For this assignment, download the Harry Potter Books data from the following link (PDF is also attached):

https://ztcprep.com/library/story/Harry_Potter/Harry_Potter_(www.ztcprep.com).pdf

Extract Data:

Select the book which corresponds to your birth month. For birth month 8-12, divide by 2 and round up.

Once you selected the book, go to page number that corresponds to your birth date (1-31) and extract next 10 pages of the book to a text file (file1.txt).

Next, go to page number that corresponds to your birth year (last 2 digits). For year 2000 onwards, use 1 infront of the year number to find the page number (so year 2000 becomes 100, 2001 - 101 and so on). Extract next 10 pages into another text file (file2.txt).

Write Code to analyze data:

1. Write Python code and use MapReduct to count occurrences of each word in the first text file (file.txt). How many times each word is repeated?
2. From the second text file (file2.txt), write Python code and use MapReduct to count how many times non-English words (names, places, spells etc.) were used. List those words and how many times each was repeated. There are multiple ways of doing this. You can use pyenchant (https://pypi.org/project/pyenchant/), pyspellchecker (https://pyspellchecker.readthedocs.io/en/latest/) or just download a list of words (http://www.gwicks.net/dictionaries.htm) and search through them.

DateOfBirth : 25-04-1999

- Taking book number 4 as my book because my birth month is 4

Question 1: Write Python code and use MapReduct to count occurrences of each word in the first text file (file.txt). How many times each word is repeated?

In [15]:
```
pip install mrjob
```

```
Requirement already satisfied: mrjob in c:\users\saigo\anaconda3\lib\site-packages
(0.7.4)
Requirement already satisfied: PyYAML>=3.10 in c:\users\saigo\anaconda3\lib\site-pack
ages (from mrjob) (6.0)
Note: you may need to restart the kernel to use updated packages.
```

In [16]:
```
%%file wordcount.py

from mrjob.job import MRJob

class wordcount(MRJob):

    def mapper(self, _, line):
        line = line.strip()
        words = line.split()
        for word in words:
            yield word, 1

    def reducer(self, word, counts):
        yield word, sum(counts)

if __name__ == '__main__':
    wordcount.run()
```

```
Overwriting wordcount.py
```

In [17]:
```
import wordcount
mr_job = wordcount.wordcount(args=["file1.txt"])
with mr_job.make_runner() as runner:
    runner.run()
    for key, value in mr_job.parse_output(runner.cat_output()):
        print(key, value)
```

```
No configs specified for inline runner
```

```
(Harry 1
(Hermione’s 1
(however 1
(which 1
A 2
About 1
Allowing 1
And 1
Arthur, 1
As 1
Aunt 16
Bewildered, 1
Bit 1
Britain 1
Buckbeak 1
But 2
By 1
Cup 1
Daily 1
Dear 2
Department 1
Diddy 1
Do 1
Dudley 7
Dudley, 2
Dudley’s 4
Dudley’s. 1
Dursley 1
Dursley, 1
Dursleys 4
Dursleys’ 1
Errol, 2
Every 1
Express,” 1
Fire 8
Games 1
Goblet 8
Hagrid, 2
Hagrid’s 1
Harry 34
Harry, 4
Harry. 5
Harry’s 5
He 9
Hedwig 4
Hermione, 1
Hermione’s 1
His 2
Hog 1
Hogwarts 1
Hogwarts. 1
Hoping 1
However, 1
I 7
INVITATION 1
I’ll 1
I’m 1
If 1
It 1
J.K. 8
```

```
Last 1
Magical 1
Mail, 1
Mega-Mutilation 1
Molly 1
Monday 1
Mr. 1
Mrs. 7
Muggle 1
My 2
No 1
Nobody 1
None 1
Now.” 1
On 1
Other 1
P.S. 1
Part 1
Petunia 13
Petunia. 1
Petunia’s 2
PlayStation 1
Poor 1
Potter 8
Quick 1
Quidditch 1
Ron, 1
Ron. 1
Ron’s 1
Rowling 8
Say 1
Seemed 1
She 3
Sirius, 1
Sirius. 1
Smeltings 1
So 1
Sports. 1
THE 1
Thanks 1
That 1
That’s 2
The 5
Their 1
Then 2
There 1
They 2
Things 1
This 1
Three 1
To 2
Uncle 28
Vernon 22
Vernon. 1
Vernon’s 5
Voldemort 1
We 2
Weasley 2
Weasley, 1
Weasley. 1
```

```
Weasley’s 3
Weasleys’ 1
When 1
Who 2
Without 1
World 1
Yes, 1
Yours 1
– 8
— 9
“A 1
“Dumpy 1
“He’s 1
“Hogwarts 1
“In 1
“Is 1
“Load 1
“Look 1
“She 1
“She’s 1
“So 2
“So,” 1
“So. 1
“The 1
“There 1
“This 1
“Very 1
“Who 1
“You,” 1
“You’ve 1
“dumpy,” 1
“he 1
“rabbit 1
” 2
a 35
about 7
accusations 1
achieved 1
address 1
after 3
afterward? 1
again. 2
against 1
age 1
all 3
all, 1
all. 1
allow 1
allowing 1
almost 1
aloud 1
aloud: 1
already 3
also 1
always 2
am 2
an 2
and 60
angry 1
answer 1
```

```
answering 1
anymore. 1
anyone 3
anything 2
anything. 1
anyway.” 1
anywhere 1
are 3
are, 1
arguments 1
around 2
arrest. 1
arrived 1
arrived,” 1
as 13
aside 2
ask 1
asked. 1
assorted 1
at 20
ate 1
aunt 1
away. 1
back 3
back. 1
back; 1
bad 1
barked 1
bars 1
bats 1
battle 1
be 8
because 3
become 1
bedroom 1
been 5
before 1
began 1
begun. 1
behind 3
best 2
better 1
big 1
big-boned, 1
bird 1
birthday 2
bit 2
boisterous 1
both 1
bottom 1
box 1
boy 3
boy, 1
brandished 1
break 1
breakfast 1
breakfast. 1
breast 1
bristled. 1
bullying 1
```

```
burgers 1
bushy 1
but 5
by 3
by. 1
cakes, 2
cakes. 1
call 1
called 1
came 4
can 4
carrot 1
chair 1
children 1
chocolate 1
chucked 1
cleared 1
closed 1
closed, 1
come 2
come, 1
comings 1
comments 1
complaint. 1
completely 1
computer 1
conflict. 1
confusion 1
connected 1
connections 1
continued 1
cooking.) 1
could 5
course 1
covered 1
crossed 1
crumpled 1
cup 1
curiously 1
curse 1
curtly. 1
cut 1
cutting 1
darling,” 1
days 1
deal 1
dearly 1
decided 1
deep 1
delivered 1
dentists.) 1
desk 1
desk, 1
did 2
did, 1
didn’t 9
diet 4
disappear 1
disapproval 1
distantly) 1
```

```
distaste. 1
do 4
doing 1
don’t 1
done 1
door 2
door, 1
doorbell 1
doorbell. 1
doughnuts 1
down 7
down. 1
dream; 1
dressed 1
drew 1
drinks 1
each 1
earlier 1
early 1
earth 1
eat 1
elderly 1
else. 1
emptied 1
end 1
end-of-year 1
enormous 2
enormous; 1
enough 3
entered 1
entire 1
envelope 1
especially 1
even 5
ever 1
except 1
excuses 1
expected 1
experience 1
explain 1
expression 1
extra 1
extremely 1
eyeing 1
eyes 2
eyes. 1
face 3
face. 1
fact 2
family 2
far 1
fat, 1
favorite 1
fear 1
feeble, 1
feel 3
feet, 1
few 1
fight 1
filled 1
```

```
final 1
finally 1
finally. 1
find 3
fingerprints 1
finished 3
fireplace 1
five 1
fizzy 1
flash, 1
flashed. 1
floor, 1
floorboard 1
fly!” 1
folded 1
follow 1
followed 1
following 1
food, 1
food. 1
food.” 1
for 17
forward 1
found 1
four 1
fridge, 1
friend 1
friends 1
from 10
from, 1
front 1
front, 1
frowned. 1
fruit 1
fruitcake 1
full 3
fundamental 1
funny. 1
furious 2
fuss 1
gamekeeper, 1
games 1
gave 2
get 6
gets 1
gifted 1
give 1
glad 1
glancing 1
glare. 1
glared 1
glaring 1
gleaming 1
glowered 1
go 2
going 3
goings 1
gold, 1
got 7
grapefruit 6
```

```
grapefruit. 1
great 2
gritted 1
growing 1
growled 1
growled. 1
grumpily 1
had 34
hadn’t 1
hair?” 1
hall. 2
hand 1
hand, 1
happened 2
happy, 1
hard 1
hardly 1
has 2
hasn’t 2
hated 1
have 10
having 1
he 42
he? 1
he’d 1
heard 3
heaved 1
held 1
hello 1
help, 1
her 3
her,” 1
her. 1
here. 1
hidden 2
him 8
him, 1
him? 1
himself 2
himself. 1
his 31
holidays, 1
home 1
homemade 1
hope 2
hoped, 1
horselike 1
hosted 1
house 1
house, 1
house. 1
household. 1
how 2
however, 1
hunt¬ing 1
hurt 3
husband, 1
idea 1
if 3
ignored) 1
```

```
in 13
inch 1
increased. 1
insisted 2
instead 1
instincts 1
interested 1
into 7
introduced, 1
is 2
is. 1
isn’t 1
it 13
it, 1
it?” 1
journey. 1
just 3
keep 2
keeps 1
kettle, 1
killer 1
kitchen 1
kitchen, 1
knew 1
knickerbockers 1
know 5
knows 1
lack 1
laid 2
large 3
large, 1
last 2
laugh. 1
laughing, 1
least, 1
left, 1
letter 4
letter. 3
letters 1
life 1
lifetime. 1
like 1
lips 1
little 3
lived 1
livid. 1
living 1
long 2
look 2
look, 1
looked 8
looking 2
loose 1
lot 1
loved 1
magnificently. 1
mainly 1
maintained 1
make 3
make. 1
```

```
making 1
managed 2
many 3
marching 1
marks 1
match, 1
matter 1
me 1
me. 1
meat 1
meeting 1
mentioned 1
merely 1
might 4
mind 1
minute 1
minute, 1
mistake 1
moment 1
moment. 1
money 1
morale 1
more 2
more. 1
morning, 2
morning’s 1
most 2
mother 1
mother, 1
movement 1
much 2
mustache 1
mustache, 1
mustache: 1
my 4
name 1
nancy 1
near 1
needed 2
needing 1
neighbors 1
neutral. 1
never 2
new 1
next 1
night, 1
no 2
nodded 1
normal 1
not 5
noticed 1
noticed,” 1
nourishment, 1
now 2
now, 1
nurse 3
obliged 1
observing 1
occasion 1
occupied 1
```

```
of 52
off 5
okay, 1
on 8
on, 1
on. 2
once 1
once-in-a-lifetime 1
one 1
onto 2
opened 1
opportunity; 1
or 2
ordinary. 1
other 1
our 1
out 6
out; 1
outfitters 1
over 3
owl, 1
own 4
paper 3
parchment 2
parents 1
passed 1
people 2
picked 1
piece 1
pieces 1
pies. 1
piggy 1
place 1
plate 1
play 1
pleas 1
plenty 1
pocket 1
pocket, 1
point 1
pointedly 1
politely 1
possible 1
postman 2
postman? 1
poundage 1
prime 1
pronounce 1
puppy 1
purple 2
pursed 1
put 4
putting 1
puzzled. 1
quarter 3
quarter. 1
quarters, 1
quickly 1
rang 1
rang. 1
```

```
reached 1
read 1
reading, 1
ready 1
real 1
really 3
really, 1
received 1
reckon 1
recover 1
red 2
reflection, 1
refused 1
regime 1
remainder 1
remained 1
remember 1
report 2
report. 1
reread 1
rest 2
returned 1
returned. 1
rich 1
rid 1
right. 1
ripping 1
risen 1
rock 1
room 1
room, 1
room. 2
roughly 1
ruffled 1
sack 1
safely 1
said 7
said, 3
same 1
sat 1
say 3
saying 1
scar 1
scars 1
school 6
school. 1
screwed 1
seated 1
see 3
see: 1
seemed 2
seemed, 1
seen 2
send 2
sent 4
set 2
settled 1
severe 2
sharp 1
sharply 1
```

```
she 2
she’s 1
sheet 1
shook 1
should 1
side 1
sigh, 1
signature 1
silence. 1
simply 1
since 2
sincerely, 1
size 1
skated 1
slight 1
slightly 1
smaller 1
smuggling 1
snacks. 1
sniff 1
so 2
so, 1
some 1
somehow 1
someone 2
something 3
something, 2
sometimes 1
son 2
son, 1
soon, 1
sort 2
sorts 1
sound 2
sounds 1
sour 1
space 1
spasm 1
spoon. 1
spotting 1
square 2
squeezed 1
stamps 3
stamps, 1
staring 1
started 1
stay 1
sticks, 1
still 2
stock 1
stole 1
strain 1
stretched, 1
struggled 1
stuffed 1
stupid 1
stupid, 1
sugar-free 1
sulky, 1
summer 4
```

```
superb 1
supposed 1
sure 4
survive 1
swotty 1
table 2
table. 1
take 2
taken 1
takes 1
taking 1
talking 1
tall 1
tantrums, 1
taped 1
teachers 1
teapot 1
tearfully. 1
tears 1
teeth. 2
temper 2
term.” 1
terrified 1
tested 1
than 5
that 23
the 95
them 3
them. 1
then 3
then,” 1
then?” 1
there 1
therefore 1
these; 1
they 3
they’d 1
thing 2
things 2
things. 1
think 1
thinking 1
thirteen 1
thirty 1
this 9
this,” 1
though 4
though. 1
thought 3
threatening 1
three 1
three, 1
through 3
tickets 2
time 2
time, 2
to 59
to. 2
told 2
too 5
```

```
too. 1
took 1
touched 1
touchy 1
train 2
treat 1
tremulous 1
tried 1
trying 2
turn 3
turned 1
turning 1
two 3
uncle’s 2
under 3
understand 2
unpleasant 1
unpleasant. 1
unsweetened 1
up 10
up. 1
upstairs, 1
upstairs. 1
us 2
usual 1
usual. 1
usual: 1
vegetables 1
very 3
wailed 1
wait 1
waited 1
walls 1
walls, 1
want 2
wardrobe 1
was 34
was. 1
way 2
way, 1
we’ve 1
weeks 1
weight 1
weird 1
well-chosen 1
well. 1
were 6
whale. 1
what 5
what?” 1
when 7
where 3
which 4
while 2
who 5
whole 1
whose 1
why 2
wider 1
will 1
```

```
wind 1
window. 2
with 15
without 1
woman?” 2
wondering 1
worried. 1
worst 1
would 7
wouldn’t 1
writing 2
writing. 1
years 1
years, 1
years. 1
yesterday. 1
you 7
you, 1
you.” 1
young 1
your 2
```

Question 2

In [18]:
```
pip install pyspellchecker
```

```
Requirement already satisfied: pyspellchecker in c:\users\saigo\anaconda3\lib\site-pa
ckages (0.8.1)
Note: you may need to restart the kernel to use updated packages.
```

In [19]:
```
from spellchecker import SpellChecker
spell = SpellChecker()
```

In [20]:
```
from nltk.tokenize import word_tokenize
import string
import codecs
import re
with codecs.open("english3.txt", 'r', encoding='utf-8',errors='ignore') as f1:
    f1 = f1.read()
    f1v = word_tokenize(f1)
```

In [21]:
```
from nltk.tokenize import word_tokenize
def tokenize(word):
    tokens = word_tokenize(word)
    tokens = [w.lower() for w in tokens]
    return set(tokens)
```

In [22]:
```
def mapper(txt):
    for word in tokenize(txt):
        yield(word, 1)

def word_count(txt):
    x=[]
    for word, count in mapper(txt):
        if word not in f1v:
            print(word)
```

```
            x.append(word)
    return x
```

In [23]:
```
file2 = open("file2.txt", "r" , encoding = "utf8")
file2 = file2.read()
li=word_count(file2)
```

```
ll
down.
209
moody.
right-hand
discontinued.
slytherin
;
horror-struck
j.k.
205
hogwarts
beauxbatons
thirty-
quidditch
204
!
whole-hearted
triwizard
–
.
yo-yos
hagrid
wouldn
201
208
house-elf
“
couldn
’
’
er-my-knee
wasn
rowling
albus
?
durmstrang
(
)
names.
206
207
hogsmeade
year.
move.
202
mr.
hmph
weasley
it.
|
isn
champion.
inter-house
ever-bashing
house-elves
203
hundred.
short-listed
...
```

```
—
out-of-bounds
less-than-
www.ztcprep.com
ve
underage
money.
”
:
mad-eye
```

In [24]:
```
d ={}
for i in range(len(li)-1):
    x=li[i]
    c=0
    for j in range(i,len(li)):
        if li[j]==li[i]:
            c=c+1
    count=dict({x:c})
    if x not in d.keys():
        d.update(count)
print (d)
```

```
{'ll': 1, 'down.': 1, '209': 1, 'moody.': 1, 'right-hand': 1, 'discontinued.': 1, 'sl
ytherin': 1, ';': 1, 'horror-struck': 1, 'j.k.': 1, '205': 1, 'hogwarts': 1, 'beauxba
tons': 1, 'thirty-': 1, 'quidditch': 1, '204': 1, '!': 1, 'whole-hearted': 1, 'triwiz
ard': 1, '–': 1, '.': 1, 'yo-yos': 1, 'hagrid': 1, 'wouldn': 1, '201': 1, '208': 1,
'house-elf': 1, '“': 1, 'couldn': 1, ',': 1, '’': 1, 'er-my-knee': 1, 'wasn': 1, 'row
ling': 1, 'albus': 1, '?': 1, 'durmstrang': 1, '(': 1, ')': 1, 'names.': 1, '206': 1,
'207': 1, 'hogsmeade': 1, 'year.': 1, 'move.': 1, '202': 1, 'mr.': 1, 'hmph': 1, 'wea
sley': 1, 'it.': 1, '|': 1, 'isn': 1, 'champion.': 1, 'inter-house': 1, 'ever-bashin
g': 1, 'house-elves': 1, '203': 1, 'hundred.': 1, 'short-listed': 1, '...': 1, '—':
1, 'out-of-bounds': 1, 'less-than-': 1, 'www.ztcprep.com': 1, 've': 1, 'underage': 1,
'money.': 1, '”': 1, ':': 1}
```

In [ ]:
```

```