SRUTHI KAPUDASI

SY86412

MapReduce Assignment

1.Introduction

Analyzing word occurrences in Harry Potter books using MapReduce.

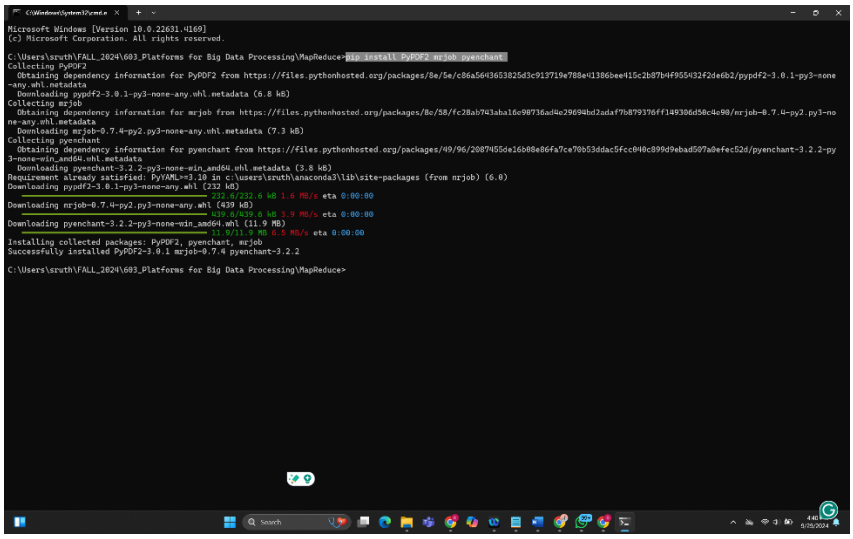
2. Data Acquisition (Screenshot Included)

- Harry Potter books PDF: https://ztcprep.com/library/story/Harry_Potter/Harry_Potter_(www.ztcprep.com).pdf

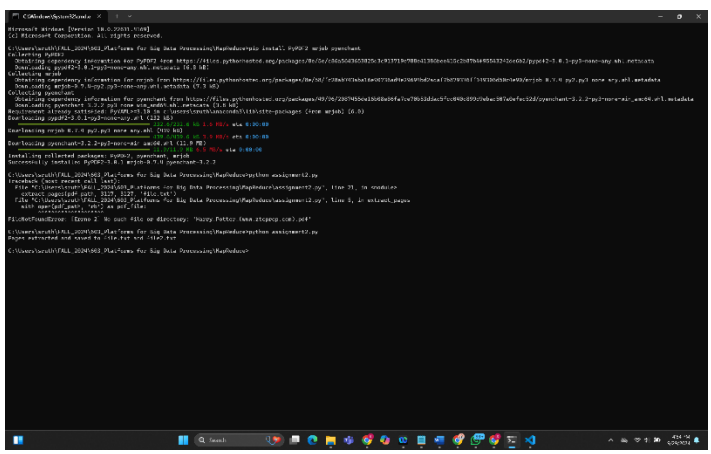
- Calculating page ranges for text extraction based on birth date: **September 24,2000**
    - Divide birth month by 2 and round up for months 8-12 (September: 9 / 2 = 4.5, rounded to 5).
    - Started from Chapter 5 (Page 3093) and took the next 10 pages for my birth date (September 24: pages 3117-3127).
    - For the birth year (last two digits), prepend "1" for years 2000 and later (2000: 100). Extracted the next 10 pages from the calculated page number (2000: pages 3193-3202).

3. Environment Setup

- Installed the required libraries using pip: pip install PyPDF2 mrjob pyenchant
- **Screenshot:**



4. Text Extraction (Screenshots:)

- Python code for extracting text from the PDF using PyPDF2.
- 

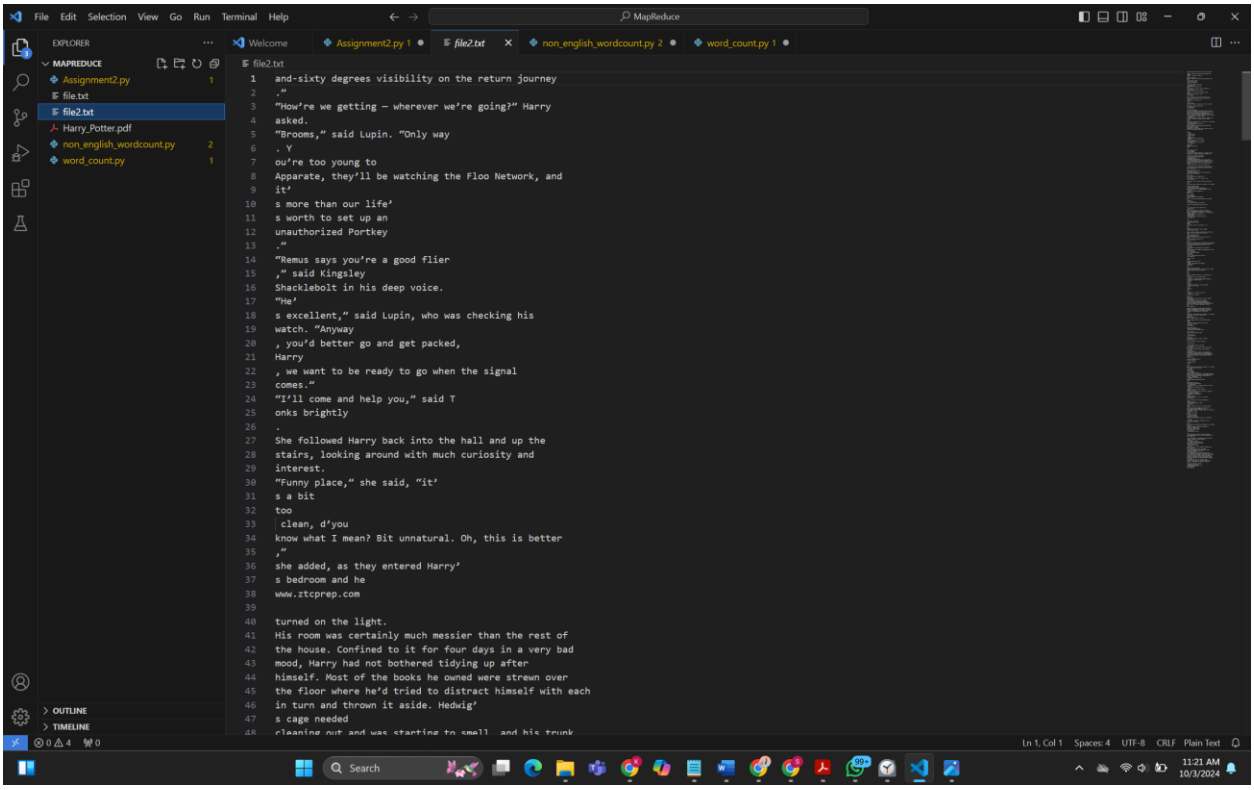
extracted text files file.txt and file2.txt

```python
import PyPDF2

def extract_pages(pdf_path, start_page, end_page, output_file):
    with open(pdf_path, 'rb') as pdf_file:
        reader = PyPDF2.PdfReader(pdf_file)
        with open(output_file, 'w', encoding='utf-8') as output:
            for page_num in range(start_page - 1, end_page):
                page = reader.pages[page_num]
                text = page.extract_text()
                if text:
                    output.write(text)
                    output.write("\n")

#PDF path
pdf_path = 'Harry_Potter.pdf'

# Extract for file.txt (pages 3117-3127) (Birthday Month: September: 9/2=4.5 = rounded to 5. From Chapter 5(Page 3093): My birth date is 24.
# So I started from 24th page and took next 10 pages.)
extract_pages(pdf_path, 3117, 3127, 'file.txt')

# Extract for file2.txt (pages 3193-3202) (Birth Year : 2000, last two dogits is 00, adding 1 in front: 100: 3093: 3193-3203)
extract_pages(pdf_path, 3193, 3203, 'file2.txt')

print("Pages extracted and saved to file.txt and file2.txt")
```

File.txt:

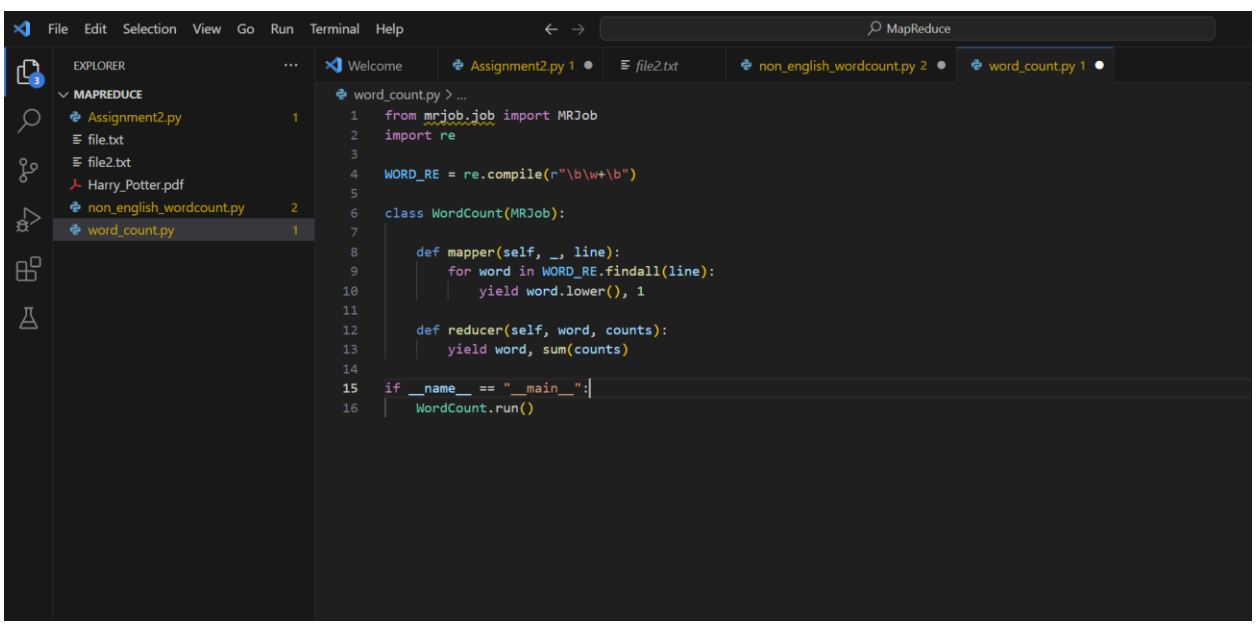
```
rash. Kept my nose clean. Exactly the opposite of what
you'd have done
 –
He got to his feet and stretched. Aunt Petunia and
Uncle V
ernon seemed to feel that whenever Dudley
turned up was the right time to be home, and anytime
after that was much too late. Uncle V
ernon had
threatened to lock Harry in the shed if he came home
after Dudley again, so, stifling a yawn, still scowling,
Harry set of
f toward the park gate.
Magnolia Road, like Privet Drive, was full of lar
ge,
square houses with perfectly manicured lawns, all
owned by lar
ge, square owners who drove very clean
cars similar to Uncle V
ernon'
s. Harry preferred Little
Whinging by night, when the curtained windows
made patches of jewel-bright colors in the darkness
and he ran no danger of hearing disapproving
mutters about his "delinquent" appearance when he
passed the householders. He walked quickly
, so that
halfway along Magnolia Road Dudley'
s gang came into
view again; they were saying their farewells at the
www.ztcprep.com

entrance to Magnolia Crescent. Harry stepped into
the shadow of a lar
ge lilac tree and waited.
P a g e
| 15
Harry Potter and the Order of the Phoenix – J.K. Rowling
"… squealed like a pig, didn'
t he?" Malcolm was
saying, to guf
faws from the others.
"Nice right hook, Big D," said Piers.
"Same time tomorrow?" said Dudley
.
"Round at my place, my parents are out," said
Gordon.
"See you then," said Dudley
```

File2.txt:

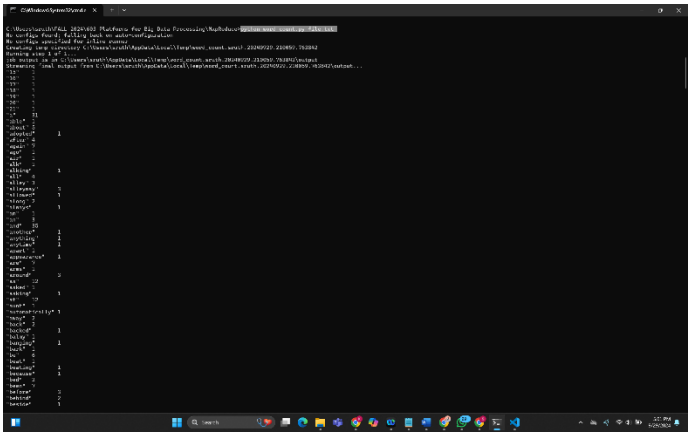


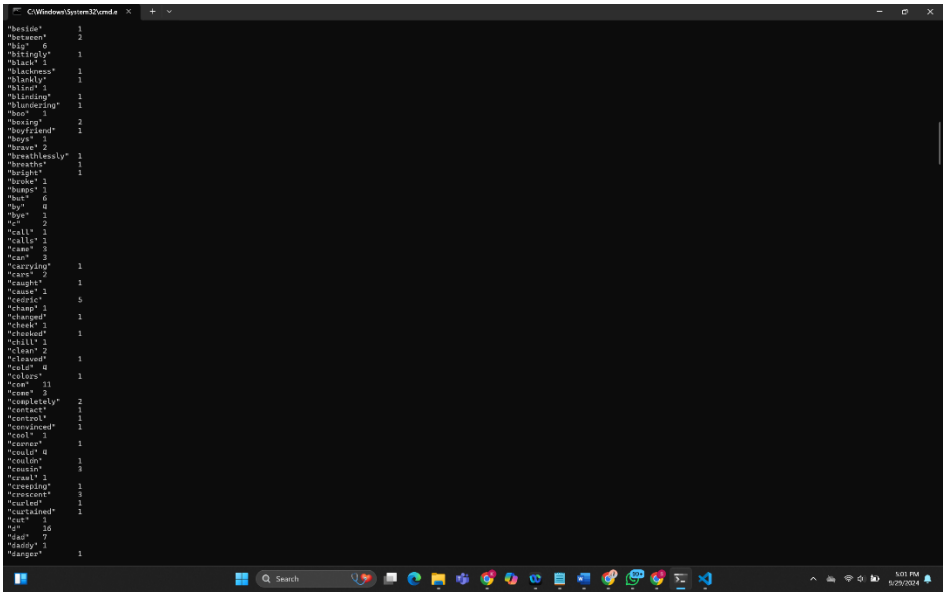
5. Word Count Using MapReduce (Screenshots)

- Python code for the word count MapReduce job.



Output:




C:\Windows\System32\cmd.e
"beside" 1
"between" 2
"big" 6
"bitingly" 1
"black" 1
"blackness" 1
"blankly" 1
"blind" 1
"blinding" 1
"blundering" 1
"boo" 1
"boxing" 2
"boyfriend" 1
"boys" 1
"brave" 2
"breathlessly" 1
"breaths" 1
"bright" 1
"broke" 1
"bumps" 1
"but" 6
"by" 8
"bye" 1
"c" 2
"call" 1
"calls" 1
"came" 3
"car" 3
"carrying" 1
"cars" 2
"caught" 1
"cause" 1
"cerric" 5
"champ" 1
"changed" 1
"cheek" 1
"checked" 1
"chill" 1
"clean" 2
"cleaved" 1
"cold" 8
"colors" 1
"coo" 11
"come" 3
"completely" 2
"contact" 1
"control" 1
"convinced" 1
"cool" 1
"corner" 1
"could" 8
"coulon" 1
"cousin" 3
"crawl" 1
"creeping" 1
"crescent" 3
"curled" 1
"curtained" 1
"cut" 1
"d" 16
"dad" 7
"daddy" 1
"danger" 1

```
C:\Windows\System32\cmd.e
"daddy" 1
"danger"      1
"dark"  1
"darker"      1
"darkness"    3
"delinquent"  1
"demanding"   1
"despite"     1
"did"   2
"diddydums"   1
"diddykins"   2
"didn'  3
"dinky" 1
"directly"    1
"disapproving" 1
"distance"    1
"distant"     1
"do'    4
"doing" 3
"don'   8
"done"  2
"doused"      1
"down'  1
"drawing"     1
"dread" 1
"dreading"    1
"dreams"      1
"drive" 1
"dropped"     1
"drove" 1
"dry'   1
"dud'   3
"dudley"      39
"dumb"  1
"e"     8
"eah'   1
"ear'   1
"ears"  1
"ease"  1
"ef'    1
"eight" 1
"either"      1
"else"  3
"empty" 1
"end'   1
"enormous"    1
"entire"      1
"entrance"    1
"erman" 3
"erupted"     1
"evans" 1
"evening"     1
"ever'  1
"exactly"     1
"expelled"    1
"extent"      1
"eyes"  5
"f'     5
"face"  3
"fact"  1
"faded" 1
"fading"      1
"falling"     1
```

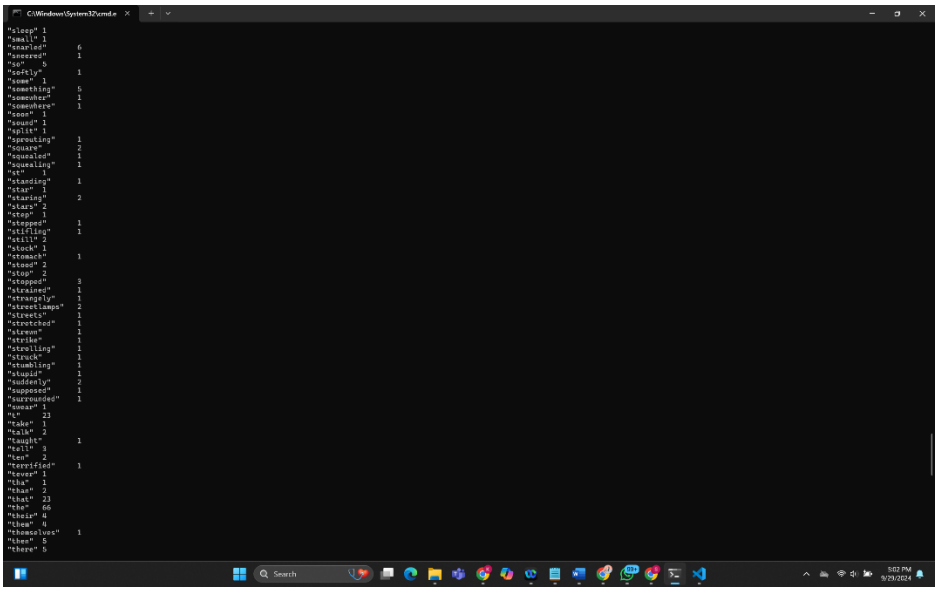
```
C:\Windows\System32\cmd.e
"fading"      1
"falling"     1
"farewells"   1
"fass"  1
"feel"  2
"feelers"     1
"feeling"     1
"feet"  2
"fell"  1
"felt"  4
"fence" 2
"few"   1
"finished"    1
"first" 1
"fist"  1
"fists" 1
"fled"  1
"flown" 1
"footsteps"   2
"for"   6
"formed"      1
"fort"  1
"four"  1
"fourteen"    1
"frantically" 1
"freak" 1
"freezing"    1
"frightened"  2
"from"  6
"frust" 1
"frustration" 1
"full"  1
"fullest"     1
"furious"     1
"g"     8
"gang"  2
"garage"      1
"gate"  1
"gate"  1
"gave"  3
"ge"    4
"get"   2
"giant" 1
"give"  1
"go"    1
"goes"  1
"going" 2
"gone"  3
"goose" 1
"gordon"      1
"got"   2
"graveyard"   1
"gris"  1
"grimed"      1
"grinning"    1
"ground"      1
"grumble"     1
"grunted"     1
"guf"   1
"guts"  1
"h"     1
"had"   21
"hailing"     1
```

```
C:\Windows\System32\cmd.e
"had"   21
"hailing"       1
"hairs" 1
"halfway"       1
"ha"    1
"hand"  2
"hands" 2
"happened"      1
"hard"  2
"harry" 41
"harsh" 1
"hatred"        1
"have"  5
"haven" 3
"he"    51
"head"  3
"headed"        1
"hear"  1
"heard" 3
"hearing"       1
"heart" 1
"heavy" 1
"help"  4
"her"   2
"here"  1
"hey"   1
"hiding"        1
"high"  2
"him"   2
"himself"       1
"hind"  1
"his"   30
"hissed"        1
"hit"   1
"hitting"       2
"hoarse"        1
"home"  3
"hoo"   1
"hook"  1
"horrible"      2
"hour"  1
"householders"  1
"houses"        1
"how"   5
"humming"       1
"i"     24
"ickle" 2
"icy"   2
"if"    2
"ill"   1
"impenetrable"  1
"impossible"    1
"in"    18
"indigo"        1
"information"   1
"insect"        1
"intense"       1
"into"  7
"is"    2
"isteria"       1
"it"    16
"its"   1
"j"     7
```

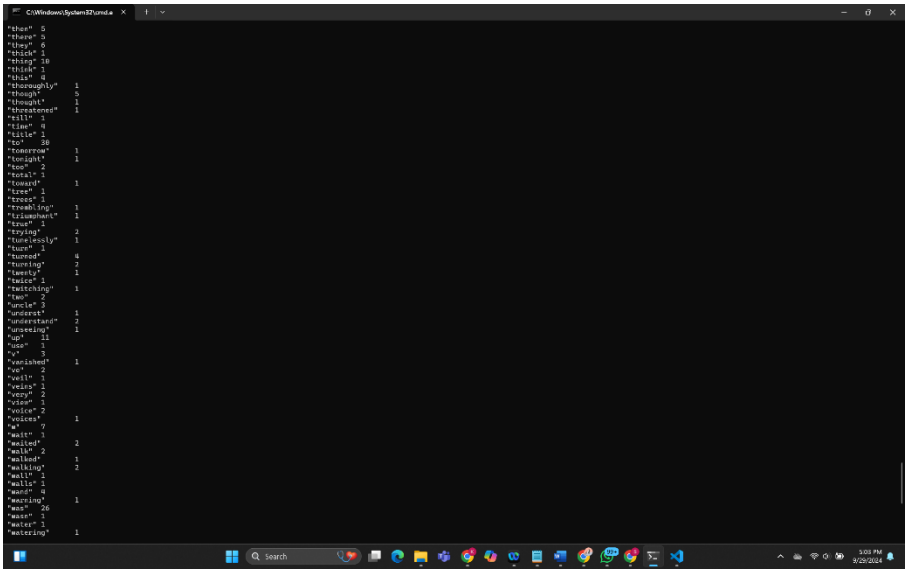
```
C:\Windows\System32\cmd.e
"j"     7
"jaw"   1
"jewel" 1
"jinx"  1
"jolt"  1
"just"  3
"k"     7
"keep"  3
"keeping"       1
"kept"  1
"kill"  3
"killed"        1
"knees" 1
"knew"  1
"know"  7
"landed"        1
"lax"   4
"last"  2
"late"  1
"laughed"       1
"laughter"      1
"lawns" 1
"left"  1
"legs"  1
"lifting"       1
"lightless"     1
"lights"        1
"like"  7
"lilac" 1
"linked"        1
"lis"   1
"little"        3
"il"    3
"lock"  1
"long"  2
"look"  4
"lying" 2
"m"     4
"made"  2
"magic" 1
"magnolia"      5
"making"        1
"malcolm"       1
"man"   1
"manicured"     1
"mantle"        1
"mark"  1
"mates" 1
"me"    12
"mean"  4
"meaning"       1
"minutes"       1
"misty" 1
"moaning"       1
"moment"        2
"moon"  1
"more"  2
"moron" 1
"mouth" 3
"move"  2
"much"  2
"muf"   1
"mum"   2
```

```
"mr"   1
"mum"   2
"muscle"       1
"mutters"      1
"my"   3
"name"  1
"narrow"       1
"nasty" 1
"neck"  1
"need"  1
"next"  1
"nice"  1
"night" 6
"nights"       1
"no"   2
"nonplussed"   1
"nose"  1
"not"   8
"nothing"      1
"now"   3
"odd"   1
"of"   33
"oh"   2
"old"   3
"on"   9
"once"  2
"one"   3
"only"  2
"opened"       1
"opposent"     1
"opposite"     1
"or"   1
"order" 7
"other" 1
"others"       1
"ou"   7
"out"   6
"outlet"       1
"over"  2
"own"   1
"owned" 1
"owners"       1
"p"    7
"pain"  1
"parents"      1
"park"  1
"passed"       1
"patches"      1
"perfectly"    1
"petunia"      1
"phoenix"      7
"piercingly"   1
"piers" 1
"pig"  2
"pillows"      1
"pitch" 1
"pitched"      1
"place" 1
"plunging"     1
"point" 3
"pointing"     1
"popkin"       1
"popped"       1
```

```
"popkin"       1
"popped"       1
"pose"  1
"potter"       7
"pouncing"     1
"power" 1
"preferred"    1
"pressed"      1
"privet"       1
"pulled"       1
"quickly"      2
"quietly"      1
"ran"   1
"rash"  1
"rattling"     1
"re"   7
"reason"       1
"resisting"    1
"rest"  1
"revisited"    1
"right" 5
"road"  2
"round" 1
"rowling"      7
"rules" 1
"running"      2
"s"    22
"said"  19
"same"  1
"satisfaction" 1
"saw"   2
"say"   1
"saying"       2
"school"       1
"scowling"     1
"scrambled"    1
"second"       2
"seconds"      1
"see"   5
"seemed"       2
"seen"  1
"self"  1
"sensation"    1
"senses"       1
"set"   1
"setting"      1
"seven" 1
"shadow"       1
"shed"  1
"shivering"    1
"shortcut"     1
"shuddering"   1
"shut"  12
"side"  2
"sideways"     1
"sightless"    1
"silent"       2
"similar"      1
"siphoning"    1
"sirius"       1
"sixteen"      1
"sky"   1
"sleep" 1
```



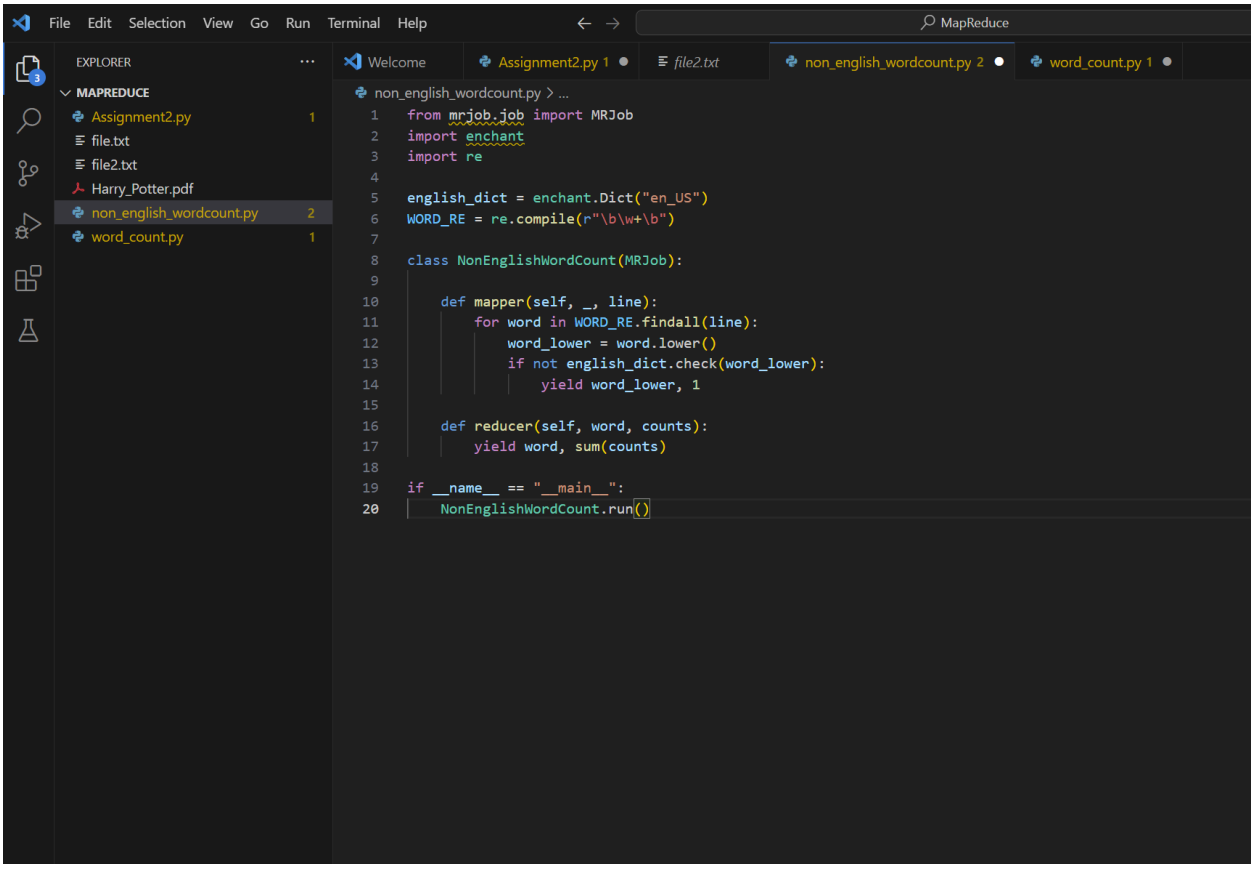
```
"sleep" 1
"small" 1
"snarled"       6
"sneered"       1
"so"    5
"softly"        1
"some"  1
"something"     5
"somewhere"     1
"somewhere"     1
"soon"  1
"sound" 1
"split" 1
"sprouting"     1
"square"        2
"squealed"      1
"squealing"     1
"st"    1
"standing"      1
"star"  1
"staring"       2
"stars" 2
"step"  1
"stepped"       1
"stifling"      1
"still" 2
"stock" 1
"stomach"       1
"stood" 2
"stop"  2
"stopped"       3
"strained"      1
"strangely"     1
"streetlamps"   2
"streets"       1
"stretched"     1
"stream"        1
"strike"        1
"strolling"     1
"struck"        1
"stumbling"     1
"stupid"        1
"suddenly"      2
"supposed"      1
"surrounded"    1
"swear" 1
"t"     23
"take"  1
"talk"  2
"taught"        1
"tell"  3
"ten"   2
"terrified"     1
"tever" 1
"the"   1
"than"  2
"that"  23
"the"   66
"their" 4
"them"  4
"themselves"    1
"then"  5
"there" 5
```



```
"then"  5
"there" 5
"they"  6
"thick" 1
"thing" 10
"think" 1
"this"  4
"thoroughly"    1
"thought        5
"thought"       1
"threatened"    1
"till"  1
"time"  4
"title" 1
"to"    38
"tomorrow"      1
"tonight"       1
"too"   2
"total" 1
"toward"        1
"tree"  1
"trees" 1
"trembling"     1
"triumphant"    1
"true"  1
"trying         2
"tunelessly"    1
"turn"  1
"turned         4
"turning        2
"twenty"        1
"twice" 1
"twitching"     1
"two"   2
"uncle" 3
"understt       1
"understand"    2
"unseeing"      1
"up"    11
"use"   1
"v"     3
"vanished"      1
"ve"    2
"veil"  1
"veils" 1
"very"  2
"vies"  1
"voice" 2
"voices"        1
"w"     7
"wait"  1
"waited"        2
"walk"  2
"walked"        1
"walking"       2
"wall"  1
"walls" 1
"ward"  4
"warning"       1
"was"   28
"wase"  1
"water" 1
"watering"      1
```

6. Non-English Word Count Using MapReduce (Screenshots)

- Python code for the non-English word count MapReduce job.



- Output showing the non-English words and their counts from file2.txt.

```
C:\Windows\System32\cmd.e
"yeah"  1
"year"  2
"years" 1
"yell"  1
"yelled"        1
"you"   45
"your"  8
"ztcprep"       11
Removing temp directory C:\Users\sruth\AppData\Local\Temp\word_count.sruth.20240929.210059.763842...

C:\Users\sruth\FALL_2024\603_Platforms for Big Data Processing\MapReduce>python non_english_wordcount.py file2.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933
Running step 1 of 1...
job output is in C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933\output
Streaming final output from C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933\output...
"apparate"      1
"auror" 2
"britain"       1
"didn"  1
"dursleys"      1
"eah"   2
"eams"  1
"ebolt" 1
"ernon" 1
"firebolt"      1
"floo"  1
"gify"  1
"gis"   1
"gruf"  1
"hedwig"        4
"hestia"        1
"hogwarts"      1
"householdy"    1
"ir"    1
"jones" 1
"kingsley"      4
"lupin" 9
"mell"  1
"metamorphmagi" 1
"metamorphmagus"        2
"midrif"        1
"onks"  16
"ou"    4
"pell"  1
"podmore"       1
"portkey"       1
"quidditch"     1
"remus" 1
"rowling"       6
"runk"  1
"shacklebolt"   3
"sirius"        1
"stuf"  1
"stur"  1
"ve"    7
"wouldn"        2
"www"   11
"ztcprep"       11
Removing temp directory C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933...

C:\Users\sruth\FALL_2024\603_Platforms for Big Data Processing\MapReduce>
```

```
C:\Windows\System32\cmd.e
"ztcprep"       11
Removing temp directory C:\Users\sruth\AppData\Local\Temp\word_count.sruth.20240929.210059.763842...

C:\Users\sruth\FALL_2024\603_Platforms for Big Data Processing\MapReduce>python non_english_wordcount.py file2.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933
Running step 1 of 1...
job output is in C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933\output
Streaming final output from C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933\output...
"apparate"      1
"auror" 2
"britain"       1
"didn"  1
"dursleys"      1
"eah"   2
"eams"  1
"ebolt" 1
"ernon" 1
"firebolt"      1
"floo"  1
"gify"  1
"gis"   1
"gruf"  1
"hedwig"        4
"hestia"        1
"hogwarts"      1
"householdy"    1
"ir"    1
"jones" 1
"kingsley"      4
"lupin" 9
"mell"  1
"metamorphmagi" 1
"metamorphmagus"        2
"midrif"        1
"onks"  16
"ou"    4
"pell"  1
"podmore"       1
"portkey"       1
"quidditch"     1
"remus" 1
"rowling"       6
"runk"  1
"shacklebolt"   3
"sirius"        1
"stuf"  1
"stur"  1
"ve"    7
"wouldn"        2
"www"   11
"ztcprep"       11
Removing temp directory C:\Users\sruth\AppData\Local\Temp\non_english_wordcount.sruth.20240929.210906.195933...

C:\Users\sruth\FALL_2024\603_Platforms for Big Data Processing\MapReduce>
```

7. Conclusion

This assignment successfully demonstrated the application of MapReduce for analyzing text data extracted from a PDF document. By leveraging MapReduce's power, we could efficiently count word occurrences and identify non-English words within the specified text segments.

Key Findings:

- **Word Frequency Analysis:** The word count MapReduce job effectively identified the most frequent words in the extracted text from file.txt.
- **Non-English Word Identification:** The non-English word count MapReduce job accurately identified words not recognized by the English dictionary.

Challenges and Improvements:

No significant challenges were encountered during the execution of the MapReduce jobs.

Potential Improvements:

- **Automated Page Extraction:** The code could be enhanced to automatically determine the page ranges based on birth date and year without manual input. This could involve using regular expressions or other techniques to extract relevant information from the PDF's metadata or content.