Class

1. Loading data

In [270…
```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-packages
(3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packages
(from nltk) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.12/dist-packages
(from nltk) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.12/dist-
packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (f
rom nltk) (4.67.1)
```

In [271…
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call dri
ve.mount("/content/drive", force_remount=True).
```

In [272…
```
import os

file_path = '/content/drive/MyDrive/CPE393/books.txt'
print(os.path.exists(file_path))
```

```
True
```

In [273…
```
import nltk
# nltk.download('all')
# nltk.download() #for the missing library
nltk.download('punkt_tab')
dir = '/content/drive/MyDrive/CPE393/'
f = open(dir+'books.txt', encoding='utf-8')
raw = f.read()
tokens = nltk.word_tokenize(raw)
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
```

In [274…
```
print(tokens[1000:1100])
```

```
[',', 'eating', 'banquets', 'in', 'the', 'Great', 'Hall', ',', 'sleep\xading', 'i
n', 'his', 'four-poster', 'bed', 'in', 'the', 'tower', 'dormitory', ',', 'visitin
g', 'the', 'gamekeeper', ',', 'Hagrid', ',', 'in', 'his', 'cabin', 'next', 'to',
'the', 'Forbidden', 'Forest', 'in', 'the', 'grounds', ',', 'and', ',', 'especiall
y', ',', 'Quidditch', ',', 'the', 'most', 'popular', 'sport', 'in', 'the', 'wizar
ding', 'world', '(', 'six', 'tall', 'goal', 'posts', ',', 'four', 'flying', 'ball
s', ',', 'and', 'four\xadteen', 'players', 'on', 'broomsticks', ')', '.', 'All',
'Harry', '’', 's', 'spellbooks', ',', 'his', 'wand', ',', 'robes', ',', 'cauldro
n', ',', 'and', 'top-of-the-line', 'Nimbus', 'Two', 'Thousand', 'broomstick', 'ha
d', 'been', 'locked', 'in', 'a', 'cupboard', 'under', 'the', 'stairs', 'by', 'Unc
le', 'Vernon', 'the', 'instant']
```

2. Regular expression

In [275…
```
tokens = [w.lower() for w in tokens]
```

In [276…
```
import re

[w for w in tokens if re.search('ed$', w)][:10]
```

Out[276…
```
['reserved',
 'published',
 'registered',
 'related',
 'reproduced',
 'stored',
 'transmitted',
 'opened',
 'released',
 'printed']
```

---

1. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **ed$ หมายถึง ค้นหาคำที่ลงท้ายด้วยตัวอักษร "ed"**
ดังนั้นคำที่แสดงออกมาจะเป็นคำที่ลงท้ายด้วย "ed" เช่น reserved, published
ในทางกลับกันคำที่ไม่ตรงกับรูปแบบนี้ก็คือคำที่ไม่ได้ลงท้ายด้วย "ed" เช่น

- education (มี ed อยู่ข้างหน้า ไม่ใช่ท้ายคำ)
- fast (ไม่ลงท้ายด้วย ed)

In [277…
```
[w for w in tokens if re.search('^..j..t..$', w)]
```

Out[277…
```
['dejected',
 'adjusted',
 'rejected',
 'rejected',
 'dejected',
 'adjusted',
 'adjusted',
 'adjusted',
 'adjusted']
```

---

2. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **^..j..t..$ หมายถึง ค้นหาคำที่มี 8 ตัว ซึ่งตัวที่ 3 เป็น "j" และตัวที่ 6 เป็น "t" ส่วนตัวที่เหลือเป็นอะไรก็ได้**
ดังนั้นคำที่แสดงออกมาจะเป็นคำที่มี 8 ตัว ซึ่งตัวที่ 3 เป็น "j" และตัวที่ 6 เป็น "t" ส่วนตัวที่เหลือเป็นอะไรก็ได้เช่น dejected, adjusted

ในทางกลับกันคำที่ไม่ตรงกับรูปแบบนี้ก็คือคำที่มีตัวอักษรไม่เท่ากับ 8 ตัวหรือตัวที่ 3 ไม่เป็น "j" และตัวที่ 6 ไม่เป็น "t"

- reject (ความยาวแค่ 6 ตัว)
- projected (ความยาวเกิน 8 ตัว)
- objected (ตำแหน่งที่ 3 ไม่ใช่ j)

In [278…
```
[w for w in tokens if re.search('^spo.ts?$', w)][:10]
```

Out[278…
```
['sport',
 'sports',
 'sport',
 'sport',
 'sports',
 'sport',
 'sports',
 'sports',
 'sports',
 'sports']
```

---

3. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **^spo.ts?$ หมายถึง ค้นหาคำที่ขึ้นต้นด้วย "spo" ตามด้วยตัวอักษรใด ๆ หนึ่งตัว และตามด้วย "t" จากนั้นอาจจะมี "s" 1 ตัวหรือไม่มีเลยก็ได้**
ดังนั้นคำที่ตรงกับรูปแบบนี้จะเป็นคำที่มีความยาว 5 หรือ 6 ตัวอักษร เช่น sport, spots
ในทางกลับกัน คำที่ไม่ตรงคือคำที่ไม่ได้ขึ้นต้นด้วย "spo" หรือไม่มี "t" เป็นตัวรองสุดท้าย เช่น

- spoon (ไม่มีตัว t)
- spore (ไม่มีตัว t)
- sporting (เกิน 6 ตัวอักษร)

In [279…
```
[w for w in tokens if re.search('^[b-f]', w)][:10]
```

Out[279…
```
['chamber',
 'by',
 'by',
 'books',
 'for',
 'f.',
 'driver',
 'foul-weather',
 'friend',
 'copyright']
```

---

4. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **^[b-f] หมายถึง ค้นหาคำที่ขึ้นต้นด้วยตัวอักษรใดก็ได้ระหว่าง b ถึง f (รวม b, c, d, e, f)**

ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำที่อักษรตัวแรกอยู่ในช่วง b–f เช่น banana, cat, dog, elephant, fish
ในทางกลับกันคำที่ไม่ตรงคือคำที่ไม่ได้ขึ้นต้นด้วย b–f เช่น

- apple (ขึ้นต้นด้วย a)
- grape (ขึ้นต้นด้วย g)
- zebra (ขึ้นต้นด้วย z)

In [280…
```
[w for w in tokens if re.search('^[bdf][aue][rts]$', w)][:10]
```

Out[280…
```
['but', 'but', 'far', 'but', 'but', 'but', 'fat', 'but', 'fat', 'but']
```

---

5. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **^[bdf][aue][rts]$ หมายถึง ค้นหาคำที่มีความยาว 3 ตัวอักษร โดยตัวแรกต้องเป็น b, d หรือ f ตัวที่สองต้องเป็น a, u หรือ e และตัวที่สามต้องเป็น r, t หรือ s**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำที่มี 3 ตัวอักษร และตรงตามเงื่อนไขแต่ละตัว เช่น but, bet, fur, dus
ในทางกลับกันคำที่ไม่ตรงคือคำที่ไม่ใช่ 3 ตัวอักษร หรือไม่ตรงตามเงื่อนไขตัวอักษรแต่ละตำแหน่ง เช่น

- cat (ตัวแรกไม่ใช่ b, d, f)
- fun (ตัวสามไม่ใช่ r, t, s)
- dusted (ความยาวเกิน 3 ตัว)

In [281…
```
set([w for w in tokens if re.search('noo+$', w)])
```

Out[281…
```
{'nooo',
 'noooo',
 'nooooo',
 'noooooo',
 'nooooooo',
 'noooooooo',
 'nooooooooooo',
 'noooooooooooooooo'}
```

---

6. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **noo+$ หมายถึง ค้นหาคำที่ลงท้ายด้วยตัวอักษร "no" ตามด้วย "o" ซ้ำอย่างน้อยหนึ่งตัว**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำที่ลงท้ายด้วย "no" + "o" ซ้ำ เช่น nooo, nooooo, technoo
ในทางกลับกันคำที่ไม่ตรงคือคำที่ไม่ลงท้ายด้วย "no" + "o" ซ้ำ เช่น

- no (ไม่ลงท้ายด้วย "no" + "o"อย่างน้อยหนึ่งตัว)
- noon (ไม่ลงท้ายด้วย "no" + "o"อย่างน้อยหนึ่งตัว)

- banana (ไม่ลงท้ายด้วย "no" + "o"อย่างน้อยหนึ่งตัว)

In [282…
```
set([w for w in tokens if re.search('noo*$', w)])
```

Out[282…
```
{'albino',
 'dunno',
 'gemino',
 'inferno',
 'n-no',
 'no',
 'nooo',
 'noooo',
 'nooooo',
 'noooooo',
 'nooooooo',
 'noooooooo',
 'nooooooooooo',
 'noooooooooooooooo',
 'oppugno',
 'palomino',
 'patrono',
 'piano',
 'pi\xadano',
 'rhino',
 'xeno'}
```

---

7. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **noo*$ หมายถึง ค้นหาคำที่ลงท้ายด้วยตัวอักษร "no" แล้วตามด้วย "o" ตั้งแต่ 0 ตัวขึ้นไป**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำที่ลงท้ายด้วย "no" แล้วตามด้วย "o" ตั้งแต่ 0 ตัวขึ้นไป เช่น no, noo, nooo, techno
ในทางกลับกันคำที่ไม่ตรงคือคำที่ไม่ลงท้ายด้วย "no" แล้วตามด้วย "o" ตั้งแต่ 0 ตัวขึ้นไป เช่น

- na (ไม่ลงท้ายด้วย "no" แล้วตามด้วย "o" ตั้งแต่ 0 ตัวขึ้นไป)
- banana (ไม่ลงท้ายด้วย "no" แล้วตามด้วย "o" ตั้งแต่ 0 ตัวขึ้นไป)
- note (ไม่ลงท้ายด้วย "no" แล้วตามด้วย "o" ตั้งแต่ 0 ตัวขึ้นไป)

In [283…
```
[w for w in tokens if re.search('\.+$', w)][:10]
```

```
<>:1: SyntaxWarning: invalid escape sequence '\.'
<>:1: SyntaxWarning: invalid escape sequence '\.'
/tmp/ipython-input-707925584.py:1: SyntaxWarning: invalid escape sequence '\.'
  [w for w in tokens if re.search('\.+$', w)][:10]
```

Out[283…
```
['j.', 'k.', 'a.', 'p.', 'f.', '.', 'j.', 'k.', '.', 'bros.']
```

---

8. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **.+$ หมายถึง ค้นหาคำหรือสตริงที่ลงท้ายด้วยจุด (.) อย่างน้อย 1 ตัว**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำหรือสตริงที่ลงท้ายด้วย "." เช่น ., end., hello..., version.
ในทางกลับกันคำที่ไม่ตรงคือคำหรือสตริงที่ไม่ลงท้ายด้วย "." เช่น

- .hello (ไม่มีจุดลงท้าย)
- world (ไม่มีจุดลงท้าย)
- test (ไม่มีจุดลงท้าย)

In [284…
```
[w for w in tokens if re.search('^[A-Z]+\.$', w)]
```

```
<>:1: SyntaxWarning: invalid escape sequence '\.'
<>:1: SyntaxWarning: invalid escape sequence '\.'
/tmp/ipython-input-1844296225.py:1: SyntaxWarning: invalid escape sequence '\.'
  [w for w in tokens if re.search('^[A-Z]+\.$', w)]
```

Out[284…
```
[]
```

---

9. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **^[A-Z]+.$ หมายถึง ค้นหาคำที่ประกอบด้วยตัวอักษรพิมพ์ใหญ่ (A–Z) อย่างน้อยหนึ่งตัว และลงท้ายด้วยจุด (.)**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำที่เป็นตัวพิมพ์ใหญ่ทั้งหมดและลงท้ายด้วย "." เช่น HELLO., WORLD., TEST.
ในทางกลับกันคำที่ไม่ตรงคือคำหรือสตริงที่ไม่ใช่ตัวพิมพ์ใหญ่ทั้งหมด หรือไม่ลงท้ายด้วย "." เช่น

- Hello. (มีตัวพิมพ์เล็ก)
- WORLD (ไม่มีจุดลงท้าย)
- test. (เป็นตัวพิมพ์เล็ก)

In [285…
```
[w for w in tokens if re.search('^[0-9]{4}$', w)][:10]
```

Out[285…
```
['1999',
 '1999',
 '1920',
 '1999',
 '1999',
 '1875',
 '1492',
 '1289',
 '2007',
 '2007']
```

---

10. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **^[0-9]{4}$ หมายถึง ค้นหาคำที่ประกอบด้วยตัวเลข 0–9 จำนวน 4 ตัว**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือตัวเลข 4 ตัว เช่น 2025, 1234, 0001
ในทางกลับกันคำที่ไม่ตรงคือสตริงที่ไม่ใช่ตัวเลข 4 ตัว เช่น

- 123 (มีแค่ 3 ตัว)
- 12345 (มี 5 ตัว)
- ab12 (มีตัวอักษร)

In [286…
```
set([w for w in tokens if re.search('(ed|ing)$', w)][:10])
```

Out[286…
```
{'published',
 'publishing',
 'registered',
 'related',
 'reproduced',
 'reserved',
 'rowling',
 'stored'}
```

---

11. What kind of input that matches, and what do not? Also, give a brief explanation.

Ans: **(ed|ing)$ หมายถึง ค้นหาคำที่ลงท้ายด้วย "ed" หรือ "ing"**
ดังนั้นคำที่ตรงกับรูปแบบนี้คือคำที่ลงท้ายด้วย "ed" หรือ "ing" เช่น walked, jumping, played, running
ในทางกลับกันคำที่ไม่ตรงคือคำที่ไม่ลงท้ายด้วย "ed" หรือ "ing" เช่น

- walk (ไม่มี "ed" หรือ "ing")
- run (ไม่มี "ed" หรือ "ing")
- play (ไม่มี "ed" หรือ "ing")

---

3. Regular expression tokenizer

3.1 Basic regex tokenizer

In [287…
```
s = ("Good muffins cost $3.88\nin New York.  Please buy me\n"
     "two of them.\n\nThanks.")
s2 = ("Alas, it has not rained today. When, do you think, "
     "will it rain again?")
s3 = ("<p>Although this is <b>not</b> the case here, we must "
     "not relax our vigilance!</p>")
```

In [288…
```
s2
```

Out[288…
```
'Alas, it has not rained today. When, do you think, will it rain again?'
```

In [289…
```
nltk.regexp_tokenize(s2, r'[,\.\?!"]\s*', gaps=True)
```

Out[289…
```
['Alas',
 'it has not rained today',
 'When',
 'do you think',
 'will it rain again']
```

12.What happened after applying the above RegEx tokenizer to the input string?

Ans: **แยกคำโดยใช้เครื่องหมาย , . ? ! " เป็นตัวแบ่งคำ**

In [290…
```
s3
```

Out[290…
```
'<p>Although this is <b>not</b> the case here, we must not relax our vigilance!
</p>'
```

In [291…
```
nltk.regexp_tokenize(s3, r'</?.>', gaps=False)
```

Out[291…
```
['<p>', '<b>', '</b>', '</p>']
```

13.What happened after applying the above RegEx tokenizer to the input string?

Ans: **ค้นหาคำที่มีเครื่องหมาย < >ประกบ เช่น ,**

In [292…
```
nltk.regexp_tokenize(s3, r'</?.>', gaps=True)
```

Out[292…
```
['Although this is ',
 'not',
 ' the case here, we must not relax our vigilance!']
```

14.What happened after applying the above RegEx tokenizer to the input string?

Ans: **แยกคำโดยใช้เครื่องหมาย < > เป็นตัวแบ่งคำ****

In [293…
```
tokens[:10]
```

Out[293…
```
['harry', 'potter', 'and', 'the', 'chamber', 'of', 'secrets', 'by', 'j.', 'k.']
```

Harry Potter series

Cleaning text

In [294…
```
# 1. Lowercase all tokens
tokens = [w.lower() for w in tokens]
```

In [295…
```
# 2. Show punctuation and non-alphanumeric characters
from collections import Counter
punctuation_chars = []

# find
for w in tokens:
```

```
    punctuation_chars.extend(re.findall(r'[^a-zA-Z0-9]', w))

# counter
punctuation_count = Counter(punctuation_chars)

# result
for char, count in punctuation_count.items():
    print(f"{char} : {count}")
```

```
. : 65759
é : 30
á : 1
- : 5660
© : 23
, : 86310
/ : 10
: : 1474
[ : 6
— : 9711
] : 6
’ : 34979
“ : 37060
! : 7637
” : 36750
? : 10865
 : 9622
… : 7957
; : 3482
‘ : 511
( : 329
) : 331
* : 57
' : 3
& : 2
” : 1
ß : 1
¬ : 1
– : 9
ù : 1
```

In [296…

```
# 3. Remove punctuation and non-alphanumeric characters from each token
tokens = [re.sub(r'[^a-zA-Z0-9]', '', w) for w in tokens]
```

In [297…

```
# 4. Show space or empty tokens
empty_count = sum(w.strip() == '' for w in tokens)
print(f"Empty or space-only tokens count: {empty_count}")
```

```
Empty or space-only tokens count: 294308
```

In [298…

```
# 5. Remove space or empty tokens
tokens = [w for w in tokens if w.strip() != '']
```

In [299…

```
# 6. Show single characters b-z
single_chars = []

# find
for w in tokens:
    single_chars.extend(re.findall(r'^[a-z]{1}$', w))

# counter
```

```
single_count = Counter(single_chars)

# result
for char, count in sorted(single_count.items()):
    print(f"'{char}' : {count}")
```

```
'a' : 21038
'b' : 13
'c' : 61
'd' : 1984
'e' : 52
'f' : 5
'g' : 5
'h' : 9
'i' : 13557
'j' : 22
'k' : 19
'l' : 4
'm' : 1358
'n' : 25
'o' : 258
'p' : 15
'r' : 7
's' : 14750
't' : 7703
'u' : 1
'v' : 1
'w' : 6
'x' : 2
'y' : 20
'z' : 1
```

In [300…

```
# 7. Remove single characters b-z
tokens = [re.sub(r'\b[b-z]\b', '', w) for w in tokens]
```

In [301…

```
print("Clean tokens:", tokens[:10])
```

```
Clean tokens: ['harry', 'potter', 'and', 'the', 'chamber', 'of', 'secrets', 'by',
'', '']
```

EDA

จำนวนบทของหนังสือ

In [302…

```
cleaned_text = ' '.join(tokens)  # รวม tokens เป็น string

# ใช้ regex chapter [\d][\d]* → chapter ตามด้วยตัวเลข 1 หลักขึ้นไป
chapter_matches = re.findall(r'chapter [\d][\d]*', cleaned_text)

# นับจำนวน occurrence
chapter_count = len(chapter_matches)

print(f"The phrase 'chapter xx' appears {chapter_count} times in the text.")
```

```
The phrase 'chapter xx' appears 199 times in the text.
```

จำนวนเล่มของหนังสือ

In [303…
```
# ใช้ regex หา "chapter 1"
chapter_1_matches = re.findall(r'\bchapter 1\b', cleaned_text)

# นับจำนวน occurrence
chapter_1_count = len(chapter_1_matches)

print(f"The phrase 'chapter 1' appears {chapter_1_count} times in the text.")
```

```
The phrase 'chapter 1' appears 7 times in the text.
```

พิจารณา book_2 เนื่องจากจำนวนบทไม่ตรง เพราะ chapter 18 อ้างอิง chapter 22 ในเนื้อหา

หมายเหตุ !

ตอนแรกคิดว่าจะรวม chapter 18 กับ chapter 19 แล้วลบ chapter 19 ออกเหมือนใน comment

In [304…
```
book_index = 2  # enumerate start from 0
book_2 = books[book_index]

# split chapters
chapters_in_book2 = re.split(r'(?i)(?=chapter \d+\b)', book_2)
chapters_in_book2 = [ch.strip() for ch in chapters_in_book2 if ch.strip()]

# merge chapter 18 and 19, then drop chapter 19
# if len(chapters_in_book2) >= 19:
#     chapters_in_book2[17] = chapters_in_book2[17] + "\n\n" + chapters_in_book2
#     del chapters_in_book2[18]  # remove chapter 19

# results
for j, chapter in enumerate(chapters_in_book2, 1):
    print(f"chapter {j}: {chapter[:100]}...")  # preview first 100 char
```

chapter 1: chapter 1 the dark lord ascending the two men appeared out of nowhere a few yards apart in the narro...
chapter 2: chapter 2 in memoriam harry was bleeding clutching his right hand in h is left and swearing under his...
chapter 3: chapter 3 the dursleys departing the sound of the front door slamming echoed up the stairs and a voi...
chapter 4: chapter 4 the seven potters harry ran back upstairs to his bedroom arr iving at the window just in ti...
chapter 5: chapter 5 fallen warrior hagrid harry struggled to raise himself out o f the debris of metal and leat...
chapter 6: chapter 6 the ghoul in pajamas the shock of losing madeye hung over th e house in the days that follo...
chapter 7: chapter 7 the will of albus dumbledore he was walking along a mountain road in the cool blue light o...
chapter 8: chapter 8 the wedding three clock on the following afternoon found ha rry ron fred and george standi...
chapter 9: chapter 9 a place to hide everything seemed fuzzy slow harry and hermi one jumped to their feet and d...
chapter 10: chapter 10 kreacher tale harry woke early next morning wrapped in a sleeping bag on the drawing roo...
chapter 11: chapter 11 the bribe if kreacher could escape a lake full of inferi h arry was confident that the cap...
chapter 12: chapter 12 magic is might as august wore on the square of unkempt gra ss in the middle of grimmauld p...
chapter 13: chapter 13 the muggleborn registration commission ah mafalda said umb ridge looking at hermione trave...
chapter 14: chapter 14 the thief harry opened his eyes and was dazzled by gold an d green he had no idea what had...
chapter 15: chapter 15 the goblin revenge early next morning before the other tw o were awake harry left the ten...
chapter 16: chapter 16 godric hollow when harry woke the following day it was se veral seconds before he remembe...
chapter 17: chapter 17 bathilda secret harry stop what wrong they had only just reached the grave of the unkno...
chapter 18: chapter 18 the life and lies of albus dumbledore the sun was coming u p the pure colorless vastness o...
chapter 19: chapter 22 what caused this abrupt rupture had dumbledore come to his senses had he told grindelwald...
chapter 20: chapter 19 the silver doe it was snowing by the time hermione took ov er the watch at midnight harry ...
chapter 21: chapter 20 xenophilius lovegood harry had not expected hermione ange r to abate overnight and was th...
chapter 22: chapter 21 the tale of the three brothers harry turned to look at ron and hermione neither of them s...
chapter 23: chapter 22 the deathly hallows harry fell panting onto grass and scra mbled up at once they seemed to...
chapter 24: chapter 23 malfoy manor harry looked around at the other two now mere outlines in the darkness he sa...
chapter 25: chapter 24 the wandmaker it was like sinking into an old nightmare fo r an instant harry knelt again ...
chapter 26: chapter 25 shell cottage bill and fleur cottage stood alone on a cli ff overlooking the sea its wall...
chapter 27: chapter 26 gringotts their plans were made their preparations complet e in the smallest bedroom a sin...
chapter 28: chapter 27 the final hiding place there was no means of steering the dragon could not see where it w...
chapter 29: chapter 28 the missing mirror harry feet touched road he saw the ach ingly familiar hogsmeade high s...
chapter 30: chapter 29 the lost diadem neville what the how but neville had spott ed ron and hermione and with ye...

```
chapter 31: chapter 30 the sacking of severus snape the moment her finger touched
the mark harry  scar burned sa...
chapter 32: chapter 31 the battle of hogwarts the enchanted ceiling of the great
hall was dark and scattered wit...
chapter 33: chapter 32 the elder wand the world had ended so why had the battle n
ot ceased the castle fallen sil...
chapter 34: chapter 33 the prince  tale harry remained kneeling at snape  side si
mply staring down at him until ...
chapter 35: chapter 34 the forest again finally the truth lying with his face pre
ssed into the dusty carpet of t...
chapter 36: chapter 35 king  cross he lay facedown listening to the silence he wa
s perfectly alone nobody was wa...
chapter 37: chapter 36 the flaw in the plan he was lying facedown on the ground a
gain the smell of the forest fi...
```

แต่คิดวิธีที่ดีกว่าได้คือการแบ่ง chapter แบบไล่ลำดับ :)

Comparable length each book and all book

In [305…

```python
import re

# เก็บข้อมูล chapter ที่สั้นที่สุดและยาวที่สุดของทุกเล่ม
shortest_chapter_overall = {"book": None, "chapter": None, "words": float('inf')}
longest_chapter_overall = {"book": None, "chapter": None, "words": 0}

for i, book in enumerate(books, 0):
    # สร้าง pattern ไล่ลำดับ
    chapters = []
    current_pos = 0
    chapter_num = 1

    while True:
        next_chapter_pattern = f'chapter {chapter_num}'
        match = re.search(next_chapter_pattern, book[current_pos:], re.IGNORECAS
        if not match:
            remaining_text = book[current_pos:].strip()
            if remaining_text:
                chapters.append(remaining_text)
            break
        start = current_pos + match.start()

        if chapter_num > 1:
            chapters.append(book[current_pos:start].strip())

        current_pos = start
        chapter_num += 1

    if i > 0:
        print(f"book_{i} has {len(chapters)} chapters")

        chapter_lengths = [len(ch.split()) for ch in chapters]

        for j, length in enumerate(chapter_lengths, 1):
            print(f"  chapter {j} has {length} words")
            # update the shortest chapter
            if length < shortest_chapter_overall["words"]:
                shortest_chapter_overall = {"book": i, "chapter": j, "words": le
            # update the longest chapter
```

```
            if length > longest_chapter_overall["words"]:
                longest_chapter_overall = {"book": i, "chapter": j, "words": len

        # Check chapter is the shortest and which is the longest each books
        min_length = min(chapter_lengths)
        max_length = max(chapter_lengths)
        min_index = chapter_lengths.index(min_length) + 1
        max_index = chapter_lengths.index(max_length) + 1

        if min_length != max_length:
            print(f"  -> Shortest chapter in this book: chapter {min_index} ({mi
            print(f"  -> Longest chapter in this book: chapter {max_index} ({max

        print()

# Check chapter is the shortest and which is the longest all books
print(f"Overall shortest chapter: book_{shortest_chapter_overall['book']}, "
      f"chapter {shortest_chapter_overall['chapter']} "
      f"({shortest_chapter_overall['words']} words)")

print(f"Overall longest chapter: book_{longest_chapter_overall['book']}, "
      f"chapter {longest_chapter_overall['chapter']} "
      f"({longest_chapter_overall['words']} words)")
```

```
book_1 has 18 chapters
  chapter 1 has 2560 words
  chapter 2 has 2843 words
  chapter 3 has 4487 words
  chapter 4 has 5733 words
  chapter 5 has 5395 words
  chapter 6 has 4521 words
  chapter 7 has 4529 words
  chapter 8 has 4360 words
  chapter 9 has 5085 words
  chapter 10 has 5282 words
  chapter 11 has 5937 words
  chapter 12 has 5392 words
  chapter 13 has 5441 words
  chapter 14 has 3911 words
  chapter 15 has 4634 words
  chapter 16 has 5626 words
  chapter 17 has 5341 words
  chapter 18 has 4131 words
  -> Shortest chapter in this book: chapter 1 (2560 words)
  -> Longest chapter in this book: chapter 11 (5937 words)

book_2 has 36 chapters
  chapter 1 has 3083 words
  chapter 2 has 4038 words
  chapter 3 has 3291 words
  chapter 4 has 5349 words
  chapter 5 has 5813 words
  chapter 6 has 6535 words
  chapter 7 has 6498 words
  chapter 8 has 6134 words
  chapter 9 has 4104 words
  chapter 10 has 6594 words
  chapter 11 has 5517 words
  chapter 12 has 6053 words
  chapter 13 has 5810 words
  chapter 14 has 4182 words
  chapter 15 has 6935 words
  chapter 16 has 5184 words
  chapter 17 has 5391 words
  chapter 18 has 3445 words
  chapter 19 has 6618 words
  chapter 20 has 4423 words
  chapter 21 has 4802 words
  chapter 22 has 5797 words
  chapter 23 has 7786 words
  chapter 24 has 6651 words
  chapter 25 has 4312 words
  chapter 26 has 6809 words
  chapter 27 has 2628 words
  chapter 28 has 4473 words
  chapter 29 has 4515 words
  chapter 30 has 4844 words
  chapter 31 has 7937 words
  chapter 32 has 5485 words
  chapter 33 has 8113 words
  chapter 34 has 3864 words
  chapter 35 has 4862 words
  chapter 36 has 8982 words
  -> Shortest chapter in this book: chapter 27 (2628 words)
```

```
  -> Longest chapter in this book: chapter 36 (8982 words)

book_3 has 37 chapters
  chapter 1 has 4121 words
  chapter 2 has 2840 words
  chapter 3 has 3179 words
  chapter 4 has 2990 words
  chapter 5 has 3774 words
  chapter 6 has 2386 words
  chapter 7 has 5259 words
  chapter 8 has 5804 words
  chapter 9 has 7175 words
  chapter 10 has 3165 words
  chapter 11 has 3235 words
  chapter 12 has 5446 words
  chapter 13 has 3896 words
  chapter 14 has 4840 words
  chapter 15 has 5247 words
  chapter 16 has 6078 words
  chapter 17 has 4066 words
  chapter 18 has 6607 words
  chapter 19 has 6342 words
  chapter 20 has 7081 words
  chapter 21 has 5556 words
  chapter 22 has 4450 words
  chapter 23 has 8142 words
  chapter 24 has 6181 words
  chapter 25 has 5501 words
  chapter 26 has 8008 words
  chapter 27 has 6989 words
  chapter 28 has 7303 words
  chapter 29 has 4334 words
  chapter 30 has 6253 words
  chapter 31 has 7910 words
  chapter 32 has 2007 words
  chapter 33 has 3931 words
  chapter 34 has 2892 words
  chapter 35 has 5819 words
  chapter 36 has 6133 words
  chapter 37 has 5100 words
  -> Shortest chapter in this book: chapter 32 (2007 words)
  -> Longest chapter in this book: chapter 23 (8142 words)

book_4 has 30 chapters
  chapter 1 has 5038 words
  chapter 2 has 4810 words
  chapter 3 has 4716 words
  chapter 4 has 5892 words
  chapter 5 has 5883 words
  chapter 6 has 5875 words
  chapter 7 has 6464 words
  chapter 8 has 4275 words
  chapter 9 has 5841 words
  chapter 10 has 5944 words
  chapter 11 has 5095 words
  chapter 12 has 5360 words
  chapter 13 has 5351 words
  chapter 14 has 6467 words
  chapter 15 has 5763 words
  chapter 16 has 6059 words
```

```
  chapter 17 has 6311 words
  chapter 18 has 6487 words
  chapter 19 has 6174 words
  chapter 20 has 6503 words
  chapter 21 has 5703 words
  chapter 22 has 5739 words
  chapter 23 has 5701 words
  chapter 24 has 5903 words
  chapter 25 has 4970 words
  chapter 26 has 6261 words
  chapter 27 has 4448 words
  chapter 28 has 3583 words
  chapter 29 has 5434 words
  chapter 30 has 5587 words
  -> Shortest chapter in this book: chapter 28 (3583 words)
  -> Longest chapter in this book: chapter 20 (6503 words)

book_5 has 38 chapters
  chapter 1 has 5750 words
  chapter 2 has 6019 words
  chapter 3 has 5223 words
  chapter 4 has 5670 words
  chapter 5 has 5345 words
  chapter 6 has 6748 words
  chapter 7 has 4564 words
  chapter 8 has 4172 words
  chapter 9 has 7993 words
  chapter 10 has 5780 words
  chapter 11 has 5950 words
  chapter 12 has 8294 words
  chapter 13 has 8829 words
  chapter 14 has 8039 words
  chapter 15 has 6719 words
  chapter 16 has 5877 words
  chapter 17 has 6570 words
  chapter 18 has 6854 words
  chapter 19 has 6666 words
  chapter 20 has 5987 words
  chapter 21 has 7291 words
  chapter 22 has 7978 words
  chapter 23 has 7368 words
  chapter 24 has 8007 words
  chapter 25 has 8054 words
  chapter 26 has 8499 words
  chapter 27 has 7427 words
  chapter 28 has 7973 words
  chapter 29 has 7401 words
  chapter 30 has 8219 words
  chapter 31 has 7943 words
  chapter 32 has 6126 words
  chapter 33 has 3840 words
  chapter 34 has 5220 words
  chapter 35 has 7728 words
  chapter 36 has 3786 words
  chapter 37 has 7902 words
  chapter 38 has 7759 words
  -> Shortest chapter in this book: chapter 36 (3786 words)
  -> Longest chapter in this book: chapter 13 (8829 words)

book_6 has 22 chapters
```

```
  chapter 1 has 3683 words
  chapter 2 has 3840 words
  chapter 3 has 4357 words
  chapter 4 has 5077 words
  chapter 5 has 6533 words
  chapter 6 has 6672 words
  chapter 7 has 4213 words
  chapter 8 has 5008 words
  chapter 9 has 5243 words
  chapter 10 has 7009 words
  chapter 11 has 5529 words
  chapter 12 has 4750 words
  chapter 13 has 4226 words
  chapter 14 has 5497 words
  chapter 15 has 5484 words
  chapter 16 has 4343 words
  chapter 17 has 4164 words
  chapter 18 has 2237 words
  chapter 19 has 5056 words
  chapter 20 has 1924 words
  chapter 21 has 7210 words
  chapter 22 has 4892 words
  -> Shortest chapter in this book: chapter 20 (1924 words)
  -> Longest chapter in this book: chapter 21 (7210 words)

book_7 has 17 chapters
  chapter 1 has 4569 words
  chapter 2 has 3436 words
  chapter 3 has 3833 words
  chapter 4 has 3640 words
  chapter 5 has 6511 words
  chapter 6 has 6235 words
  chapter 7 has 4443 words
  chapter 8 has 3042 words
  chapter 9 has 4896 words
  chapter 10 has 4275 words
  chapter 11 has 3311 words
  chapter 12 has 5454 words
  chapter 13 has 3171 words
  chapter 14 has 3444 words
  chapter 15 has 5096 words
  chapter 16 has 6371 words
  chapter 17 has 5355 words
  -> Shortest chapter in this book: chapter 8 (3042 words)
  -> Longest chapter in this book: chapter 5 (6511 words)

Overall shortest chapter: book_6, chapter 20 (1924 words)
Overall longest chapter: book_2, chapter 36 (8982 words)
```

Find the characters

In [306…

```
characters = ['harry', 'ron', 'hermione', 'dobby', 'malfoy', 'voldemort|you know

for name in characters:
    matches = re.findall(name, cleaned_text)
    print(f"{name.upper()} occurences is {len(matches)} times.")
```

```
HARRY occurences is 18254 times.
RON occurences is 8200 times.
HERMIONE occurences is 5365 times.
DOBBY occurences is 470 times.
MALFOY occurences is 1363 times.
VOLDEMORT|YOU KNOW WHO occurences is 1266 times.
SNAPE occurences is 1829 times.
DUMBLEDORE occurences is 3375 times.
HAGRID occurences is 2044 times.
SIRIUS occurences is 1137 times.
```

Punchaya Chancharoen
65070507236