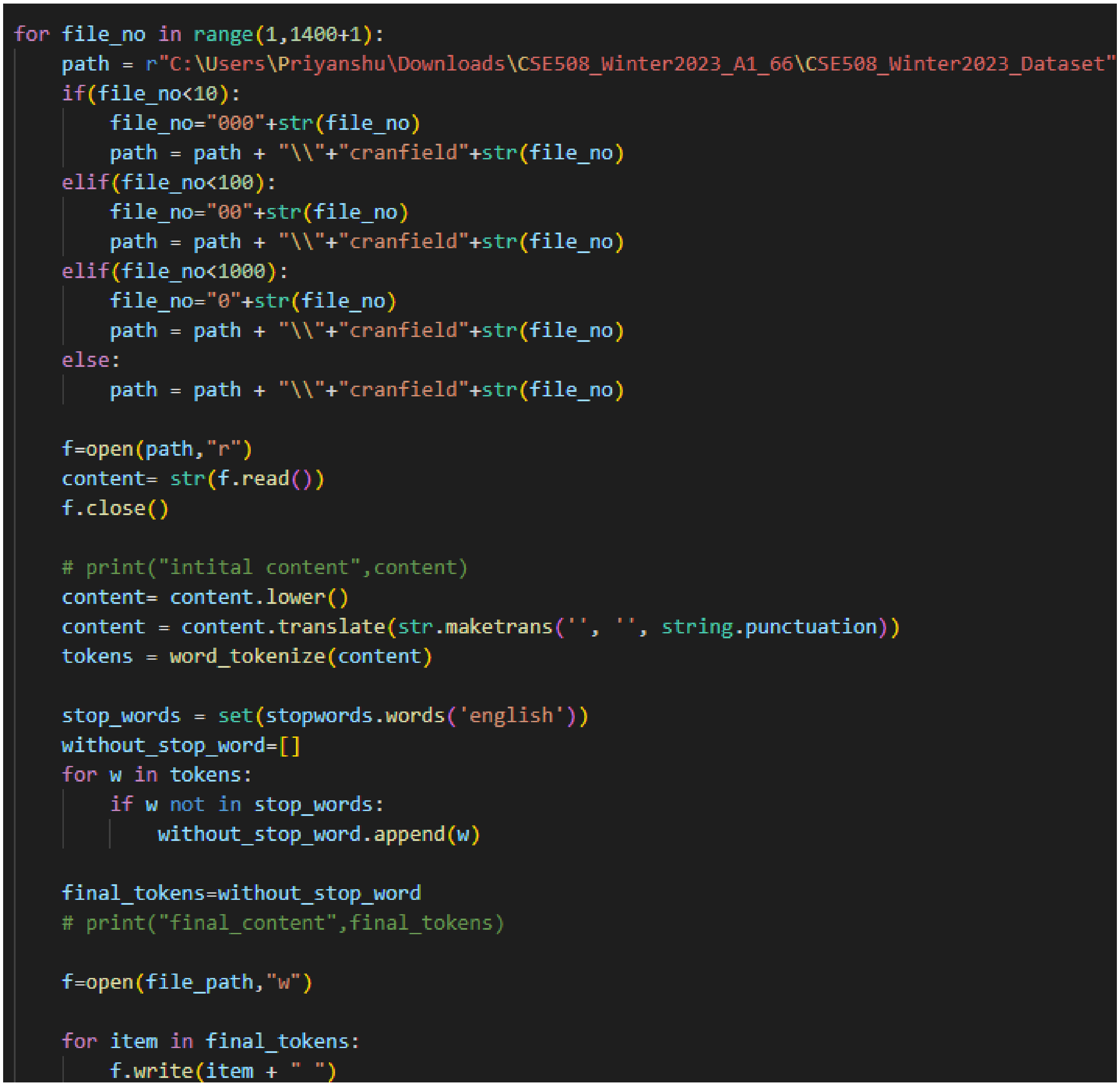
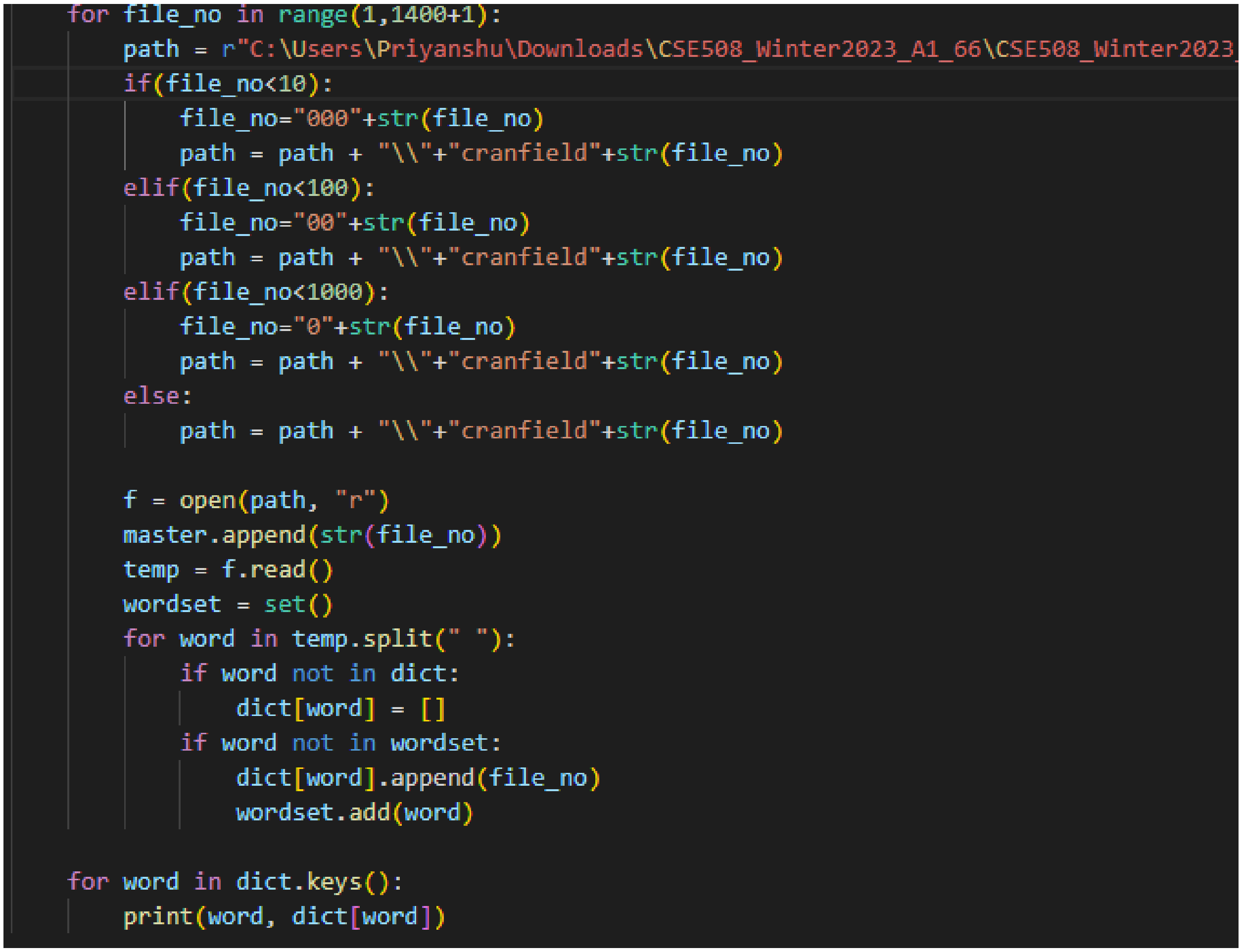
Assignment 1:

Priyanshu(2019473)
Himanshu Sehrawat(2019468)
Abhishek Goyal(2019136)

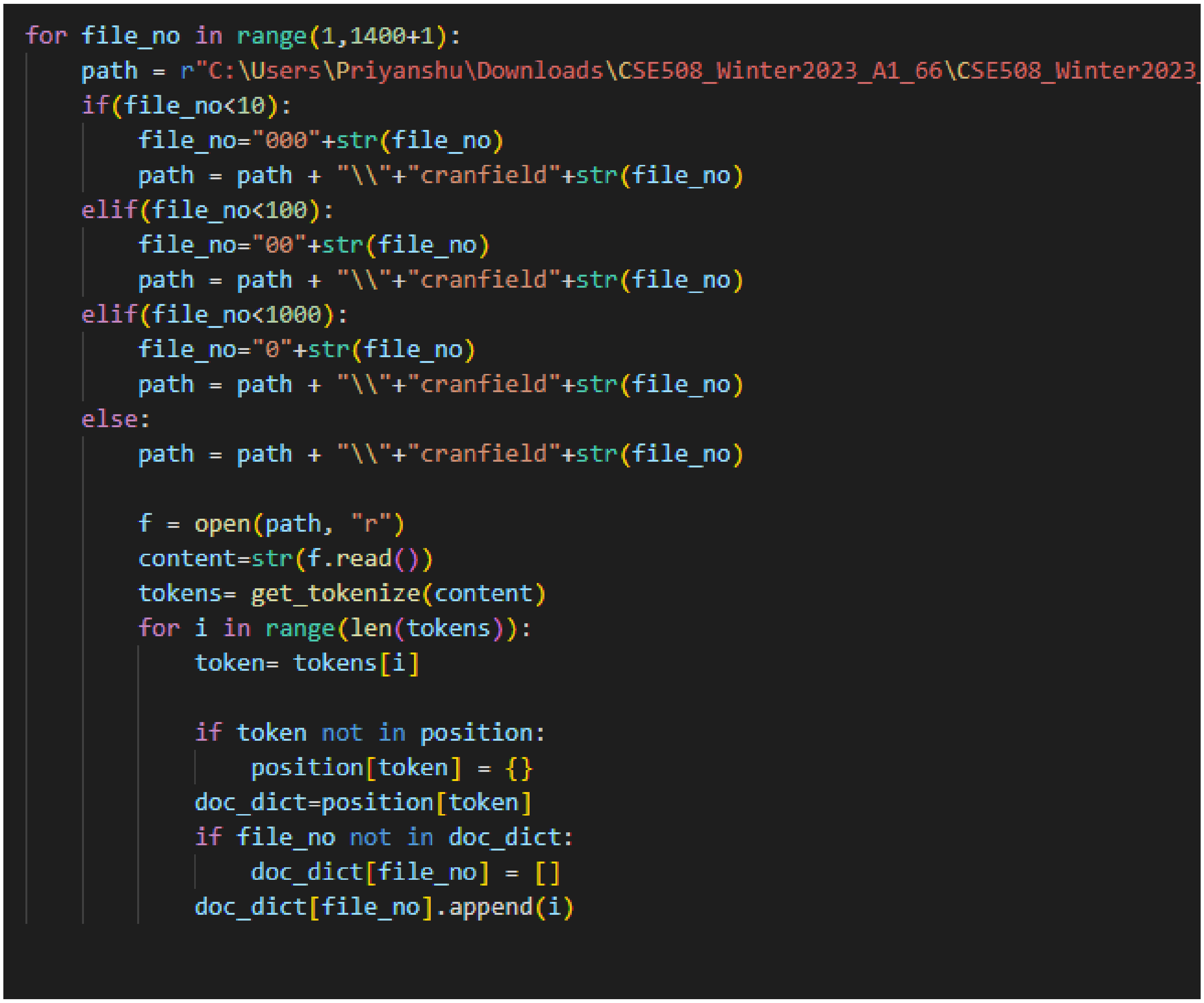
Code snippets:



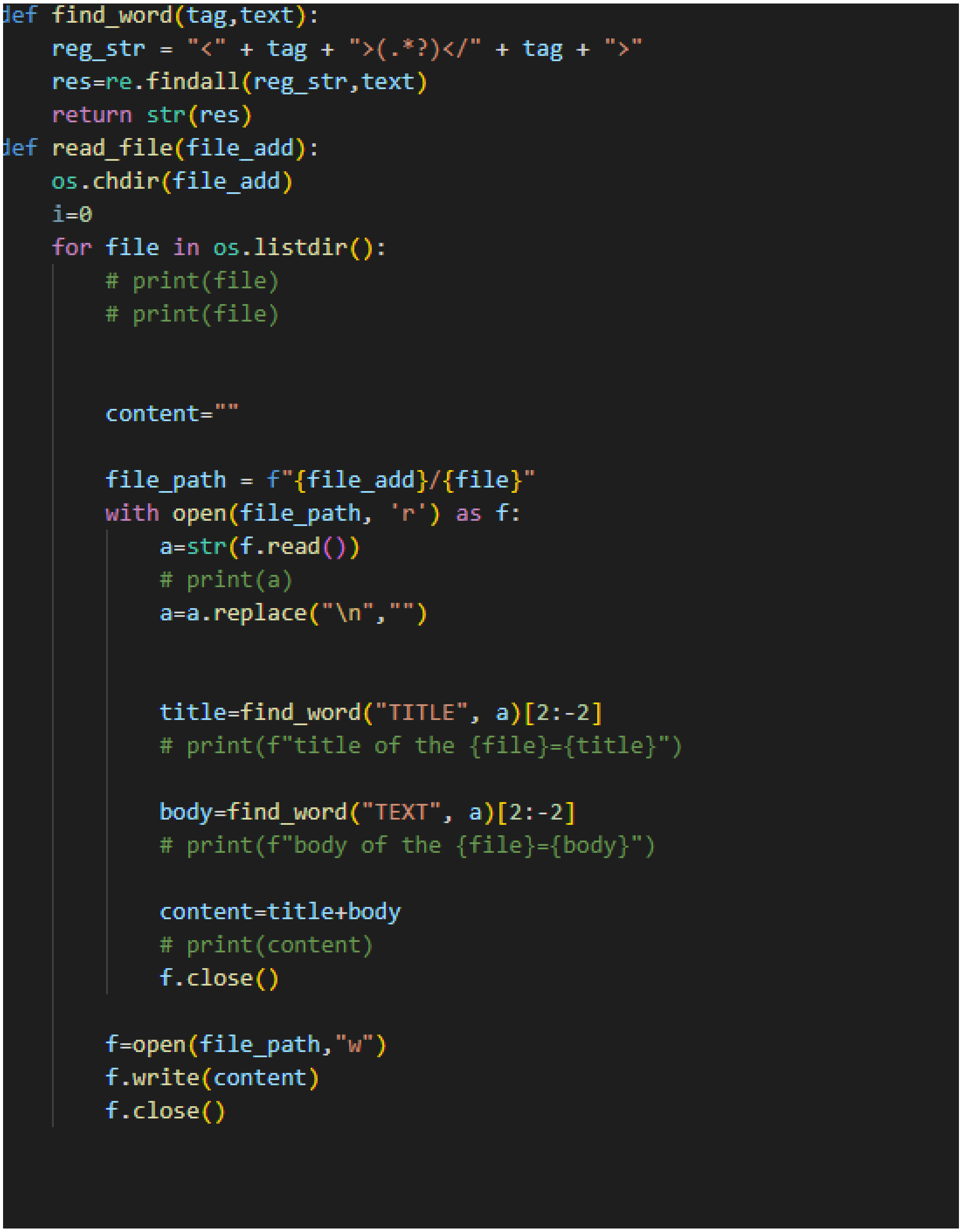
```
for file_no in range(1,1400+1):
    path = r"C:\Users\Priyanshu\Downloads\CSE508_Winter2023_A1_66\CSE508_Winter2023_Dataset"
    if(file_no<10):
        file_no="000"+str(file_no)
        path = path + "\\"+"cranfield"+str(file_no)
    elif(file_no<100):
        file_no="00"+str(file_no)
        path = path + "\\"+"cranfield"+str(file_no)
    elif(file_no<1000):
        file_no="0"+str(file_no)
        path = path + "\\"+"cranfield"+str(file_no)
    else:
        path = path + "\\"+"cranfield"+str(file_no)

    f=open(path,"r")
    content= str(f.read())
    f.close()

    # print("intital content",content)
    content= content.lower()
    content = content.translate(str.maketrans('', '', string.punctuation))
    tokens = word_tokenize(content)

    stop_words = set(stopwords.words('english'))
    without_stop_word=[]
    for w in tokens:
        if w not in stop_words:
            without_stop_word.append(w)

    final_tokens=without_stop_word
    # print("final_content",final_tokens)

    f=open(file_path,"w")

    for item in final_tokens:
        f.write(item + " ")
```



```
    for file_no in range(1,1400+1):
        path = r"C:\Users\Priyanshu\Downloads\CSE508_Winter2023_A1_66\CSE508_Winter2023_
        if(file_no<10):
            file_no="000"+str(file_no)
            path = path + "\\"+"cranfield"+str(file_no)
        elif(file_no<100):
            file_no="00"+str(file_no)
            path = path + "\\"+"cranfield"+str(file_no)
        elif(file_no<1000):
            file_no="0"+str(file_no)
            path = path + "\\"+"cranfield"+str(file_no)
        else:
            path = path + "\\"+"cranfield"+str(file_no)

        f = open(path, "r")
        master.append(str(file_no))
        temp = f.read()
        wordset = set()
        for word in temp.split(" "):
            if word not in dict:
                dict[word] = []
            if word not in wordset:
                dict[word].append(file_no)
                wordset.add(word)

    for word in dict.keys():
        print(word, dict[word])
```



```
for file_no in range(1,1400+1):
    path = r"C:\Users\Priyanshu\Downloads\CSE508_Winter2023_A1_66\CSE508_Winter2023_
    if(file_no<10):
        file_no="000"+str(file_no)
        path = path + "\\"+"cranfield"+str(file_no)
    elif(file_no<100):
        file_no="00"+str(file_no)
        path = path + "\\"+"cranfield"+str(file_no)
    elif(file_no<1000):
        file_no="0"+str(file_no)
        path = path + "\\"+"cranfield"+str(file_no)
    else:
        path = path + "\\"+"cranfield"+str(file_no)

        f = open(path, "r")
        content=str(f.read())
        tokens= get_tokenize(content)
        for i in range(len(tokens)):
            token= tokens[i]

            if token not in position:
                position[token] = {}
            doc_dict=position[token]
            if file_no not in doc_dict:
                doc_dict[file_no] = []
            doc_dict[file_no].append(i)
```



```
def find_word(tag,text):
    reg_str = "<" + tag + ">(.*?)</" + tag + ">"
    res=re.findall(reg_str,text)
    return str(res)
def read_file(file_add):
    os.chdir(file_add)
    i=0
    for file in os.listdir():
        # print(file)
        # print(file)


        content=""

        file_path = f"{file_add}/{file}"
        with open(file_path, 'r') as f:
            a=str(f.read())
            # print(a)
            a=a.replace("\n","")


            title=find_word("TITLE", a)[2:-2]
            # print(f"title of the {file}={title}")

            body=find_word("TEXT", a)[2:-2]
            # print(f"body of the {file}={body}")

            content=title+body
            # print(content)
            f.close()

        f=open(file_path,"w")
        f.write(content)
        f.close()
```