



ASSIGNMENT F1

Analyzing Shakespeare for CS4B

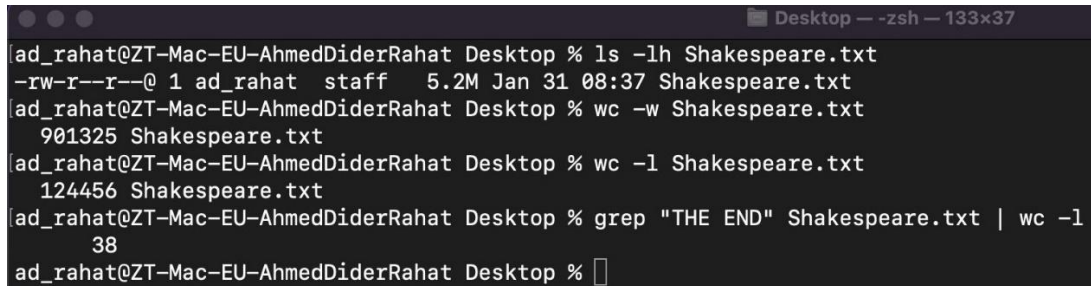
Submitted By: Ahmed Dider Rahat

Matriculation Number: 916146

Assignment Code: After implementing all the assignment I pushed my code on my [Github](#) link.

Answer to the question no. 1

In this assignment I used my MacBook terminal for running the UNIX command. The screen short are given bellow:

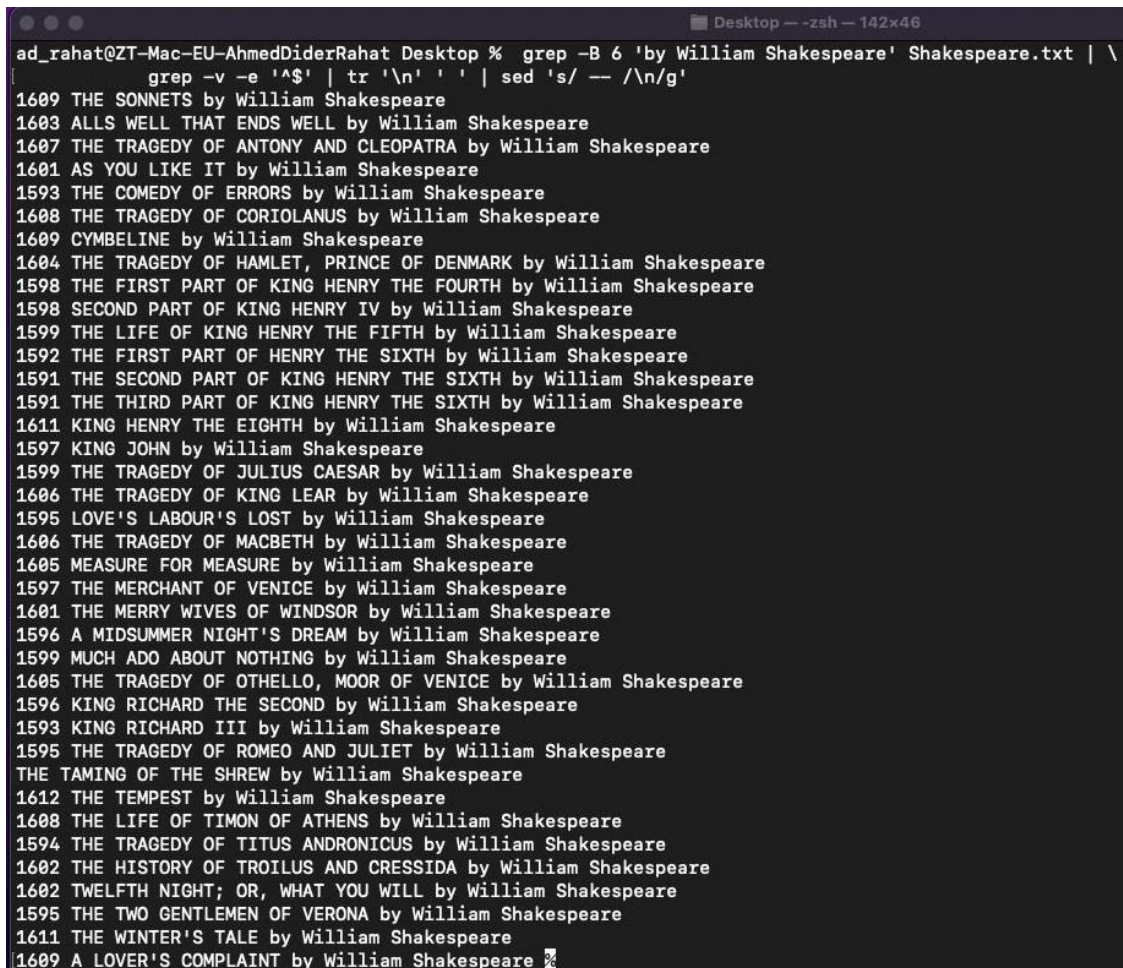


```
ad_rahah@ZT-Mac-EU-AhmedDiderRahat Desktop % ls -lh Shakespeare.txt
-rw-r--r--@ 1 ad_rahah  staff   5.2M Jan 31 08:37 Shakespeare.txt
ad_rahah@ZT-Mac-EU-AhmedDiderRahat Desktop % wc -w Shakespeare.txt
901325 Shakespeare.txt
ad_rahah@ZT-Mac-EU-AhmedDiderRahat Desktop % wc -l Shakespeare.txt
124456 Shakespeare.txt
ad_rahah@ZT-Mac-EU-AhmedDiderRahat Desktop % grep "THE END" Shakespeare.txt | wc -l
38
ad_rahah@ZT-Mac-EU-AhmedDiderRahat Desktop %
```

To find the number of play I used the count of THE END statements.

Answer to the question no. 2

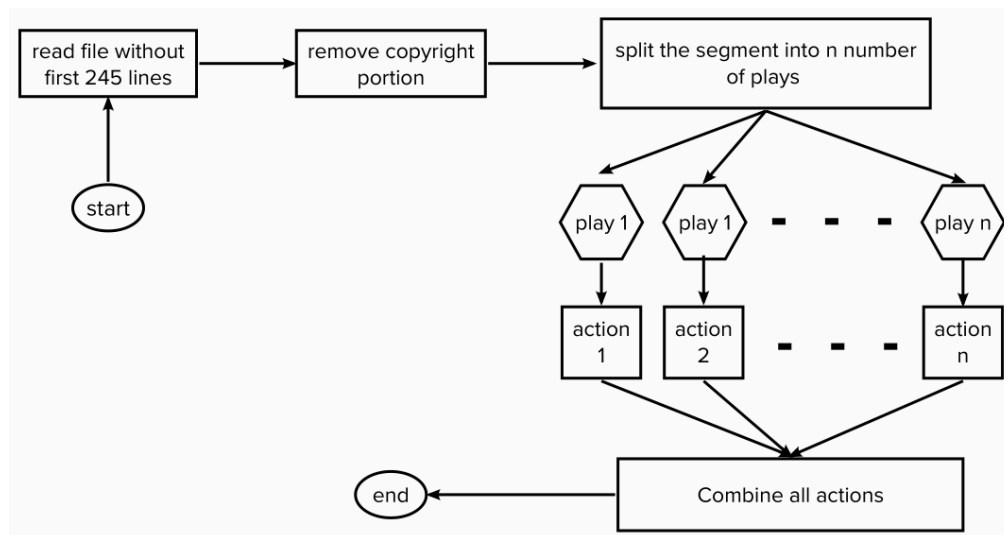
I also run the following command in my MacBook and get the given output.



```
ad_rahah@ZT-Mac-EU-AhmedDiderRahat Desktop % grep -B 6 'by William Shakespeare' Shakespeare.txt | \
grep -v -e '^$' | tr '\n' ' ' | sed 's/ -- /\n/g'
1609 THE SONNETS by William Shakespeare
1603 ALLS WELL THAT ENDS WELL by William Shakespeare
1607 THE TRAGEDY OF ANTONY AND CLEOPATRA by William Shakespeare
1601 AS YOU LIKE IT by William Shakespeare
1593 THE COMEDY OF ERRORS by William Shakespeare
1608 THE TRAGEDY OF CORIOLANUS by William Shakespeare
1609 CYMBELINE by William Shakespeare
1604 THE TRAGEDY OF HAMLET, PRINCE OF DENMARK by William Shakespeare
1598 THE FIRST PART OF KING HENRY THE FOURTH by William Shakespeare
1598 SECOND PART OF KING HENRY IV by William Shakespeare
1599 THE LIFE OF KING HENRY THE FIFTH by William Shakespeare
1592 THE FIRST PART OF HENRY THE SIXTH by William Shakespeare
1591 THE SECOND PART OF KING HENRY THE SIXTH by William Shakespeare
1591 THE THIRD PART OF KING HENRY THE SIXTH by William Shakespeare
1611 KING HENRY THE EIGHTH by William Shakespeare
1597 KING JOHN by William Shakespeare
1599 THE TRAGEDY OF JULIUS CAESAR by William Shakespeare
1606 THE TRAGEDY OF KING LEAR by William Shakespeare
1595 LOVE'S LABOUR'S LOST by William Shakespeare
1606 THE TRAGEDY OF MACBETH by William Shakespeare
1605 MEASURE FOR MEASURE by William Shakespeare
1597 THE MERCHANT OF VENICE by William Shakespeare
1601 THE MERRY WIVES OF WINDSOR by William Shakespeare
1596 A MIDSUMMER NIGHT'S DREAM by William Shakespeare
1599 MUCH ADO ABOUT NOTHING by William Shakespeare
1605 THE TRAGEDY OF OTHELLO, MOOR OF VENICE by William Shakespeare
1596 KING RICHARD THE SECOND by William Shakespeare
1593 KING RICHARD III by William Shakespeare
1595 THE TRAGEDY OF ROMEO AND JULIET by William Shakespeare
THE TAMING OF THE SHREW by William Shakespeare
1612 THE TEMPEST by William Shakespeare
1608 THE LIFE OF TIMON OF ATHENS by William Shakespeare
1594 THE TRAGEDY OF TITUS ANDRONICUS by William Shakespeare
1602 THE HISTORY OF TROILUS AND CRESSIDA by William Shakespeare
1602 TWELFTH NIGHT; OR, WHAT YOU WILL by William Shakespeare
1595 THE TWO GENTLEMEN OF VERONA by William Shakespeare
1611 THE WINTER'S TALE by William Shakespeare
1609 A LOVER'S COMPLAINT by William Shakespeare
```

Answer to the question no. 3

For creating the design I used an online site named app.mural.co. The out is given bellow:



Answer to the question no. 4-6

All the question from 4-6 are written implemented in Jupyter noyebok. The codes are attached in both '.ipynb' and '.py' file format.

Screen short of 4 (a):

```
In [4]: # remove header
def remove_hearder():
    all_text = read_text_file()

    # text to array -> took each line of the text book to each element of array
    text_to_array = all_text.split("\n")

    # take all the lines after 245
    text_to_array = text_to_array[244: ]

    # add all lines toghater
    all_text = ("\n".join(text_to_array)).strip()

    return all_text
```

Screen short of 4 (b):

```
In [5]: import re

# remove the autogenerated text
def remove_automatic_text():
    all_text = remove_hearder()

    # remove the auto generated text from the text by regular expression
    all_text = re.sub(r'<<[^\>]*>>', '', all_text)

    return all_text
```

Screen short of 4 (c):

```
In [6]: # split into segments

def split_to_segments():
    all_text = remove_automatic_text()

    # split all plays by THE END
    all_plays = all_text.strip().split("THE END")

    # remove the last portion End of ...
    all_plays.pop()

    all_plays = [plays.strip() for plays in all_plays]

    return all_plays
```

Screen short of 5:

```
In [7]: # the play count function that execute for each of the parallel call
def play_counts(play):

    # summary dictionary
    summary = {}

    # split a single play into each line
    plays_arr = play.strip().split("\n")

    # remove the line which contains 0 or 1 characters
    plays_arr = [line for line in plays_arr if len(line.strip()) > 1]

    # remove the number before play and the title from original play from the line list
    play_title = plays_arr[1]
    plays_arr = plays_arr[2:]

    # worlds list
    words_arr = (" ".join(plays_arr)).split(" ")

    # remove all empty space from the sentence
    words_arr = [word for word in words_arr if len(word.strip())>0]

    return {"title": play_title, "lines": len(plays_arr), "words": len(words_arr)}
```

```
In [29]: import pandas as pd

# initialize spark session
spark = SparkSession.builder.appName("Shakespeare Book Summary").getOrCreate()

# get all segments/plays
all_plays = split_to_segments()

# execute parallel call
summary = spark.sparkContext.parallelize(all_plays, len(all_plays)).map(play_counts)

# convert to data frame for easy access and sorting
df = summary.toDF().toPandas()

# sort by line number
df = df.sort_values(by="lines", ascending=False)

for ind, row in df.iterrows():
    print(f'{row["title"]}, {row["lines"]} lines, {row["words"]} words')

spark.stop()
```

Screen short of 6:

THE TRAGEDY OF HAMLET, PRINCE OF DENMARK, 4161 lines, 31946 words
KING RICHARD III, 4152 lines, 31084 words
THE TRAGEDY OF CORIOLANUS, 3909 lines, 29208 words
THE TRAGEDY OF ANTONY AND CLEOPATRA, 3841 lines, 26453 words
CYMBELINE, 3837 lines, 28782 words
THE HISTORY OF TROILUS AND CRESSIDA, 3632 lines, 27528 words
THE TRAGEDY OF KING LEAR, 3629 lines, 27505 words
THE TRAGEDY OF OTHELLO, MOOR OF VENICE, 3628 lines, 27856 words
KING HENRY THE EIGHTH, 3487 lines, 25810 words
THE WINTER'S TALE, 3366 lines, 25943 words
THE SECOND PART OF KING HENRY THE SIXTH, 3314 lines, 26773 words
THE LIFE OF KING HENRY THE FIFTH, 3311 lines, 27437 words
SECOND PART OF KING HENRY IV, 3282 lines, 27622 words
THE TRAGEDY OF ROMEO AND JULIET, 3280 lines, 25780 words
THE THIRD PART OF KING HENRY THE SIXTH, 3186 lines, 25822 words
THE FIRST PART OF KING HENRY THE FOURTH, 3072 lines, 25710 words
THE FIRST PART OF HENRY THE SIXTH, 3052 lines, 22766 words
KING RICHARD THE SECOND, 2977 lines, 23290 words
ALLS WELL THAT ENDS WELL, 2953 lines, 24365 words
MEASURE FOR MEASURE, 2873 lines, 22867 words
LOVE'S LABOUR'S LOST, 2826 lines, 22881 words
KING JOHN, 2785 lines, 21699 words
THE MERRY WIVES OF WINDSOR, 2783 lines, 23352 words
THE TAMING OF THE SHREW, 2771 lines, 22151 words
THE TRAGEDY OF JULIUS CAESAR, 2758 lines, 20862 words
THE TRAGEDY OF TITUS ANDRONICUS, 2753 lines, 21626 words
THE MERCHANT OF VENICE, 2736 lines, 22247 words
AS YOU LIKE IT, 2686 lines, 22805 words
THE LIFE OF TIMON OF ATHENS, 2581 lines, 19625 words
THE TRAGEDY OF MACBETH, 2566 lines, 18226 words
MUCH ADO ABOUT NOTHING, 2545 lines, 22443 words
TWELFTH NIGHT; OR, WHAT YOU WILL, 2510 lines, 21144 words
THE TEMPEST, 2434 lines, 17410 words
THE TWO GENTLEMEN OF VERONA, 2314 lines, 18285 words
THE SONNETS, 2301 lines, 17730 words
A MIDSUMMER NIGHT'S DREAM, 2239 lines, 17226 words
THE COMEDY OF ERRORS, 1919 lines, 16195 words
A LOVER'S COMPLAINT, 330 lines, 2562 words

Answer to the question no. 7

Stop words: Stop words are a set of commonly used words in a language. Some of the examples of stop words in English are “a”, “the”, “is”, “are” and etc.

Links are given bellow:

1. <https://gist.github.com/sebleier/554280>
2. <https://www.ranks.nl/stopwords>
3. <https://countwordsfree.com/stopwords>