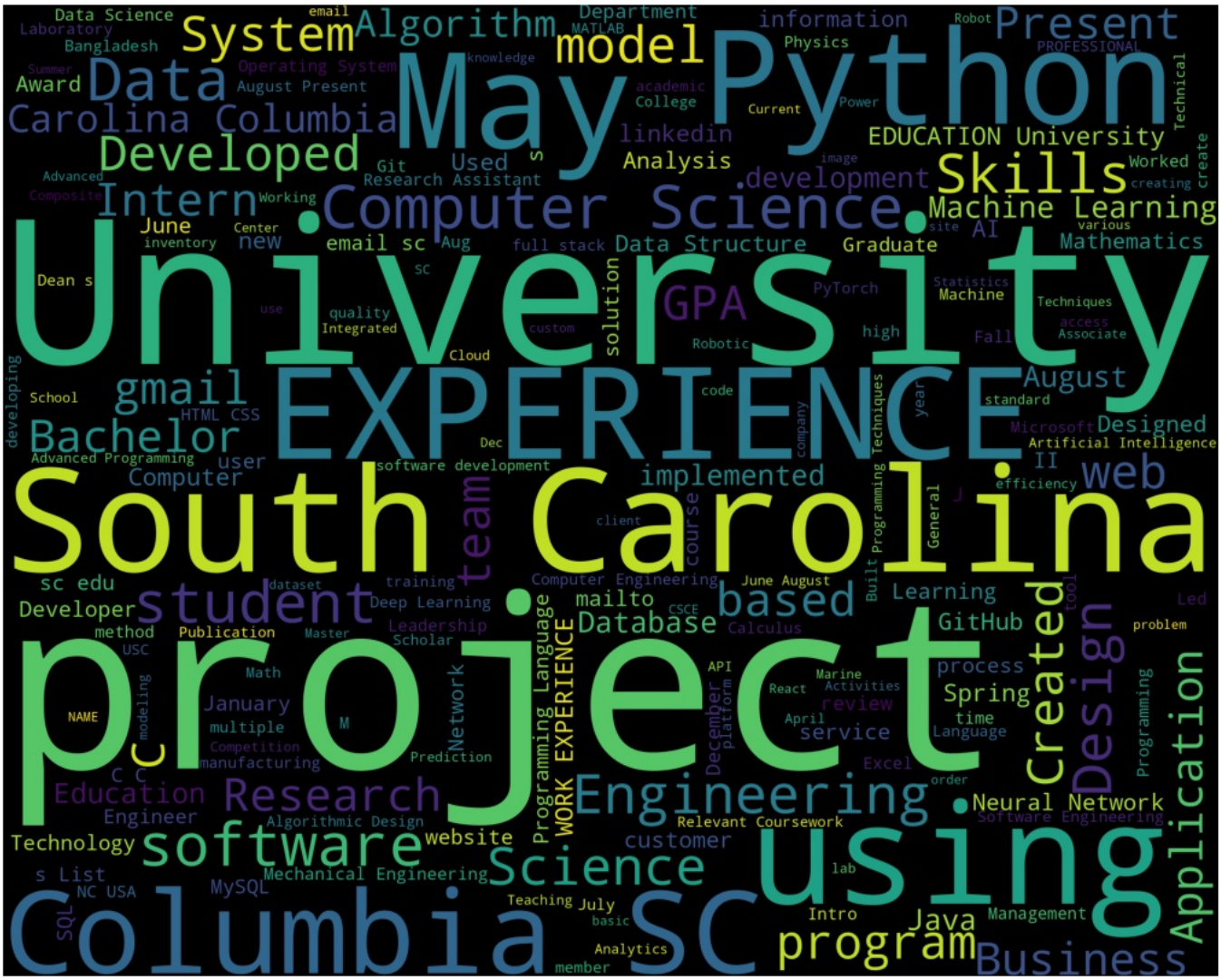
```
# Get content in all files into one string
pathFilesToUse = 'out/'

all_content = ''
count = 0
for file in glob.glob(pathFilesToUse + '*.txt'):
    print("file = " + file)
    file_handle = open(file, 'r')
    content = str( file_handle.read()).split()
    #content_as_str = " ".join(sorted(set(content), key=content.index))
    # Has duplicates
    content_as_str = " ".join(content)
    # All together
    all_content = all_content + content_as_str
    #all_content.append(content)
    count = count + 1
    file_handle.close()

print('INFO: processed total files = ' + str(count))
```

```
file = out/Resume_Kolipaka,Pranavi.txt
file = out/August 2024 Ryan Karbowniczak Resume.txt
file = out/Thien_Le_Resume.txt
file = out/Resume(4-2-2024).txt
file = out/JoshuaKolbuszResume.txt
file = out/Rowen_Burney_Resume.txt
file = out/Jacob T. Frierson Resume.txt
file = out/Resume_Nafisa_Mehtaj.txt
file = out/Résumé Zak Elguindi.txt
file = out/Murphy_Keenan_resume_copy.txt
file = out/AndyWaters-Resume2024 - 08.22.24.txt
file = out/BCS Resume.txt
file = out/Ritvik_G.txt
file = out/JohnAydin-2024-Resume.txt
file = out/Resume_Protik_Nag.txt
file = out/Belcher__Taylor_Resume (3).txt
file = out/tylerbeasley_resume.txt
file = out/Resume - Duayne Wright Jr.txt
file = out/Dudley Hallie MSBA Resume final June 2024.txt
file = out/Nayeem Mohammad.txt
file = out/CS Resume.txt
file = out/Khan Waleed Resume.txt
file = out/Resume_Burkholder,Eric.txt
file = out/Resume - Eli Bryson.txt
file = out/resume-nicholas-miklaucic.txt
file = out/TrevorSeestedt_Resume.txt
file = out/MalikSiddResume2024.txt
file = out/Jordan_Wood_Resume.txt
file = out/Francis_Resume_24.txt
INFO: processed total files = 29
```

In [25]:
```
# Now do word tag cloud
wordcloud_draw(all_content.split())
```



Observation

The coolest thing to me is seeing my name be the most frequent word in my word cloud since all of my socials include my name in their url. Another thing I noticed is that in my word cloud you can see the individual technologies that I put on my resume, but when looking at the whole class, everyone's unique experiences are drowned out by general words like technology or development. It's also interesting to

see that enough people mentioned SQL and Git on their resume for it to show on the class's word cloud.