

Map Reduce of Harry Potter

Created by Lilly on September 23rd, 2025

I extracted 10 pages from page 669 to 679 in the Harry Potter starting from Page 100 in Book 2.

```
In [9]: # open the file, harrypotter_LSA and inspect it
with open('harrypotter_LSA_extract.txt', 'r') as file:
    # read the contents of the file
    contents = file.read()
    # print the first 10 lines of the file
    # print(contents)
    # split the contents into lines
    lines = contents.split('\n')
    # print the first 5 lines
    for line in lines[:5]:
        print(line)
    # print the number of lines in the file
    print(f'\nThe file has {len(lines)} lines.')
```

Sprout, who was looking distinctly disgruntled, not at all her usual cheerful self.

There was a murmur of interest. They had only ever worked in greenhouse one before – greenhouse three housed far more interesting and dangerous plants. Professor Sprout took a large key from her belt and unlocked the door. Harry caught a whiff of damp earth and fertilizer mingling with the heavy perfume of some giant, umbrella-sized flowers dangling from the ceiling. He was about to follow Ron and Hermione inside when Lockhart's hand shot out.

"Harry! I've been wanting a word – you don't mind if he's a couple of minutes late, do you, Professor Sprout?"

The file has 102 lines.

Create the necessary Map & Reducer helper scripts

These scripts will help:

- Mapper: maps each of the words in the line to a '1'
- Shuffle: this is an intermediary step that shuffles all the key,value pairs into a dictionary
- Reducer: reduces the the keys to a singular value (a sum) of all the words found

```
In [ ]: # import necessary libraries
from collections import defaultdict # to create a dictionary of lists
from collections import Counter # to count the number of words
```

In [61]: *# helper functions*

```
# mapper – for every word in a line, create a tuple (word, 1)
def mapper(line):
    for word in line.strip().split():
        yield (word.lower(), 1)

# shuffle (intermediate step) – for every key, value pair from the mapper, c
# create a dictionary where the key is the word and the value is a list of c
def shuffle(mapped):
    shuffled = defaultdict(list)
    for key, value in mapped:
        shuffled[key].append(value)
    return shuffled

# reducer – for every key, value pair from the shuffle step, sum the counts
def reducer(key, values):
    return (key, sum(values))
```

In [60]: *# main function to run the MapReduce framework on the Harry Potter file and*
it will read the file and then call the mapper, shuffle, and reducer funct

```
# run the script
with open("harrypotter_LSA_extract.txt") as f:
    lines = f.readlines()

mapped = []
for line in lines:
    mapped.extend(mapper(line))

# Shuffle
grouped = shuffle(mapped)

# Reduce
results = [reducer(k, v) for k, v in grouped.items()]

# Print a sorted list of results
results.sort(key=lambda x: x[1], reverse=True)

for word, count in results:
    print(word, count)
```

the 72
a 56
of 48
and 45
to 41
was 30
harry 26
– 25
in 22
he 22
his 19
as 19
you 18
i 18
said 17
it 17
had 14
professor 14
her 12
but 12
for 12
she 11
have 11
all 10
they 10
out 10
that 10
at 9
harry, 9
who 8
ron 8
when 8
by 8
your 8
on 8
with 7
hermione 7
p 7
g 7
e 7
| 7
potter 7
chamber 7
secrets 7
j.k. 7
rowling 7
be 7
were 7
into 7
is 7
their 7
sprout 6
" 6
didn't 6
there 5
one 5

more 5
been 5
what 5
are 5
it's 5
everyone 5
over 5
like 5
wand 5
large 4
about 4
lockhart's 4
hand 4
course, 4
went 4
wasn't 4
see 4
just 4
right 4
mandrakes 4
tell 4
sprout. 4
earmuffs 4
ron, 4
beetle 4
not 3
only 3
greenhouse 3
from 3
if 3
minutes 3
do 3
"when 3
could 3
know 3
flying 3
stood 3
how 3
got 3
me 3
want 3
first 3
time 3
you're 3
an 3
i'd 3
gave 3
hermione's 3
or 3
though 3
ten 3
mandrake 3
up 3
another 3
here 3
little 3

looked 3
pair 3
them 3
sure 3
will 3
own 3
plant 3
him 3
off 3
i'm 3
down 3
back 3
looking 2
usual 2
before 2
far 2
took 2
damp 2
some 2
shot 2
i've 2
he's 2
you, 2
lockhart 2
said, 2
door 2
lockhart, 2
teeth 2
head. 2
"harry, 2
completely 2
heard 2
well, 2
no 2
idea 2
talking 2
say 2
so 2
on, 2
knew 2
once 2
done 2
every 2
even 2
"gave 2
taste 2
"oh, 2
-" 2
shoulder. 2
"i 2
go 2
head 2
can't 2
get 2
yes, 2
him, 2

famous 2
wizard 2
much 2
few 2
people 2
harry's 2
quite 2
supposed 2
place 2
hermione, 2
can 2
"the 2
again. 2
take 2
we 2
still 2
very 2
spoke, 2
tufty 2
growing 2
tried 2
put 2
make 2
give 2
up. 2
snapped 2
let 2
small, 2
extremely 2
earth. 2
leaves 2
top 2
compost 2
several 2
back, 2
tray 2
pots 2
sharp 2
dark 2
boy 2
always 2
howler 2
read 2
made 2
think 2
trying 2
transfiguration. 2
last 2
seemed 2
lunch 2
step 2
sprout, 1
distinctly 1
disgruntled, 1
cheerful 1
self. 1

murmur 1
interest. 1
ever 1
worked 1
three 1
housed 1
interesting 1
dangerous 1
plants. 1
key 1
belt 1
unlocked 1
door. 1
caught 1
whiff 1
earth 1
fertilizer 1
mingling 1
heavy 1
perfume 1
giant, 1
umbrella-sized 1
flowers 1
dangling 1
ceiling. 1
follow 1
inside 1
out. 1
"harry! 1
wanting 1
word 1
don't 1
mind 1
couple 1
late, 1
sprout?" 1
100 1
judging 1
sprout's 1
scowl, 1
did 1
mind, 1
"that's 1
ticket," 1
closed 1
face. 1
"harry," 1
white 1
gleaming 1
sunlight 1
shook 1
harry." 1
nonplussed, 1
nothing. 1
my 1
fault. 1

kicked 1
myself." 1
about. 1
"don't 1
shocked. 1
car 1
hogwarts! 1
why 1
you'd 1
it. 1
mile. 1
harry. 1
remarkable 1
show 1
those 1
brilliant 1
talking. 1
publicity, 1
i?" 1
lockhart. 1
bug. 1
onto 1
front 1
page 1
paper 1
couldn't 1
wait 1
again." 1
no, 1
professor, 1
harry," 1
reaching 1
grasping 1
understand. 1
natural 1
bit 1
you've 1
blame 1
myself 1
giving 1
that, 1
because 1
bound 1
here, 1
young 1
man, 1
start 1
cars 1
try 1
yourself 1
noticed. 1
calm 1
down, 1
right? 1
plenty 1
older. 1

thinking! 1
'it's 1
internationally 1
already!' 1
101 1
twelve, 1
nobody 1
now. 1
fact, 1
nobody! 1
mean, 1
haven't 1
they? 1
business 1
he-who-must-not-be-named!" 1
glanced 1
lightning 1
scar 1
forehead. 1
know, 1
good 1
winning 1
witch 1
weekly's 1
most-charming-smile 1
award 1
five 1
times 1
row, 1
start, 1
start. 1
hearty 1
wink 1
strode 1
off. 1
stunned 1
seconds, 1
then, 1
remembering 1
greenhouse, 1
opened 1
slid 1
inside. 1
standing 1
behind 1
trestle 1
bench 1
center 1
greenhouse. 1
twenty 1
pairs 1
different-colored 1
ear-muffs 1
lying 1
bench. 1
taken 1

between 1
"we'll 1
repotting 1
today. 1
now, 1
properties 1
mandrake?" 1
nobody's 1
surprise, 1
air. 1
"mandrake, 1
mandragora, 1
powerful 1
restorative," 1
sounding 1
swallowed 1
textbook. 1
"it 1
used 1
return 1
transfigured 1
cursed 1
original 1
state." 1
"excellent. 1
points 1
gryffindor," 1
forms 1
essential 1
part 1
most 1
antidotes. 1
also, 1
however, 1
dangerous. 1
why?" 1
narrowly 1
missed 1
glasses 1
102 1
cry 1
fatal 1
anyone 1
hears 1
it," 1
promptly. 1
"precisely. 1
points," 1
"now, 1
young." 1
pointed 1
row 1
deep 1
trays 1
shuffled 1
forward 1

better 1
look. 1
hundred 1
plants, 1
purplish 1
green 1
color, 1
rows. 1
unremarkable 1
slightest 1
meant 1
"cry" 1
mandrake. 1
"everyone 1
 earmuffs," 1
scramble 1
seize 1
pink 1
fluffy. 1
ears 1
covered," 1
safe 1
remove 1
them, 1
thumbs- 1
on. 1
ears. 1
shut 1
sound 1
completely. 1
pink, 1
fluffy 1
ears, 1
rolled 1
sleeves 1
robes, 1
grasped 1
plants 1
firmly, 1
pulled 1
hard. 1
gasp 1
surprise 1
hear. 1
instead 1
roots, 1
muddy, 1
ugly 1
baby 1
popped 1
pale 1
green, 1
103 1
mottled 1
skin, 1
clearly 1

bawling 1
lungs. 1
pot 1
under 1
table 1
plunged 1
it, 1
burying 1
dark, 1
until 1
tufted 1
visible. 1
dusted 1
hands, 1
thumbs-up, 1
removed 1
 earmuffs. 1
"as 1
our 1
seedlings, 1
cries 1
won't 1
kill 1
yet," 1
calmly 1
she'd 1
nothing 1
exciting 1
than 1
water 1
begonia. 1
"however, 1
knock 1
hours, 1
none 1
miss 1
day 1
securely 1
while 1
work. 1
attract 1
attention 1
pack 1
"four 1
supply 1
sacks 1
careful 1
venomous 1
tentacula, 1
teething." 1
slap 1
spiky, 1
red 1
making 1
draw 1
long 1

feelers 1
inching 1
sneakily 1
joined 1
curly-haired 1
hufflepuff 1
sight 1
never 1
spoken 1
to. 1
"justin 1
finch-fletchley," 1
brightly, 1
shaking 1
hand. 1
"know 1
are, 1
potter. 1
... 1
granger 1
everything" 1
(hermione 1
beamed 1
shaken 1
too) 1
"- 1
weasley. 1
car?" 1
104 1
smile. 1
obviously 1
mind. 1
"that 1
something, 1
isn't 1
he?" 1
justin 1
happily 1
began 1
filling 1
dragon 1
dung 1
compost. 1
"awfully 1
brave 1
chap. 1
books? 1
died 1
fear 1
cornered 1
telephone 1
booth 1
werewolf, 1
stayed 1
cool 1
zap 1

fantastic. 1
"my 1
name 1
eton, 1
know. 1
glad 1
am 1
came 1
instead. 1
mother 1
slightly 1
disappointed, 1
since 1
books 1
she's 1
begun 1
useful 1
it'll 1
fully 1
trained 1
family. 1
..." 1
after 1
chance 1
talk. 1
needed 1
concentrate 1
mandrakes. 1
look 1
easy, 1
wasn't. 1
coming 1
earth, 1
seem 1
either. 1
squirmed, 1
kicked, 1
flailed 1
fists, 1
gnashed 1
teeth; 1
spent 1
whole 1
squash 1
particularly 1
fat 1
pot. 1
end 1
class, 1
else, 1
sweaty, 1
aching, 1
covered 1
traipsed 1
castle 1
quick 1

wash 1
then 1
gryffindors 1
hurried 1
mcgonagall's 1
classes 1
hard 1
work, 1
today 1
especially 1
difficult. 1
everything 1
learned 1
year 1
leaked 1
during 1
summer. 1
turning 1
button, 1
managed 1
lot 1
exercise 1
scuttled 1
desktop 1
avoiding 1
wand. 1
105 1
having 1
worse 1
problems. 1
patched 1
borrowed 1
spellotape, 1
damaged 1
beyond 1
repair. 1
kept 1
crackling 1
sparking 1
odd 1
moments, 1
transfigure 1
engulfed 1
thick 1
gray 1
smoke 1
smelled 1
rotten 1
eggs. 1
unable 1
doing, 1
accidentally 1
squashed 1
elbow 1
ask 1
new 1

one. 1
mcgonagall 1
pleased. 1
relieved 1
hear 1
bell. 1
brain 1
felt 1
wring 1
sponge. 1
filed 1
classroom 1
except 1
whacking 1
furiously 1
desk. 1
"stupid 1
useless 1
thing 1
"write 1
home 1
one," 1
suggested 1
volley 1
bangs 1
firecracker. 1
yeah, 1
back," 1
stuffing 1
now 1
hissing 1
bag. 1
' 1
fault 1
-' 1
lunch, 1
where 1
ron's 1
mood 1
improved 1
showing 1
handful 1
perfect 1
coat 1
buttons 1
produced 1
"what've 1
this 1
afternoon?" 1
hastily 1
changing 1
subject. 1
"defense 1
against 1
arts," 1
once. 1

"why, 1
demanded 1
seizing 1
schedule, 1
"have 1
outlined 1
lessons 1
hearts?" 1
106 1
snatched 1
schedule 1
blushing 1
furiously. 1
finished 1
outside 1
overcast 1
courtyard. 1
sat 1
stone 1
buried 1
nose 1
voyages 1
vampires 1
quidditch 1
became 1
aware 1
being 1
closely 1
watched. 1
up, 1
saw 1
mousy-haired 1
he'd 1
seen 1
sorting 1
hat 1
night 1
staring 1
transfixed. 1
clutching 1
ordinary 1
muggle 1
camera, 1
moment 1
bright 1
red. 1
"all 1
right, 1
harry? 1
colin 1
creevey," 1
breathlessly, 1
taking 1
tentative 1
forward. 1
"i'm 1

```
gryffindor, 1
too. 1
d'you 1
would 1
```

In [59]: *# only use the NON-ENGLISH (Harry Potter universe) words*

```
#open the file
with open("engmix.txt", "r", encoding="latin-1") as file:
    common_words = file.read().splitlines()

# convert the list to a set for faster lookup
common_words_set = set(common_words)
print(f'The set has {len(common_words_set)} words.')

# define the characters to exclude; if a ' is found, then truncate the word
exclude = {"'", ".", "'", ',', '-', '|', '!', '?', ';', ':', '(', ')', '[',
# exclude numbers too
exclude.update({str(i) for i in range(10)})

# uncommon words only
uncommon_results = [
    (word, count)
    for word, count in results
    if word.lower() not in common_words_set and not any(ch in word for ch in

print("\nCount of Uncommon Words:")
for word, count in uncommon_results:
    print(word, count)

print(f"\nThe file has {len(uncommon_results)} uncommon words.")
```

The set has 84097 words.

Count of Uncommon Words:

```
harry 26
hermione 7
potter 7
rowling 7
lockhart 2
tufty 2
repotting 1
hufflepuff 1
gryffindors 1
mcgonagall 1
quidditch 1
muggle 1
```

The file has 12 uncommon words.