

In [2]: !pip install PyPDF2

```
Collecting PyPDF2
  Downloading pypdf2-3.0.1-py3-none-any.whl (232 kB)
      232.6/232.6 kB 2.1 MB/s eta 0:00:00
Installing collected packages: PyPDF2
Successfully installed PyPDF2-3.0.1
```

In [4]: !pip install pyenchant

```
Collecting pyenchant
  Downloading pyenchant-3.2.2-py3-none-any.whl (55 kB)
      55.7/55.7 kB 927.2 kB/s eta 0:00:00
Installing collected packages: pyenchant
Successfully installed pyenchant-3.2.2
```

In [33]: import PyPDF2

```
# My Date of birth is 26th October 2000, so defining my birth month, date and year
birth_month = 10
birth_date = 26
birth_year = 100

# Calculating the pages to extract data for file1.txt and file2.txt
file1_start_page = birth_date
file1_end_page = birth_date + 9
file2_start_page = birth_year
file2_end_page = birth_year + 9

# Function to extract text from a range of pages
pdf_path = '/content/Harry_Potter_(www.ztcprep.com).pdf'
pdf_file = open(pdf_path, 'rb')
pdf_reader = PyPDF2.PdfReader(pdf_file)
def extract_text_from_range(pdf_reader, start_page, end_page):
    text = ''
    for page_num in range(start_page, end_page+1):
        page = pdf_reader.pages[page_num]
        text += page.extract_text()
    return text

# Writing the text into the files
file1_text = extract_text_from_range(pdf_reader, file1_start_page, file1_end_page)
file2_text = extract_text_from_range(pdf_reader, file2_start_page, file2_end_page)
with open('file1.txt', 'w') as file1:
    file1.write(file1_text)
with open('file2.txt', 'w') as file2:
    file2.write(file2_text)
pdf_file.close()
```

In [35]:

```
def mapred(file1_path):
    wordcount = {} # Initiating wordcounts variable to count word occurrences
    with open(file1_path, 'r', encoding='utf-8') as f1:
        for line in f1:
            mapped_data = map(line)
            wordcount = reduce(wordcount, mapped_data)
    return wordcount

# Defining the map and reduce functions
def map(line):
    count = {} # Creating count dictionary to get the word and no.of its occurrences
    line = line.strip()
    words = line.split()
```

```
    for word in words:
        word = word.lower()
        count[word] = count.get(word, 0) + 1
    return count

def reduce(past_count, new_count):
    for word, count in new_count.items():
        past_count[word] = past_count.get(word, 0) + count
    return past_count

if __name__ == "__main__":
    file1_path = "file1.txt"
    wordcount = mapred(file1_path)
    for word, count in sorted(wordcount.items()):
        print(f"{word}:{count}")
```

,:13
,":4
.:15
."":4
15:1
16:1
17:1
18:1
19:1
20:1
21:1
?":1
a:47
about:3
above:1
after:2
again.:1
age.:1
ago,:1
air:2
all:7
all.:1
allowed,:1
almost:4
always:1
anashig:1
and:51
and,:1
another:1
an':2
arm:1
arms:2
around:2
around.:1
as:16
asked:1
asleep:2
asleep.:1
astonishing:1
astride:1
at:9
aunt:4
awake:1
away:2
baby:3
back:5
bacon.:1
bag:1
ball:1
balls:1
be:4
beach:1
beard:1
because:2
bed:2
been:7
before:1
before.:1
being:3
beneath:1
bent:2
beside:1
best:1
better:1

bicycle,:1
big:1
bigger:1
bike:2
bike.:1
birthday:3
black:3
blankets:2
blankets,:1
blankets.:2
blew:1
blinked:1
blond:1
bolt:1
bonnets:1
boots:1
both:1
bottles,:1
boy:5
brass:1
breeze:1
bright:1
bringing:1
bristol.”:1
broken:1
bundle:3
bundle;:1
burn,:1
burying:1
bushy:1
business:1
but:9
by:3
c-c-:1
came:1
can:2
can’:2
carefully:1
careless.:1
carousel:1
cat:1
catch:1
celebrations.”:1
changed:2
clicked:1
climbing:1
cloak:1
cloak,:2
closed:1
clothes:1
come:1
computer:2
corner:2
could:5
could,:1
couldn’:2
country:1
course:1
course.:1
cousin:1
crept:1
cupboard:1
cupboard,:1
curiously:1

cut,:1
dare:1
dark:1
day:1
dead:1
demanded.:1
destroyed,:1
did:2
didn':1
dif:1
do:2
does:1
dog.:1
dolphins.:1
don':1
door:5
door,:1
doorstep,:1
down:3
dream:2
dressed:1
drive:2
drive,:1
duddy':1
dudley:6
dudley':4
dumbledore:5
dumbledore,:4
dumbledore.:2
dumbledore?":2
dumbledore':1
dursley:2
dursleys:1
dursleys':2
dursley':1
e:8
eah,":1
ell:1
ell,:1
ell,":1
en:1
end:1
engine:1
es:1
es,:1
es,":1
et:1
even:1
everything:1
exactly:2
exercise:1
expect:1
expect,:1
eyed:1
eyes:2
e've:1
f:4
face:1
face,:2
fair:1
famous,:1
fast:1
fast.:1
fat:1

fateful:1
father:1
favorite:1
feeling:1
feet:1
fell:2
ferent-colored:1
few:2
finally:1
find:1
first:2
five:1
fled:2
flying:1
flyin':1
for:6
forehead:1
forever:1
forward:1
found:1
found,"":1
four:3
from:2
front:6
frying:1
full:2
funny:1
furiously:1
g:7
game:1
garden:1
gardens:1
gave:1
ge:2
ge,:1
gently:1
get:6
getting:1
giant,:1
gingerly:1
give:1
glass:1
glasses:1
glowed:1
go:1
gone:1
gone.:1
good:1
good-bye:1
got:3
gotten:1
gotten?:1
great,:1
green:1
grew:1
grip:1
groaned.:1
ground.:1
growing:1
grudgingly:1
g'night,:1
had:19
hagrid:7
hagrid,:2

hagrid.:1
hagrid':1
hair:3
hall:1
hand:1
handkerchief:1
hands:1
handy:1
happen.:1
hardly:1
harry:27
harry':1
hated:1
have:5
having.:1
he:38
head:1
headlight;:1
heard:1
heart:1
hedges:1
heel:1
held:1
her:6
here,:2
here.:1
he':2
he'd:1
hid:1
hidden:1
hiding:1
him:4
him,:2
him.:1
him.”:1
himself:1
his:26
hissed:1
holding:2
hours':1
house:1
house,:1
house.:1
how:3
howl:1
huge:1
huge,:1
hugged:1
hushed:1
i:10
if:2
important:1
in:15
inky:1
inside:2
inside,:1
into:5
involved:1
iping:1
is:2
isn':1
it:17
it,:3
it.:4

it?":1
it':1
i've:1
j.k.:7
jacket:1
james:1
jet-black:1
join:1
just:2
kicked:1
kiss.:1
kissed:1
kitchen:1
kitchen.:1
knee:1
knees,:1
knobbly:1
know:1
knowing:3
laid:1
lamps:1
landed:1
lar:3
last:1
last.:1
lay:1
least:1
leather:1
left:1
lent:1
let:2
letter:2
lids,:1
life,":1
life;:1
light:2
lightning.:1
like:4
lily:1
lit:1
little:2
live:1
lived:1
lived!":1
living:2
london:1
long:1
long.:1
longer:1
look:2
looked:7
looking:1
lots:1
louder:1
low:2
luck,:1
made:1
make:1
man:2
mantelpiece:1
map:1
may:1
mcgonagall:7
mcgonagall,:1

mcgonagall,"":1
mcgonagall.:1
me.:1
meeting:1
mention:1
might:1
milk:1
mind,:1
minute:1
moment,:2
most:1
mother:1
motorcycle:5
motorcycle?":1
mouth,:1
move:1
mr:1
mrs.:1
much:1
muf:1
muggles:2
muggles!":1
murmured.:1
muscular:1
must:1
my:2
myself:1
mystery:1
nearly:1
neat:1
nephew:1
new:1
news:1
next:1
night:1
night.:1
no:3
nodding:1
noise:1
nor:1
normal:1
nose:1
not:7
nothing:2
now:1
now!":1
number:2
o:1
of:33
often:1
old:1
on:18
on,:2
on.:1
once,:1
one:3
one.:1
only:1
onto:2
opened:2
or:1
orange:1
other:2
ou:1

oung:1
out:9
out.:1
outer:1
outside:1
over:8
owls.:1
p:7
pair:1
pan:1
passed:1
passed.:1
patting:1
people:1
perfect:2
perhaps:1
petunia:1
philosophers:7
photographs:2
pictures:1
pinched:1
pink:1
place:1
place,:1
playing:1
poor:1
potter:10
presents.:1
pretend:1
privet:3
problems,:1
prodged:1
professor:12
pulling:1
punching:2
put:3
put-:1
racing:2
rapped:1
ready:1
really:2
relieved.:1
remember:1
reply:1
report:1
riding:1
right:2
right,:1
road:1
roar:2
rolled:2
room:1
room,:1
rose:2
rowling:7
ruf:1
rumbling:1
s:13
s,:1
sad,:1
said:9
said,:1
same:3
say:1

say?":1
saying:2
scar:1
scars:1
scratchy:1
scream:1
screeched.:1
second:1
secret:1
see:3
seemed:1
seen:1
shaggy:1
shall:1
shaped:1
she:5
shone:1
shook,:1
shoulders:1
showed:2
shrill:1
sign:2
silence:1
silent:1
silver:1
simply:1
since:1
sir:4
sir?":1
sirius:1
sitting:1
size:1
skinnier:1
skinny:1
sky:2
sleeve,:1
slept:1
slept.:1
slinking:1
slowly:1
small:2
smaller:1
snapped:1
so:2
sobbed:1
socks.:1
some:1
somebody:1
something:3
soon,:1
sound:2
sounding:1
special,:1
sped:1
spend:1
spider:1
spiders,:1
spoke.:1
spotted:1
stairs:1
stand:1
start.:1
started:2
staying:1

steadily:1
step:1
step,:1
stepped:1
still:1
stone:7
stood:1
stopped:1
stove.:1
streaming:1
street:2
street.:2
suddenly:3
sun:1
swallowed,:1
swarmin':1
swelled:1
swish:1
swung:1
t:9
t.:1
tabby:1
table:1
take:1
taking:1
tall:1
tangles:1
television:1
ten:1
tend:1
ter:1
than:2
that:13
that.:1
that?":1
the:91
their:5
them:2
them,:2
them.:2
then:3
then,:1
there:2
there,:1
there?":1
they:3
thin:1
things:1
think:1
this:3
this?":1
though:2
thought:1
three:1
through:1
tidy:2
time:2
times:2
to:23
too:1
too.:1
took:3
toward:2
trash:1

tried:1
trust:2
tucked:1
tuft:1
turned:3
twelve:1
twice:1
twinkling:1
two.:1
under:5
underneath:1
unless:1
until:1
up:7
up!:1
up.:1
used:1
usually:1
v:1
vast,:1
very:7
visible,:1
voice:1
voice,:1
voices::1
w:4
wake:1
waking:1
walked:2
walking:1
wall:1
want:2
wanted:1
wanted,:1
was:28
was.:1
we:1
wear:1
wearing:1
weeks:1
well:1
went:1
were:4
we'd:1
we'll:1
what:3
when:2
where:3
which:2
whiskery:1
whispered:1
whispered,:1
who:1
why:1
wide.:1
wild:1
wise:1
with:8
with.:1
without:1
woke:1
woken:2
would:4
wouldn':1

wounded:1
 www.ztcprep.com:10
 y:1
 years:2
 yes,:2
 yet?":1
 you:9
 yourself,:1
 you're:1
 |:7
 -:7
 -:18
 -":1
 "are:1
 "at:1
 "borrowed:1
 "but:2
 "could:1
 "couldn':1
 "even:1
 "good:1
 "hagrid,":1
 "hagrid':1
 "he'll:1
 "i:2
 "is:1
 "i'd:1
 "i'm:1
 "nearly:1
 "no:1
 "no,:1
 "nothing,:1
 "s-s-sorry:1
 "shhh!":1
 "t:1
 "that':1
 "up!:1
 "up!":1
 "w:2
 "what:1
 "y:6
 "you'll:1
 ...:1
 ...":1

In [16]: !pip install pyspellchecker

Collecting pyspellchecker
 Downloading pyspellchecker-0.8.1-py3-none-any.whl (6.8 MB)
 6.8/6.8 MB 21.2 MB/s eta 0:00:00
 Installing collected packages: pyspellchecker
 Successfully installed pyspellchecker-0.8.1

In [36]: *# As my birth year is 2000 I am selecting the pages 100-109 in the book*
Comparing each word occurrence with the ukenglish words data downloaded from http:
 def english_words(file_path):
 eng_words = set()
Opening the file in read mode
 with open(file_path, 'r', encoding='latin-1') as engfile:
 for line in engfile:
 word = line.strip()
 eng_words.add(word.lower())
 return eng_words

```

# defining the function to solve the question to find non-english words in the page
def non_english_words(file_path, eng_words):
    non_eng = {} # Initiating a dictionary to store the non-english words and its r
    with open(file_path, 'r', encoding='latin-1') as file:
        for line in file:
            line = line.strip()
            words = line.split()
            for word in words:
                word = word.lower()
                if word not in eng_words:
                    non_eng[word] = non_eng.get(word, 0) + 1
    return non_eng

if __name__ == "__main__":
    eng_words = english_words("/content/ukenglish.txt")
    file_path = "file2.txt"
    non_english_words_found = non_english_words(file_path, eng_words)
    for word, count in sorted(non_english_words_found.items()):
        print(f"{word}: {count}")

```

,: 15
,â: 3
. : 14
.â: 1
64: 1
65: 1
66: 1
67: 1
68: 1
69: 1
70: 1
?â: 3
albus: 2
angry?â: 1
anything?â: 1
anâ: 9
ap.: 3
arenâ: 1
asked,: 1
bank.: 1
be?: 1
beak.: 1
beard.: 1
bit,: 1
boat,: 1
boat.: 1
boots.: 1
born.: 1
bottom,: 1
boy!: 1
business.â: 1
cake,: 1
change,: 1
chuckled.: 1
coat.: 2
coins,: 1
coins.: 1
cupboard.â: 1
cupboard?: 1
daylight,: 1
deliverinâ: 1
diagon: 1
didnâ: 6
do.â: 1
donâ: 1
dream,â: 1
dream.: 1
dudleyâ: 1
dumbledore: 2
dumbledore.: 1
dumbledore.â: 1
dumbledââ: 1
eah: 2
eh?: 1
ernon: 4
ernon.: 1
ernon;: 1
everything.: 1
everythinâ?: 1
expelled,: 1
expelled?â: 1
eyebrows.: 1
eyes,: 1
eyes.: 1

f,: 1
fact,: 1
feet,: 1
fer: 7
fetchinâ: 1
fight.: 1
fire.: 1
football?: 1
gang,: 1
ge: 1
gettinâ: 2
go,: 1
goblins,: 1
goblins.â: 1
going?â: 1
goinâ: 1
gotta: 2
gringotts: 3
gringotts.: 2
had,: 1
hadnâ: 3
hagrid: 14
hagrid,: 4
hagrid.: 3
hagrid?â: 1
hagridâ: 3
haircut,: 1
havenâ: 1
he,: 1
head,: 1
head.: 1
here?â: 1
heâ: 2
heâd: 3
heâll: 3
him,: 2
him,â: 2
him.: 4
him?: 2
himself.: 1
hissed.: 1
hogwarts: 3
hogwarts!: 1
hogwarts,â: 1
hogwarts.: 1
hogwarts.â: 1
holding.: 1
house,: 1
humbugs,: 1
insul: 1
it,: 1
it.: 2
it?: 1
izards: 1
izardsâ: 1
iâll: 2
iâve: 2
j.k.: 7
keys,: 1
knuts,â: 1
lar: 1
light,: 1
mad.: 1
magic,: 1

magic.â: 1
magic?â: 1
man,: 1
me!â: 1
me,: 1
meself: 1
mistake.: 2
morning.: 1
muggle: 1
mumbled,: 1
nah,: 1
nameâ: 1
noise.: 1
on,: 1
one.: 1
ones.â: 1
open.: 1
ou: 1
oâ: 5
pain.: 1
payinâ: 1
pellets,: 1
pig,: 1
pigâ: 1
pockets.â: 2
proud,: 1
puncture.: 1
revenge,: 1
right,â: 1
right.â: 1
roared.: 1
rock.: 1
room,: 1
rowling: 7
said.: 2
sausage,: 1
school.â: 1
second,: 1
see.: 1
sinking.: 1
smiling,: 1
sofa,: 1
sofa.: 1
sort,: 1
speakinâ.: 1
squeal,: 1
storm.: 1
strange-looking: 1
stretched.: 1
string,: 1
stuf: 3
sunlight,: 1
sunlight.: 1
surprise,: 1
sâpposed: 1
teabags: 1
teh: 1
ter: 19
that,â: 2
that.: 1
that.â: 2
them,: 1
them.: 2
then.â: 1

ther: 1
there,: 1
theyâd: 1
theyâre: 1
this.: 1
thought,: 1
thundered,: 1
tight.: 1
town,: 1
tricks!â: 1
trousers.: 1
truth.: 1
up,: 1
up.: 1
up.â: 1
wait,: 1
wasnâ: 2
weâll: 1
weâve: 1
wizard,: 2
wizard.â: 1
wizard?: 1
wizards.: 1
wonâ: 3
world,: 1
world.: 1
wouldnâ: 1
www.ztcprep.com: 10
yeh: 7
yeh.â: 1
yehâd: 1
yer: 5
ying: 1
youâll: 1
|: 7
â: 7
â: 24
ââ: 5
â: 2
âcept: 1
â: 2
âall: 1
âbe: 1
âbest: 1
âbut: 2
âdonâ: 3
âdâyeh: 1
âflew: 1
âgive: 1
âgot: 1
âgotta: 1
âhagrid!â: 1
âhagrid,â: 1
âharry: 1
âhavenâ: 1
âhe: 2
âheâ: 1
âhow: 1
âi: 4
âif: 1
âit: 1
âitâ: 1
âiâm: 2
âjust: 1

âknuts?â: 1
âmm?â: 1
ânever: 1
ânot: 1
âoh,: 1
âpay: 1
âsee?â: 1
âshouldnâ: 1
âstop: 1
âthe: 1
âthereâ: 1
âthey: 1
âum: 1
âw: 1
âwhat?â: 1
âwhy: 2
ây: 3
ââ: 1
â|: 6

In []: