DATA 603: ASSIGNMENT 2

Mrunal Vinay Jadhav

Birth Date: 25th March 1998

PART 1

* **File 1 pages 25 to 35**
* **Book Name: Harry Potter and the Prisoner of Azkaban**
* **MapReduce to count the number of occurrences of each word in the first text file (file1.txt).**

Code:

```
#accessing the file 1 text and reading it
file_path = 'file1.txt'
with open(file_path, 'r', encoding='ISO-8859-1') as file:
    text = file.read()

#importing necessary files
from collections import defaultdict
import re

#removing the punctuations and reading the words
punctuation = re.findall(r'\b\w+\b', text.lower())

#mapping the words to the occurrances
def count_occurrences(punctuation):
  return [(word,1)for word in punctuation]

#shuffling the words
def word_shuffel(map):
  shufflingofwords = defaultdict(list)
  for word, count in map:
    shufflingofwords[word].append(count)
  return shufflingofwords
```

```
#reduction of the number of words appeared
def word_reduction(shuffle):
  return {word: sum(counts) for word, counts in shuffle.items()}


#getting the total count of sorted word occurrances
def map_reduce_count(text):
  mapping = count_occurrences(text)
  shuffling = word_shuffel(mapping)
  reducing = word_reduction(shuffling)
  return reducing


final_word_count = map_reduce_count(punctuation)


#printing the results
print('The total word count in file is:')
print(final_word_count)
```

Output:

The total word count in file is:

{'ripper': 3, 'began': 2, 'to': 49, 'growl': 1, 'again': 6, 'as': 15, 'harry': 55, 'sat': 3, 'down': 7, 'this': 9, 'directed': 1, 'aunt': 34, 'marge': 32, 's': 31, 'attention': 1, 'for': 13, 'the': 115, 'first': 2, 'time': 4, 'so': 5, 'she': 22, 'barked': 2, 'still': 6, 'here': 4, 'are': 4, 'you': 28, 'yes': 4, 'said': 18, 'p': 11, 'a': 65, 'g': 11, 'e': 11, 'i': 30, '25': 1, 'potter': 12, 'and': 70, 'prisoner': 12, 'of': 64, 'azkaban': 11, 'j': 11, 'k': 11, 'rowling': 11, 'don': 5, 't': 22, 'say': 3, 'in': 37, 'that': 18, 'ungrateful': 2, 'tone': 2, 'growled': 1, 'it': 32, 'damn': 1, 'good': 3, 'vernon': 23, 'petunia': 14, 'keep': 2, 'wouldn': 1, 'have': 9, 'done': 2, 'myself': 1, 'd': 11, 'gone': 3, 'straight': 1, 'an': 9, 'orphanage': 2, 'if': 9, 'been': 4, 'dumped': 1, 'on': 20, 'my': 1, 'doorstep': 1, 'was': 40, 'bursting': 2, 'he': 73, 'rather': 2, 'live': 1, 'than': 3, 'with': 22, 'dursleys': 2, 'but': 18, 'thought': 4, 'hogsmeade': 3, 'form': 3, 'stopped': 2, 'him': 18, 'forced': 1, 'his': 55, 'face': 7, 'into': 9, 'painful': 1, 'smile': 1, 'smirk': 1, 'at': 25, 'me': 5, 'boomed': 1, 'can': 4, 'see': 4, 'haven': 1, 'improved': 1, 'since': 2, 'last': 5, 'saw': 2, 'hoped': 1, 'school': 2, 'would': 7, 'knock': 1, 'some': 1, 'manners': 1, 'took': 2, 'large': 2, 'gulp': 1, 'tea': 1, 'wiped': 1, 'her': 31, 'mustache': 1, 'where': 2, 'is': 2, 'send': 1, 'st': 2, 'brutus': 2, 'uncle': 18, 'promptly': 1, 'rate': 1, 'institution': 1, 'hopeless': 1, 'cases': 2, 'do': 7, 'they': 8, 'use': 2, 'cane': 1, 'boy': 7, 'across': 1, 'table': 4, 'er': 1, 'nodded': 1, 'curtly': 1, 'behind': 4, 'back': 6, 'then': 6, 'feeling': 1, 'might': 2, 'well': 2, 'thing': 4, 'properly': 1, 'added': 1, 'all': 9, 'excellent': 2, 'won': 1, 'namby': 1, 'pamby': 1, 'wishy': 1, 'washy': 1, 'nonsense': 1, 'about': 8, 'not': 4, 'hitting': 2, 'people': 1, 'who': 5, 'deserve': 1, 'thrashing': 1, 'what': 8, 'needed': 1, 'ninety': 1, 'nine': 1, 'out': 19, 'hundred': 1, 'beaten': 1, 'often': 1, 'oh': 1, 'yeah': 1, 'loads': 1, 'times': 2, '26': 1, 'narrowed': 1, 'eyes': 7, 'like': 8, 'your': 7, 'speak': 1, 'beatings': 1, 'casual': 1, 'way': 7, 'clearly': 2, 'aren': 1, 'hard': 3, 'enough': 2, 'write': 1, 'were': 4, 'make': 3, 'clear': 1, 'approve': 1, 'extreme': 1, 'force': 1, 'case': 2, 'perhaps': 1, 'worried': 1, 'forget': 1, 'their': 3, 'bargain': 1, 'any': 4, 'changed': 1, 'subject': 1, 'abruptly': 1, 'heard': 3, 'news': 1, 'morning': 1, 'escaped': 1, 'eh': 1, 'started': 4, 'herself': 1, 'home': 1, 'caught': 1, 'himself': 8, 'thinking': 1, 'almost': 5, 'longingly': 1, 'life': 4, 'number': 2, 'four': 1, 'without': 2, 'usually': 1, 'encouraged': 1, 'stay': 2, 'which': 3, 'only': 4, 'too': 4, 'happy': 1, 'other': 3, 'hand': 7, 'wanted': 2, 'under': 4, 'eye': 1, 'could': 4, 'boom': 1,

'suggestions': 1, 'improvement': 1, 'delighted': 1, 'comparing': 1, 'dudley': 5, 'huge': 3, 'pleasure': 1, 'buying': 1, 'expensive': 1, 'presents': 2, 'while': 1, 'glaring': 1, 'though': 3, 'daring': 1, 'ask': 1, 'why': 2, 'hadn': 1, 'got': 6, 'present': 1, 'also': 1, 'kept': 1, 'throwing': 1, 'dark': 4, 'hints': 1, 'made': 4, 'such': 1, 'unsatisfactory': 1, 'person': 1, 'mustn': 1, 'blame': 1, 'yourself': 2, 'turned': 1, 'over': 6, 'lunch': 1, 'third': 1, 'day': 3, 'there': 6, 'something': 7, 'rotten': 1, 'inside': 1, 'nothing': 3, 'anyone': 2, 'tried': 2, 'concentrate': 1, 'food': 1, 'hands': 3, 'shook': 1, 'starting': 1, 'burn': 1, 'anger': 3, 'remember': 3, 'told': 2, 'think': 2, 'anything': 1, 'rise': 2, '27': 1, 'reached': 2, 'glass': 5, 'wine': 3, 'one': 8, 'basic': 1, 'rules': 1, 'breeding': 1, 'dogs': 3, 'wrong': 2, 'bitch': 1, '11': 1, 'be': 10, 'pup': 1, 'moment': 3, 'wineglass': 1, 'holding': 2, 'exploded': 1, 'shards': 1, 'flew': 3, 'every': 1, 'direction': 1, 'sputtered': 1, 'blinked': 1, 'great': 3, 'ruddy': 1, 'dripping': 1, 'squealed': 1, 'right': 3, 'worry': 1, 'grunted': 1, 'mopping': 1, 'napkin': 1, 'must': 1, 'squeezed': 1, 'did': 2, 'saine': 1, 'colonel': 2, 'fubster': 2, 'no': 5, 'need': 1, 'fuss': 1, 'very': 6, 'firm': 1, 'grip': 1, 'both': 2, 'looking': 4, 'suspiciously': 1, 'decided': 1, 'better': 1, 'skip': 1, 'dessert': 1, 'escape': 1, 'from': 11, 'soon': 1, 'outside': 2, 'hall': 1, 'leaned': 1, 'against': 3, 'wall': 3, 'breathing': 2, 'deeply': 1, 'had': 23, 'long': 3, 'lost': 1, 'control': 1, 'explode': 1, 'couldn': 4, 'afford': 1, 'let': 1, 'happen': 2, 'wasn': 2, 'stake': 1, 'ifhe': 1, 'carried': 1, 'trouble': 1, 'ministry': 3, 'magic': 6, 'underage': 2, 'wizard': 3, 'forbidden': 1, 'by': 3, 'law': 1, 'record': 1, 'exactly': 1, 'clean': 1, 'either': 2, 'summer': 1, 'gotten': 1, 'official': 1, 'warning': 1, 'stated': 1, 'quite': 6, 'wind': 1, 'more': 8, 'privet': 1, 'drive': 1, 'expulsion': 1, 'hogwarts': 2, '28': 1, 'leaving': 1, 'hurried': 1, 'upstairs': 2, 'through': 3, 'next': 4, 'three': 1, 'days': 1, 'forcing': 1, 'handbook': 2, 'broomcare': 1, 'whenever': 1, 'worked': 1, 'seemed': 3, 'give': 1, 'glazed': 1, 'look': 3, 'because': 1, 'voicing': 1, 'opinion': 1, 'mentally': 1, 'subnormal': 1, 'final': 1, 'evening': 2, 'arrived': 1, 'cooked': 1, 'fancy': 1, 'dinner': 1, 'uncorked': 1, 'several': 3, 'bottles': 1, 'soup': 1, 'salmon': 1, 'single': 1, 'mention': 1, 'faults': 1, 'during': 1, 'lemon': 1, 'meringue': 1, 'pie': 3, 'bored': 1, 'them': 3, 'talk': 1, 'grunnings': 1, 'drill': 1, 'making': 2, 'company': 1, 'coffee': 2, 'brought': 1, 'bottle': 3, 'brandy': 6, 'tempt': 1, 'already': 2, 'lot': 1, 'red': 2, 'just': 6, 'small': 1, 'chuckled': 1, 'bit': 3, '1nore': 1, 'ticket': 1, 'eating': 1, 'fourth': 1, 'slice': 1, 'sipping': 1, 'little': 6, 'finger': 1, 'sticking': 1, 'really': 1, 'disappear': 1, 'bedromn': 1, 'met': 1, 'angry': 2, 'knew': 1, 'sit': 2, 'aah': 1, 'smacking': 1, 'lips': 1, 'putting': 1, 'empty': 2, 'nosh': 1, 'normally': 1, 'fry': 1, 'up': 11, 'twelve': 2, 'after': 2, 'burped': 1, 'richly': 1, 'patted': 2, 'tweed': 3, 'stomach': 3, '29': 1, 'pardon': 1, 'healthy': 1, 'sized': 2, 'went': 2, 'winking': 1, 'll': 2, 'proper': 1, 'man': 1, 'dudders': 1, 'father': 2, 'spot': 1, 'now': 5, 'jerked': 1, 'head': 2, 'felt': 2, 'clench': 1, 'quickly': 1, 'mean': 1, 'runty': 1, 'get': 3, 'drown': 1, 'year': 1, 'ratty': 1, 'weak': 1, 'underbred': 1, 'trying': 2, 'page': 1, 'book': 1, 'charm': 1, 'cure': 1, 'reluctant': 1, 'reversers': 1, 'comes': 1, 'blood': 2, 'saying': 2, 'bad': 2, 'will': 1, 'm': 2, 'family': 1, 'bony': 1, 'shovellike': 1, 'sister': 1, 'egg': 1, 'turn': 1, 'best': 1, 'families': 1, 'ran': 1, 'off': 3, 'wastrel': 1, 'result': 1, 'front': 2, 'us': 1, 'staring': 1, 'plate': 1, 'funny': 2, 'ringing': 1, 'ears': 1, 'grasp': 1, 'broom': 1, 'firmly': 1, 'tail': 1, 'came': 2, 'voice': 1, 'boring': 1, 'drills': 1, 'loudly': 1, 'seizing': 1, 'splashing': 1, 'tablecloth': 1, 'never': 4, 'extremely': 1, 'tense': 1, 'even': 2, 'looked': 5, 'fro1n': 1, 'gape': 1, 'parents': 3, '30': 1, 'didn': 4, 'work': 1, 'half': 1, 'glance': 1, 'unemployed': 1, 'expected': 1, 'taking': 1, 'swig': 1, 'wiping': 1, 'chin': 1, 'sleeve': 1, 'account': 1, 'lazy': 1, 'scrounger': 1, 'suddenly': 4, 'quiet': 2, 'shaking': 1, 'yelled': 2, 'white': 1, 'emptied': 1, 'snarled': 1, 'go': 6, 'bed': 2, 'hiccuped': 1, 'tiny': 2, 'bloodshot': 1, 'fixed': 1, 'proud': 1, 'themselves': 1, 'killed': 1, 'car': 3, 'crash': 3, 'drunk': 1, 'expect': 1, 'die': 1, 'found': 2, 'feet': 3, 'died': 1, 'nasty': 1, 'liar': 1, 'left': 2, 'burden': 1, 'decent': 1, 'hardworking': 1, 'relatives': 1, 'screamed': 1, 'swelling': 3, 'fury': 1, 'insolent': 1, 'speaking': 1, 'words': 1, 'failed': 1, 'inexpressible': 1, 'stop': 2, 'expand': 1, 'bulged': 1, 'mouth': 1, 'stretched': 1, 'tightly': 1, 'speech': 1, 'second': 2, 'buttons': 1, 'burst': 3, 'jacket': 1, 'pinged': 1, 'walls': 2, 'inflating': 1, 'monstrous': 1, 'balloon': 1, 'free': 1, '31': 1, 'waistband': 1, 'each': 1, 'fingers': 1, 'blowing': 1, 'salami': 1, 'together': 1, 'whole': 1, 'body': 1, 'chair': 1, 'toward': 1, 'ceiling': 1, 'entirely': 1, 'round': 1, 'vast': 1, 'buoy': 1, 'piggy': 1, 'stuck': 1, 'weirdly': 1, 'drifted': 1, 'air': 1, 'apoplectic': 1, 'popping': 1, 'noises': 1, 'skidding': 1, 'room': 3, 'barking': 1, 'madly': 1, 'nooooooo': 1, 'seized': 2, 'pull': 1, 'lifted': 1, 'floor': 1, 'later': 1, 'leapt': 1, 'forward': 1, 'sank': 2, 'teeth': 1, 'leg': 2, 'tore': 1, 'dining': 2, 'before': 3, 'heading': 1, 'cupboard': 2, 'stairs': 1, 'door': 4, 'magically': 1, 'open': 2, 'seconds': 1, 'heaved': 1, 'trunk': 11, 'sprinted': 1, 'threw': 2, 'wrenching': 1, 'loose': 1, 'floorboard': 1, 'grabbed': 1, 'pillowcase': 1, 'full': 1, 'books': 1, 'birthday': 1, 'wriggled': 1, 'hedwig': 3, 'cage': 2, 'dashed': 2, 'downstairs': 1, 'trouser': 1, 'bloody': 1, 'tatters': 1, 'come': 3, 'bellowed': 1, 'put': 1, 'reckless': 1, 'rage': 1, 'kicked': 1, 'pulled': 1, 'wand': 5, 'pointed': 1, 'deserved': 2, 'fast': 2, 'away': 2, '32': 1, 'fumbled': 1, 'hi1n': 1, 'latch': 1, 'going': 2, 've': 1, 'street': 3, 'heaving': 1, 'heavy': 1, 'arm': 2, '33': 1, 'knight': 1, 'bus': 1, 'streets': 1, 'collapsed': 1, 'onto': 1, 'low': 1, 'magnolia': 2, 'crescent': 2, 'panting': 1, 'effort': 1, 'dragging': 1, 'surging': 1,

'listening': 1, 'frantic': 1, 'thumping': 2, 'heart': 3, 'ten': 1, 'minutes': 1, 'alone': 2, 'new': 1, 'emotion': 1, 'overtook': 1, 'panic': 1, 'whichever': 1, 'worse': 1, 'fix': 1, 'stranded': 1, 'muggle': 3, 'world': 2, 'absolutely': 1, 'nowhere': 1, 'worst': 1, 'serious': 1, 'meant': 1, 'certainly': 1, 'expelled': 2, 'broken': 1, 'decree': 1, 'restriction': 1, 'wizardry': 1, 'badly': 1, 'surprised': 1, 'representatives': 1, 'weren': 1, 'swooping': 1, 'shivered': 1, 'arrested': 1, 'or': 5, 'simply': 1, 'outlawed': 1, 'wizarding': 2, 'ron': 2, '34': 1, 'hermione': 2, 'lower': 1, 'sure': 1, 'criminal': 1, 'want': 1, 'help': 1, 'abroad': 1, 'means': 1, 'contacting': 1, 'money': 3, 'gold': 1, 'bag': 1, 'bottom': 1, 'rest': 2, 'fortune': 1, 'stored': 1, 'vault': 2, 'gringotts': 1, 'bank': 1, 'london': 3, 'able': 1, 'drag': 1, 'unless': 1, 'clutching': 1, 'painfully': 1, 'hurt': 1, 'invisibility': 2, 'cloak': 3, 'inherited': 1, 'bewitched': 1, 'feather': 1, 'light': 2, 'tied': 1, 'broomstick': 2, 'covered': 1, 'begin': 1, 'outcast': 1, 'horrible': 1, 'prospect': 1, 'forever': 1, 'find': 1, 'explain': 1, 'police': 1, 'dead': 1, 'night': 1, 'trunkful': 1, 'spellbooks': 1, 'opened': 1, 'pushed': 1, 'contents': 1, 'aside': 1, 'straightened': 1, 'around': 1, 'once': 2, 'prickling': 1, 'neck': 1, 'feel': 1, 'being': 1, 'watched': 1, 'appeared': 2, 'deserted': 1, 'lights': 1, 'shone': 1, 'square': 1, 'houses': 1, 'bent': 1, 'immediately': 1, 'stood': 1, 'clenched': 1, '35': 1, 'sensed': 1, 'someone': 1, 'standing': 1, 'narrow': 1, 'gap': 1, 'between': 2, 'garage': 2, 'fence': 1, 'squinted': 1, 'black': 1, 'alleyway': 1, 'move': 1, 'know': 1, 'whether': 1, 'stray': 1, 'cat': 1, 'else': 1, 'lumas': 1, 'muttered': 1, 'end': 1, 'dazzling': 1, 'held': 1, 'high': 1, 'pebble': 1, 'two': 1, 'sparkled': 1, 'gleamed': 1, 'distinctly': 1, 'hulking': 1, 'outline': 1, 'big': 1, 'wide': 1, 'gleaming': 1, 'stepped': 1, 'backward': 1, 'legs': 1, 'hit': 1, 'tripped': 1, 'flung': 1, 'break': 1, 'fall': 1, 'landed': 1, 'gutter': 1, 'deafening': 1, 'bang': 1, 'shield': 1, 'sudden': 1, 'blinding': 1}

PART 2

* **File 2 pages 98 to 108**

* **Book Name: Harry Potter and the Prisoner of Azkaban**

* **To count the number of occurrences of each non-English word in the second text file (file2.txt).**

Code:

```
!pip install pyspellchecker #installing the pyspellchecker library

#importing necessary files

from spellchecker import SpellChecker

from collections import Counter

import re


#using the library

harrypotter_spellwords = SpellChecker()

#removing the punctuations


#accessing the file 2 text and reading it

file_path = 'file2.txt'

with open(file_path, 'r', encoding='ISO-8859-1') as file:

    text = file.read()


#removing the punctuations and reading the words
```

```
punctuation = re.findall(r'\b\w+\b', text.lower())


#counting the occurance of the non english words from the file2
def count_nonenglish(word):
  return word not in harrypotter_spellwords


counting_nonenglish_words = Counter(word for word in punctuation if count_nonenglish(word))


#printing the results
print("The spells or non english words used are:")
print(counting_nonenglish_words)
```

Output:

The spells or non english words used are:

Counter({'hermione': 20, 'azkaban': 13, 'mcgonagall': 12, 'malfoy': 10, 've': 9, 'dementors': 9, 'hagrid': 9, 'gryffindor': 7, 'pomfrey': 7, 'slytherin': 5, 'snape': 5, 'dementor': 4, 'wasn': 4, 'weasley': 3, 'll': 3, 'kettleburn': 2, 'didn': 2, 'quidditch': 2, '98': 1, '99': 1, 'ofus': 1, 'flitwick': 1, '1100': 1, 'ravenclaw': 1, 'hufflepuff': 1, 'couldn': 1, 'albus': 1, '1101': 1, '1102': 1, 'rubeus': 1, 'gamekeeping': 1, '1103': 1, 'ter': 1, '104': 1, 'gryffindors': 1, 'longbottom': 1, '1105': 1, 'slytherins': 1, 'wooooooo': 1, '1106': 1, '1107': 1, 'isn': 1, 'muggle': 1, 'arithmancy': 1, 'righ': 1, 'gettin': 1, 'everythin': 1, 'ly': 1, '1108': 1, 'ofnorth': 1, 'hadn': 1})