Devolved AI

Subject : Tokenization (Chapter 2) documentation

Date of Submission : 28th March, 2024

Submitted To,

Nathan Peterson,
Founder and CEO,
Devolved AI

Md. Nazmul Hossain,
Co-Founder & COO,
Devolved AI

Submitted By,

Md Al Amin Tokder

Machine Learning Engineer ,
Devolved AI

enization-for-llms-from-scratch-2

March 28, 2024

1 Working with Text

Packages that are being used in this notebook:

[19]:
```
!pip install tiktoken
```

```
Requirement already satisfied: tiktoken in /opt/conda/lib/python3.10/site-
packages (0.6.0)
Requirement already satisfied: regex>=2022.1.18 in
/opt/conda/lib/python3.10/site-packages (from tiktoken) (2023.12.25)
Requirement already satisfied: requests>=2.26.0 in
/opt/conda/lib/python3.10/site-packages (from tiktoken) (2.31.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.10/site-packages (from requests>=2.26.0->tiktoken)
(3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-
packages (from requests>=2.26.0->tiktoken) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/opt/conda/lib/python3.10/site-packages (from requests>=2.26.0->tiktoken)
(1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.10/site-packages (from requests>=2.26.0->tiktoken)
(2024.2.2)
```

[20]:
```
from importlib.metadata import version

import tiktoken
import torch

print("torch version:", version("torch"))
print("tiktoken version:", version("tiktoken"))
```

```
torch version: 2.1.2+cpu
tiktoken version: 0.6.0
```

- This chapter covers data preparation and sampling to get input data "ready" for the LLM

1.1 2.1 Understanding word embeddings

- No code in this section
- There are any forms of embeddings; we focus on text embeddings in this book
- LLMs work embeddings in high-dimensional spaces (i.e., thousands of dimensions)
- Since we can't visualize such high-dimensional spaces (we humans think in 1, 2, or 3 dimensions), the figure below illustrates a 2-dimensipnal embedding space

1.2 2.2 Tokenizing text

- In this section, we tokenize text, which means breaking text into smaller units, such as individual words and punctuation characters
- Load raw text we want to work with

[21]:
```
with open("/kaggle/input/verdict-nlp/the-verdict.txt", "r", encoding="utf-8")␣
↪as f:
    raw_text = f.read()

print("Total number of characters:", len(raw_text))
print(raw_text[:103])
```

```
Total number of characters: 20479
I HAD always thought Jack Gisburn rather a cheap genius--though a good fellow
enough--so it was no grea
```

- The goal is to tokenize and embed this text for an LLM
- Let's develop a simple tokenizer based on some simple sample text that we can then later apply to the text above
- The following regular expression will split on whitespaces

[22]:
```
import re

text = "Hello, world. This, is a test.Acvx"
result = re.split(r'(\s)', text)

print(result)
```

```
['Hello,', ' ', 'world.', ' ', 'This,', ' ', 'is', ' ', 'a', ' ', 'test.Acvx']
```

[23]:
```
import re
text="Hello, world. thIS is a shoukhin.Abc"
text1 = "Hello, world. Is this-- a test?"

r1=re.split(' ',text)
r2=re.split(r'()\s',text)
print(r2)
```

```
['Hello,', '', 'world.', '', 'thIS', '', 'is', '', 'a', '', 'shoukhin.Abc']
```

- We don't only want to split on whitespaces but also commas and periods, so let's modify the regular expression to do that as well

[24]:
```
result = re.split(r'([,.]|\s)', text)

print(result)
```

```
['Hello', ',', '', ' ', 'world', '.', '', ' ', 'thIS', ' ', 'is', ' ', 'a', ' ',
'shoukhin', '.', 'Abc']
```

- As we can see, this creates empty strings, let's remove them

[25]:
```
# Strip whitespace from each item and then filter out any empty strings.
result = [item for item in result if item.strip()]
print(result)
```

```
['Hello', ',', 'world', '.', 'thIS', 'is', 'a', 'shoukhin', '.', 'Abc']
```

- This looks pretty good, but let's also handle other types of punctuation, such as periods, question marks, and so on

[42]:
```
text = "Hello, world. Is this-- a test?"

result = re.split(r'([,.:;?_!"()\']|--|\s)', text)
print(result)
result = [item.strip() for item in result if item.strip()] #space remove by␣
 ↪strip() function
print(result)
```

```
['Hello', ',', '', ' ', 'world', '.', '', ' ', 'Is', ' ', 'this', '--', '', ' ',
'a', ' ', 'test', '?', '']
['Hello', ',', 'world', '.', 'Is', 'this', '--', 'a', 'test', '?']
```

- This is pretty good, and we are now ready to apply this tokenization to the raw text

[52]:
```
preprocessed = re.split(r'([,.?_!"()\']|--|\s)', raw_text) #verdict dataset␣
 ↪theke raw_text data nilam
preprocessed = [item.strip() for item in preprocessed if item.strip()] #white␣
 ↪space removed korlam

# print(preprocessed)
print(preprocessed[:30])
print("\n  Dataset er length : ", len(preprocessed))
```

```
['I', 'HAD', 'always', 'thought', 'Jack', 'Gisburn', 'rather', 'a', 'cheap',
'genius', '--', 'though', 'a', 'good', 'fellow', 'enough', '--', 'so', 'it',
'was', 'no', 'great', 'surprise', 'to', 'me', 'to', 'hear', 'that', ',', 'in']

  Dataset er length :  4649
```

- Let's calculate the total number of tokens

[28]:
```
print(len(preprocessed))
```

4649

1.3 2.3 Converting tokens into token IDs

- Next, we convert the text tokens into token IDs that we can process via embedding layers later
- From these tokens, we can now build a vocabulary that consists of all the **unique tokens**

[62]:
```
# Preprocessing on raw_text data

preprocessed = re.split(r'([,.?_!"()\']|--|\s)', raw_text) #verdict dataset theke raw_text data nilam
preprocessed = [item.strip() for item in preprocessed if item.strip()] #white space removed korlam
print(preprocessed)
print(len(preprocessed))

# print(preprocessed[:30])
# print("\n  Dataset er length : ", len(preprocessed))









#Takng Only Uique Word


all_words = sorted(list(set(preprocessed)))   #Similar to stopword in NLP
print(all_words)

#Vocab_size  =len(all_words)

print(len(all_words))



#Size reduce hoye 4649 theke 1159 hoye gelo
```

```
['I', 'HAD', 'always', 'thought', 'Jack', 'Gisburn', 'rather', 'a', 'cheap',
'genius', '--', 'though', 'a', 'good', 'fellow', 'enough', '--', 'so', 'it',
```

'was', 'no', 'great', 'surprise', 'to', 'me', 'to', 'hear', 'that', ',', 'in',
'the', 'height', 'of', 'his', 'glory', ',', 'he', 'had', 'dropped', 'his',
'painting', ',', 'married', 'a', 'rich', 'widow', ',', 'and', 'established',
'himself', 'in', 'a', 'villa', 'on', 'the', 'Riviera', '.', '(', 'Though', 'I',
'rather', 'thought', 'it', 'would', 'have', 'been', 'Rome', 'or', 'Florence',
'.', ')', '"', 'The', 'height', 'of', 'his', 'glory', '"', '--', 'that', 'was',
'what', 'the', 'women', 'called', 'it', '.', 'I', 'can', 'hear', 'Mrs', '.',
'Gideon', 'Thwing', '--', 'his', 'last', 'Chicago', 'sitter', '--', 'deploring',
'his', 'unaccountable', 'abdication', '.', '"', 'Of', 'course', 'it', "'", 's',
'going', 'to', 'send', 'the', 'value', 'of', 'my', 'picture', "'", 'way', 'up;',
'but', 'I', 'don', "'", 't', 'think', 'of', 'that', ',', 'Mr', '.', 'Rickham',
'--', 'the', 'loss', 'to', 'Arrt', 'is', 'all', 'I', 'think', 'of', '.', '"',
'The', 'word', ',', 'on', 'Mrs', '.', 'Thwing', "'", 's', 'lips', ',',
'multiplied', 'its', '_', 'rs', '_', 'as', 'though', 'they', 'were',
'reflected', 'in', 'an', 'endless', 'vista', 'of', 'mirrors', '.', 'And', 'it',
'was', 'not', 'only', 'the', 'Mrs', '.', 'Thwings', 'who', 'mourned', '.',
'Had', 'not', 'the', 'exquisite', 'Hermia', 'Croft', ',', 'at', 'the', 'last',
'Grafton', 'Gallery', 'show', ',', 'stopped', 'me', 'before', 'Gisburn', "'",
's', '"', 'Moon-dancers', '"', 'to', 'say', ',', 'with', 'tears', 'in', 'her',
'eyes:', '"', 'We', 'shall', 'not', 'look', 'upon', 'its', 'like', 'again', '"',
'?', 'Well', '!', '--', 'even', 'through', 'the', 'prism', 'of', 'Hermia', "'",
's', 'tears', 'I', 'felt', 'able', 'to', 'face', 'the', 'fact', 'with',
'equanimity', '.', 'Poor', 'Jack', 'Gisburn', '!', 'The', 'women', 'had',
'made', 'him', '--', 'it', 'was', 'fitting', 'that', 'they', 'should', 'mourn',
'him', '.', 'Among', 'his', 'own', 'sex', 'fewer', 'regrets', 'were', 'heard',
',', 'and', 'in', 'his', 'own', 'trade', 'hardly', 'a', 'murmur', '.',
'Professional', 'jealousy', '?', 'Perhaps', '.', 'If', 'it', 'were', ',', 'the',
'honour', 'of', 'the', 'craft', 'was', 'vindicated', 'by', 'little', 'Claude',
'Nutley', ',', 'who', ',', 'in', 'all', 'good', 'faith', ',', 'brought', 'out',
'in', 'the', 'Burlington', 'a', 'very', 'handsome', '"', 'obituary', '"', 'on',
'Jack', '--', 'one', 'of', 'those', 'showy', 'articles', 'stocked', 'with',
'random', 'technicalities', 'that', 'I', 'have', 'heard', '(', 'I', 'won', "'",
't', 'say', 'by', 'whom', ')', 'compared', 'to', 'Gisburn', "'", 's',
'painting', '.', 'And', 'so', '--', 'his', 'resolve', 'being', 'apparently',
'irrevocable', '--', 'the', 'discussion', 'gradually', 'died', 'out', ',',
'and', ',', 'as', 'Mrs', '.', 'Thwing', 'had', 'predicted', ',', 'the', 'price',
'of', '"', 'Gisburns', '"', 'went', 'up', '.', 'It', 'was', 'not', 'till',
'three', 'years', 'later', 'that', ',', 'in', 'the', 'course', 'of', 'a', 'few',
'weeks', "'", 'idling', 'on', 'the', 'Riviera', ',', 'it', 'suddenly',
'occurred', 'to', 'me', 'to', 'wonder', 'why', 'Gisburn', 'had', 'given', 'up',
'his', 'painting', '.', 'On', 'reflection', ',', 'it', 'really', 'was', 'a',
'tempting', 'problem', '.', 'To', 'accuse', 'his', 'wife', 'would', 'have',
'been', 'too', 'easy', '--', 'his', 'fair', 'sitters', 'had', 'been', 'denied',
'the', 'solace', 'of', 'saying', 'that', 'Mrs', '.', 'Gisburn', 'had', '"',
'dragged', 'him', 'down', '.', '"', 'For', 'Mrs', '.', 'Gisburn', '--', 'as',
'such', '--', 'had', 'not', 'existed', 'till', 'nearly', 'a', 'year', 'after',
'Jack', "'", 's', 'resolve', 'had', 'been', 'taken', '.', 'It', 'might', 'be',
'that', 'he', 'had', 'married', 'her', '--', 'since', 'he', 'liked', 'his',

```
'ease', '--', 'because', 'he', 'didn', "'", 't', 'want', 'to', 'go', 'on',
'painting;', 'but', 'it', 'would', 'have', 'been', 'hard', 'to', 'prove',
'that', 'he', 'had', 'given', 'up', 'his', 'painting', 'because', 'he', 'had',
'married', 'her', '.', 'Of', 'course', ',', 'if', 'she', 'had', 'not',
'dragged', 'him', 'down', ',', 'she', 'had', 'equally', ',', 'as', 'Miss',
'Croft', 'contended', ',', 'failed', 'to', '"', 'lift', 'him', 'up', '"', '--',
'she', 'had', 'not', 'led', 'him', 'back', 'to', 'the', 'easel', '.', 'To',
'put', 'the', 'brush', 'into', 'his', 'hand', 'again', '--', 'what', 'a',
'vocation', 'for', 'a', 'wife', '!', 'But', 'Mrs', '.', 'Gisburn', 'appeared',
'to', 'have', 'disdained', 'it', '--', 'and', 'I', 'felt', 'it', 'might', 'be',
'interesting', 'to', 'find', 'out', 'why', '.', 'The', 'desultory', 'life',
'of', 'the', 'Riviera', 'lends', 'itself', 'to', 'such', 'purely', 'academic',
'speculations;', 'and', 'having', ',', 'on', 'my', 'way', 'to', 'Monte',
'Carlo', ',', 'caught', 'a', 'glimpse', 'of', 'Jack', "'", 's', 'balustraded',
'terraces', 'between', 'the', 'pines', ',', 'I', 'had', 'myself', 'borne',
'thither', 'the', 'next', 'day', '.', 'I', 'found', 'the', 'couple', 'at',
'tea', 'beneath', 'their', 'palm-trees;', 'and', 'Mrs', '.', 'Gisburn', "'",
's', 'welcome', 'was', 'so', 'genial', 'that', ',', 'in', 'the', 'ensuing',
'weeks', ',', 'I', 'claimed', 'it', 'frequently', '.', 'It', 'was', 'not',
'that', 'my', 'hostess', 'was', '"', 'interesting', '"', ':', 'on', 'that',
'point', 'I', 'could', 'have', 'given', 'Miss', 'Croft', 'the', 'fullest',
'reassurance', '.', 'It', 'was', 'just', 'because', 'she', 'was', '_', 'not',
'_', 'interesting', '--', 'if', 'I', 'may', 'be', 'pardoned', 'the', 'bull',
'--', 'that', 'I', 'found', 'her', 'so', '.', 'For', 'Jack', ',', 'all', 'his',
'life', ',', 'had', 'been', 'surrounded', 'by', 'interesting', 'women:', 'they',
'had', 'fostered', 'his', 'art', ',', 'it', 'had', 'been', 'reared', 'in',
'the', 'hot-house', 'of', 'their', 'adulation', '.', 'And', 'it', 'was',
'therefore', 'instructive', 'to', 'note', 'what', 'effect', 'the', '"',
'deadening', 'atmosphere', 'of', 'mediocrity', '"', '(', 'I', 'quote', 'Miss',
'Croft', ')', 'was', 'having', 'on', 'him', '.', 'I', 'have', 'mentioned',
'that', 'Mrs', '.', 'Gisburn', 'was', 'rich;', 'and', 'it', 'was',
'immediately', 'perceptible', 'that', 'her', 'husband', 'was', 'extracting',
'from', 'this', 'circumstance', 'a', 'delicate', 'but', 'substantial',
'satisfaction', '.', 'It', 'is', ',', 'as', 'a', 'rule', ',', 'the', 'people',
'who', 'scorn', 'money', 'who', 'get', 'most', 'out', 'of', 'it;', 'and',
'Jack', "'", 's', 'elegant', 'disdain', 'of', 'his', 'wife', "'", 's', 'big',
'balance', 'enabled', 'him', ',', 'with', 'an', 'appearance', 'of', 'perfect',
'good-breeding', ',', 'to', 'transmute', 'it', 'into', 'objects', 'of', 'art',
'and', 'luxury', '.', 'To', 'the', 'latter', ',', 'I', 'must', 'add', ',', 'he',
'remained', 'relatively', 'indifferent;', 'but', 'he', 'was', 'buying',
'Renaissance', 'bronzes', 'and', 'eighteenth-century', 'pictures', 'with', 'a',
'discrimination', 'that', 'bespoke', 'the', 'amplest', 'resources', '.', '"',
'Money', "'", 's', 'only', 'excuse', 'is', 'to', 'put', 'beauty', 'into',
'circulation', ',', '"', 'was', 'one', 'of', 'the', 'axioms', 'he', 'laid',
'down', 'across', 'the', 'Sevres', 'and', 'silver', 'of', 'an', 'exquisitely',
'appointed', 'luncheon-table', ',', 'when', ',', 'on', 'a', 'later', 'day', ',',
'I', 'had', 'again', 'run', 'over', 'from', 'Monte', 'Carlo;', 'and', 'Mrs',
'.', 'Gisburn', ',', 'beaming', 'on', 'him', ',', 'added', 'for', 'my',
```

```
'enlightenment:', '"', 'Jack', 'is', 'so', 'morbidly', 'sensitive', 'to',
'every', 'form', 'of', 'beauty', '.', '"', 'Poor', 'Jack', '!', 'It', 'had',
'always', 'been', 'his', 'fate', 'to', 'have', 'women', 'say', 'such', 'things',
'of', 'him:', 'the', 'fact', 'should', 'be', 'set', 'down', 'in', 'extenuation',
'.', 'What', 'struck', 'me', 'now', 'was', 'that', ',', 'for', 'the', 'first',
'time', ',', 'he', 'resented', 'the', 'tone', '.', 'I', 'had', 'seen', 'him',
',', 'so', 'often', ',', 'basking', 'under', 'similar', 'tributes', '--', 'was',
'it', 'the', 'conjugal', 'note', 'that', 'robbed', 'them', 'of', 'their',
'savour', '?', 'No', '--', 'for', ',', 'oddly', 'enough', ',', 'it', 'became',
'apparent', 'that', 'he', 'was', 'fond', 'of', 'Mrs', '.', 'Gisburn', '--',
'fond', 'enough', 'not', 'to', 'see', 'her', 'absurdity', '.', 'It', 'was',
'his', 'own', 'absurdity', 'he', 'seemed', 'to', 'be', 'wincing', 'under', '--',
'his', 'own', 'attitude', 'as', 'an', 'object', 'for', 'garlands', 'and',
'incense', '.', '"', 'My', 'dear', ',', 'since', 'I', "'", 've', 'chucked',
'painting', 'people', 'don', "'", 't', 'say', 'that', 'stuff', 'about', 'me',
'--', 'they', 'say', 'it', 'about', 'Victor', 'Grindle', ',', '"', 'was', 'his',
'only', 'protest', ',', 'as', 'he', 'rose', 'from', 'the', 'table', 'and',
'strolled', 'out', 'onto', 'the', 'sunlit', 'terrace', '.', 'I', 'glanced',
'after', 'him', ',', 'struck', 'by', 'his', 'last', 'word', '.', 'Victor',
'Grindle', 'was', ',', 'in', 'fact', ',', 'becoming', 'the', 'man', 'of', 'the',
'moment', '--', 'as', 'Jack', 'himself', ',', 'one', 'might', 'put', 'it', ',',
'had', 'been', 'the', 'man', 'of', 'the', 'hour', '.', 'The', 'younger',
'artist', 'was', 'said', 'to', 'have', 'formed', 'himself', 'at', 'my',
'friend', "'", 's', 'feet', ',', 'and', 'I', 'wondered', 'if', 'a', 'tinge',
'of', 'jealousy', 'underlay', 'the', 'latter', "'", 's', 'mysterious',
'abdication', '.', 'But', 'no', '--', 'for', 'it', 'was', 'not', 'till',
'after', 'that', 'event', 'that', 'the', '_', 'rose', 'Dubarry', '_', 'drawing-
rooms', 'had', 'begun', 'to', 'display', 'their', '"', 'Grindles', '.', '"',
'I', 'turned', 'to', 'Mrs', '.', 'Gisburn', ',', 'who', 'had', 'lingered', 'to',
'give', 'a', 'lump', 'of', 'sugar', 'to', 'her', 'spaniel', 'in', 'the',
'dining-room', '.', '"', 'Why', '_', 'has', '_', 'he', 'chucked', 'painting',
'?', '"', 'I', 'asked', 'abruptly', '.', 'She', 'raised', 'her', 'eyebrows',
'with', 'a', 'hint', 'of', 'good-humoured', 'surprise', '.', '"', 'Oh', ',',
'he', 'doesn', "'", 't', '_', 'have', '_', 'to', 'now', ',', 'you', 'know;',
'and', 'I', 'want', 'him', 'to', 'enjoy', 'himself', ',', '"', 'she', 'said',
'quite', 'simply', '.', 'I', 'looked', 'about', 'the', 'spacious', 'white-
panelled', 'room', ',', 'with', 'its', '_', 'famille-verte', '_', 'vases',
'repeating', 'the', 'tones', 'of', 'the', 'pale', 'damask', 'curtains', ',',
'and', 'its', 'eighteenth-century', 'pastels', 'in', 'delicate', 'faded',
'frames', '.', '"', 'Has', 'he', 'chucked', 'his', 'pictures', 'too', '?', 'I',
'haven', "'", 't', 'seen', 'a', 'single', 'one', 'in', 'the', 'house', '.', '"',
'A', 'slight', 'shade', 'of', 'constraint', 'crossed', 'Mrs', '.', 'Gisburn',
"'", 's', 'open', 'countenance', '.', '"', 'It', "'", 's', 'his', 'ridiculous',
'modesty', ',', 'you', 'know', '.', 'He', 'says', 'they', "'", 're', 'not',
'fit', 'to', 'have', 'about;', 'he', "'", 's', 'sent', 'them', 'all', 'away',
'except', 'one', '--', 'my', 'portrait', '--', 'and', 'that', 'I', 'have', 'to',
'keep', 'upstairs', '.', '"', 'His', 'ridiculous', 'modesty', '--', 'Jack', "'",
's', 'modesty', 'about', 'his', 'pictures', '?', 'My', 'curiosity', 'was',
```

'growing', 'like', 'the', 'bean-stalk', '.', 'I', 'said', 'persuasively', 'to',
'my', 'hostess:', '"', 'I', 'must', 'really', 'see', 'your', 'portrait', ',',
'you', 'know', '.', '"', 'She', 'glanced', 'out', 'almost', 'timorously', 'at',
'the', 'terrace', 'where', 'her', 'husband', ',', 'lounging', 'in', 'a',
'hooded', 'chair', ',', 'had', 'lit', 'a', 'cigar', 'and', 'drawn', 'the',
'Russian', 'deerhound', "'", 's', 'head', 'between', 'his', 'knees', '.', '"',
'Well', ',', 'come', 'while', 'he', "'", 's', 'not', 'looking', ',', '"', 'she',
'said', ',', 'with', 'a', 'laugh', 'that', 'tried', 'to', 'hide', 'her',
'nervousness;', 'and', 'I', 'followed', 'her', 'between', 'the', 'marble',
'Emperors', 'of', 'the', 'hall', ',', 'and', 'up', 'the', 'wide', 'stairs',
'with', 'terra-cotta', 'nymphs', 'poised', 'among', 'flowers', 'at', 'each',
'landing', '.', 'In', 'the', 'dimmest', 'corner', 'of', 'her', 'boudoir', ',',
'amid', 'a', 'profusion', 'of', 'delicate', 'and', 'distinguished', 'objects',
',', 'hung', 'one', 'of', 'the', 'familiar', 'oval', 'canvases', ',', 'in',
'the', 'inevitable', 'garlanded', 'frame', '.', 'The', 'mere', 'outline', 'of',
'the', 'frame', 'called', 'up', 'all', 'Gisburn', "'", 's', 'past', '!', 'Mrs',
'.', 'Gisburn', 'drew', 'back', 'the', 'window-curtains', ',', 'moved', 'aside',
'a', '_', 'jardiniere', '_', 'full', 'of', 'pink', 'azaleas', ',', 'pushed',
'an', 'arm-chair', 'away', ',', 'and', 'said:', '"', 'If', 'you', 'stand',
'here', 'you', 'can', 'just', 'manage', 'to', 'see', 'it', '.', 'I', 'had',
'it', 'over', 'the', 'mantel-piece', ',', 'but', 'he', 'wouldn', "'", 't',
'let', 'it', 'stay', '.', '"', 'Yes', '--', 'I', 'could', 'just', 'manage',
'to', 'see', 'it', '--', 'the', 'first', 'portrait', 'of', 'Jack', "'", 's',
'I', 'had', 'ever', 'had', 'to', 'strain', 'my', 'eyes', 'over', '!', 'Usually',
'they', 'had', 'the', 'place', 'of', 'honour', '--', 'say', 'the', 'central',
'panel', 'in', 'a', 'pale', 'yellow', 'or', '_', 'rose', 'Dubarry', '_',
'drawing-room', ',', 'or', 'a', 'monumental', 'easel', 'placed', 'so', 'that',
'it', 'took', 'the', 'light', 'through', 'curtains', 'of', 'old', 'Venetian',
'point', '.', 'The', 'more', 'modest', 'place', 'became', 'the', 'picture',
'better;', 'yet', ',', 'as', 'my', 'eyes', 'grew', 'accustomed', 'to', 'the',
'half-light', ',', 'all', 'the', 'characteristic', 'qualities', 'came', 'out',
'--', 'all', 'the', 'hesitations', 'disguised', 'as', 'audacities', ',', 'the',
'tricks', 'of', 'prestidigitation', 'by', 'which', ',', 'with', 'such',
'consummate', 'skill', ',', 'he', 'managed', 'to', 'divert', 'attention',
'from', 'the', 'real', 'business', 'of', 'the', 'picture', 'to', 'some',
'pretty', 'irrelevance', 'of', 'detail', '.', 'Mrs', '.', 'Gisburn', ',',
'presenting', 'a', 'neutral', 'surface', 'to', 'work', 'on', '--', 'forming',
',', 'as', 'it', 'were', ',', 'so', 'inevitably', 'the', 'background', 'of',
'her', 'own', 'picture', '--', 'had', 'lent', 'herself', 'in', 'an', 'unusual',
'degree', 'to', 'the', 'display', 'of', 'this', 'false', 'virtuosity', '.',
'The', 'picture', 'was', 'one', 'of', 'Jack', "'", 's', '"', 'strongest', ',',
'"', 'as', 'his', 'admirers', 'would', 'have', 'put', 'it', '--', 'it',
'represented', ',', 'on', 'his', 'part', ',', 'a', 'swelling', 'of', 'muscles',
',', 'a', 'congesting', 'of', 'veins', ',', 'a', 'balancing', ',', 'straddling',
'and', 'straining', ',', 'that', 'reminded', 'one', 'of', 'the', 'circus-clown',
"'", 's', 'ironic', 'efforts', 'to', 'lift', 'a', 'feather', '.', 'It', 'met',
',', 'in', 'short', ',', 'at', 'every', 'point', 'the', 'demand', 'of',
'lovely', 'woman', 'to', 'be', 'painted', '"', 'strongly', '"', 'because',

```
'she', 'was', 'tired', 'of', 'being', 'painted', '"', 'sweetly', '"', '--',
'and', 'yet', 'not', 'to', 'lose', 'an', 'atom', 'of', 'the', 'sweetness', '.',
'"', 'It', "'", 's', 'the', 'last', 'he', 'painted', ',', 'you', 'know', ',',
'"', 'Mrs', '.', 'Gisburn', 'said', 'with', 'pardonable', 'pride', '.', '"',
'The', 'last', 'but', 'one', ',', '"', 'she', 'corrected', 'herself', '--', '"',
'but', 'the', 'other', 'doesn', "'", 't', 'count', ',', 'because', 'he',
'destroyed', 'it', '.', '"', '"', 'Destroyed', 'it', '?', '"', 'I', 'was',
'about', 'to', 'follow', 'up', 'this', 'clue', 'when', 'I', 'heard', 'a',
'footstep', 'and', 'saw', 'Jack', 'himself', 'on', 'the', 'threshold', '.',
'As', 'he', 'stood', 'there', ',', 'his', 'hands', 'in', 'the', 'pockets', 'of',
'his', 'velveteen', 'coat', ',', 'the', 'thin', 'brown', 'waves', 'of', 'hair',
'pushed', 'back', 'from', 'his', 'white', 'forehead', ',', 'his', 'lean',
'sunburnt', 'cheeks', 'furrowed', 'by', 'a', 'smile', 'that', 'lifted', 'the',
'tips', 'of', 'a', 'self-confident', 'moustache', ',', 'I', 'felt', 'to',
'what', 'a', 'degree', 'he', 'had', 'the', 'same', 'quality', 'as', 'his',
'pictures', '--', 'the', 'quality', 'of', 'looking', 'cleverer', 'than', 'he',
'was', '.', 'His', 'wife', 'glanced', 'at', 'him', 'deprecatingly', ',', 'but',
'his', 'eyes', 'travelled', 'past', 'her', 'to', 'the', 'portrait', '.', '"',
'Mr', '.', 'Rickham', 'wanted', 'to', 'see', 'it', ',', '"', 'she', 'began',
',', 'as', 'if', 'excusing', 'herself', '.', 'He', 'shrugged', 'his',
'shoulders', ',', 'still', 'smiling', '.', '"', 'Oh', ',', 'Rickham', 'found',
'me', 'out', 'long', 'ago', ',', '"', 'he', 'said', 'lightly;', 'then', ',',
'passing', 'his', 'arm', 'through', 'mine:', '"', 'Come', 'and', 'see', 'the',
'rest', 'of', 'the', 'house', '.', '"', 'He', 'showed', 'it', 'to', 'me',
'with', 'a', 'kind', 'of', 'naive', 'suburban', 'pride:', 'the', 'bath-rooms',
',', 'the', 'speaking-tubes', ',', 'the', 'dress-closets', ',', 'the', 'trouser-
presses', '--', 'all', 'the', 'complex', 'simplifications', 'of', 'the',
'millionaire', "'", 's', 'domestic', 'economy', '.', 'And', 'whenever', 'my',
'wonder', 'paid', 'the', 'expected', 'tribute', 'he', 'said', ',', 'throwing',
'out', 'his', 'chest', 'a', 'little:', '"', 'Yes', ',', 'I', 'really', 'don',
"'", 't', 'see', 'how', 'people', 'manage', 'to', 'live', 'without', 'that',
'.', '"', 'Well', '--', 'it', 'was', 'just', 'the', 'end', 'one', 'might',
'have', 'foreseen', 'for', 'him', '.', 'Only', 'he', 'was', ',', 'through',
'it', 'all', 'and', 'in', 'spite', 'of', 'it', 'all', '--', 'as', 'he', 'had',
'been', 'through', ',', 'and', 'in', 'spite', 'of', ',', 'his', 'pictures',
'--', 'so', 'handsome', ',', 'so', 'charming', ',', 'so', 'disarming', ',',
'that', 'one', 'longed', 'to', 'cry', 'out:', '"', 'Be', 'dissatisfied', 'with',
'your', 'leisure', '!', '"', 'as', 'once', 'one', 'had', 'longed', 'to', 'say:',
'"', 'Be', 'dissatisfied', 'with', 'your', 'work', '!', '"', 'But', ',', 'with',
'the', 'cry', 'on', 'my', 'lips', ',', 'my', 'diagnosis', 'suffered', 'an',
'unexpected', 'check', '.', '"', 'This', 'is', 'my', 'own', 'lair', ',', '"',
'he', 'said', ',', 'leading', 'me', 'into', 'a', 'dark', 'plain', 'room', 'at',
'the', 'end', 'of', 'the', 'florid', 'vista', '.', 'It', 'was', 'square', 'and',
'brown', 'and', 'leathery:', 'no', '"', 'effects', '"', ';', 'no', 'bric-a-
brac', ',', 'none', 'of', 'the', 'air', 'of', 'posing', 'for', 'reproduction',
'in', 'a', 'picture', 'weekly', '--', 'above', 'all', ',', 'no', 'least',
'sign', 'of', 'ever', 'having', 'been', 'used', 'as', 'a', 'studio', '.', 'The',
'fact', 'brought', 'home', 'to', 'me', 'the', 'absolute', 'finality', 'of',
```

```
'Jack', "'", 's', 'break', 'with', 'his', 'old', 'life', '.', '"', 'Don', "'",
't', 'you', 'ever', 'dabble', 'with', 'paint', 'any', 'more', '?', '"', 'I',
'asked', ',', 'still', 'looking', 'about', 'for', 'a', 'trace', 'of', 'such',
'activity', '.', '"', 'Never', ',', '"', 'he', 'said', 'briefly', '.', '"',
'Or', 'water-colour', '--', 'or', 'etching', '?', '"', 'His', 'confident',
'eyes', 'grew', 'dim', ',', 'and', 'his', 'cheeks', 'paled', 'a', 'little',
'under', 'their', 'handsome', 'sunburn', '.', '"', 'Never', 'think', 'of', 'it',
',', 'my', 'dear', 'fellow', '--', 'any', 'more', 'than', 'if', 'I', "'", 'd',
'never', 'touched', 'a', 'brush', '.', '"', 'And', 'his', 'tone', 'told', 'me',
'in', 'a', 'flash', 'that', 'he', 'never', 'thought', 'of', 'anything', 'else',
'.', 'I', 'moved', 'away', ',', 'instinctively', 'embarrassed', 'by', 'my',
'unexpected', 'discovery;', 'and', 'as', 'I', 'turned', ',', 'my', 'eye',
'fell', 'on', 'a', 'small', 'picture', 'above', 'the', 'mantel-piece', '--',
'the', 'only', 'object', 'breaking', 'the', 'plain', 'oak', 'panelling', 'of',
'the', 'room', '.', '"', 'Oh', ',', 'by', 'Jove', '!', '"', 'I', 'said', '.',
'It', 'was', 'a', 'sketch', 'of', 'a', 'donkey', '--', 'an', 'old', 'tired',
'donkey', ',', 'standing', 'in', 'the', 'rain', 'under', 'a', 'wall', '.', '"',
'By', 'Jove', '--', 'a', 'Stroud', '!', '"', 'I', 'cried', '.', 'He', 'was',
'silent;', 'but', 'I', 'felt', 'him', 'close', 'behind', 'me', ',', 'breathing',
'a', 'little', 'quickly', '.', '"', 'What', 'a', 'wonder', '!', 'Made', 'with',
'a', 'dozen', 'lines', '--', 'but', 'on', 'everlasting', 'foundations', '.',
'You', 'lucky', 'chap', ',', 'where', 'did', 'you', 'get', 'it', '?', '"', 'He',
'answered', 'slowly:', '"', 'Mrs', '.', 'Stroud', 'gave', 'it', 'to', 'me', '.',
'"', '"', 'Ah', '--', 'I', 'didn', "'", 't', 'know', 'you', 'even', 'knew',
'the', 'Strouds', '.', 'He', 'was', 'such', 'an', 'inflexible', 'hermit', '.',
'"', '"', 'I', 'didn', "'", 't', '--', 'till', 'after', '.', '.', '.', '.',
'She', 'sent', 'for', 'me', 'to', 'paint', 'him', 'when', 'he', 'was', 'dead',
'.', '"', '"', 'When', 'he', 'was', 'dead', '?', 'You', '?', '"', 'I', 'must',
'have', 'let', 'a', 'little', 'too', 'much', 'amazement', 'escape', 'through',
'my', 'surprise', ',', 'for', 'he', 'answered', 'with', 'a', 'deprecating',
'laugh:', '"', 'Yes', '--', 'she', "'", 's', 'an', 'awful', 'simpleton', ',',
'you', 'know', ',', 'Mrs', '.', 'Stroud', '.', 'Her', 'only', 'idea', 'was',
'to', 'have', 'him', 'done', 'by', 'a', 'fashionable', 'painter', '--', 'ah',
',', 'poor', 'Stroud', '!', 'She', 'thought', 'it', 'the', 'surest', 'way',
'of', 'proclaiming', 'his', 'greatness', '--', 'of', 'forcing', 'it', 'on', 'a',
'purblind', 'public', '.', 'And', 'at', 'the', 'moment', 'I', 'was', '_', 'the',
'_', 'fashionable', 'painter', '.', '"', '"', 'Ah', ',', 'poor', 'Stroud', '--',
'as', 'you', 'say', '.', 'Was', '_', 'that', '_', 'his', 'history', '?', '"',
'"', 'That', 'was', 'his', 'history', '.', 'She', 'believed', 'in', 'him', ',',
'gloried', 'in', 'him', '--', 'or', 'thought', 'she', 'did', '.', 'But', 'she',
'couldn', "'", 't', 'bear', 'not', 'to', 'have', 'all', 'the', 'drawing-rooms',
'with', 'her', '.', 'She', 'couldn', "'", 't', 'bear', 'the', 'fact', 'that',
',', 'on', 'varnishing', 'days', ',', 'one', 'could', 'always', 'get', 'near',
'enough', 'to', 'see', 'his', 'pictures', '.', 'Poor', 'woman', '!', 'She', "'",
's', 'just', 'a', 'fragment', 'groping', 'for', 'other', 'fragments', '.',
'Stroud', 'is', 'the', 'only', 'whole', 'I', 'ever', 'knew', '.', '"', '"',
'You', 'ever', 'knew', '?', 'But', 'you', 'just', 'said', '--', '"', 'Gisburn',
'had', 'a', 'curious', 'smile', 'in', 'his', 'eyes', '.', '"', 'Oh', ',', 'I',
```

```
'knew', 'him', ',', 'and', 'he', 'knew', 'me', '--', 'only', 'it', 'happened',
'after', 'he', 'was', 'dead', '.', '"', 'I', 'dropped', 'my', 'voice',
'instinctively', '.', '"', 'When', 'she', 'sent', 'for', 'you', '?', '"', '"',
'Yes', '--', 'quite', 'insensible', 'to', 'the', 'irony', '.', 'She', 'wanted',
'him', 'vindicated', '--', 'and', 'by', 'me', '!', '"', 'He', 'laughed',
'again', ',', 'and', 'threw', 'back', 'his', 'head', 'to', 'look', 'up', 'at',
'the', 'sketch', 'of', 'the', 'donkey', '.', '"', 'There', 'were', 'days',
'when', 'I', 'couldn', "'", 't', 'look', 'at', 'that', 'thing', '--', 'couldn',
"'", 't', 'face', 'it', '.', 'But', 'I', 'forced', 'myself', 'to', 'put', 'it',
'here;', 'and', 'now', 'it', "'", 's', 'cured', 'me', '--', 'cured', 'me', '.',
'That', "'", 's', 'the', 'reason', 'why', 'I', 'don', "'", 't', 'dabble', 'any',
'more', ',', 'my', 'dear', 'Rickham;', 'or', 'rather', 'Stroud', 'himself',
'is', 'the', 'reason', '.', '"', 'For', 'the', 'first', 'time', 'my', 'idle',
'curiosity', 'about', 'my', 'companion', 'turned', 'into', 'a', 'serious',
'desire', 'to', 'understand', 'him', 'better', '.', '"', 'I', 'wish', 'you',
"'", 'd', 'tell', 'me', 'how', 'it', 'happened', ',', '"', 'I', 'said', '.',
'He', 'stood', 'looking', 'up', 'at', 'the', 'sketch', ',', 'and', 'twirling',
'between', 'his', 'fingers', 'a', 'cigarette', 'he', 'had', 'forgotten', 'to',
'light', '.', 'Suddenly', 'he', 'turned', 'toward', 'me', '.', '"', 'I', "'",
'd', 'rather', 'like', 'to', 'tell', 'you', '--', 'because', 'I', "'", 've',
'always', 'suspected', 'you', 'of', 'loathing', 'my', 'work', '.', '"', 'I',
'made', 'a', 'deprecating', 'gesture', ',', 'which', 'he', 'negatived', 'with',
'a', 'good-humoured', 'shrug', '.', '"', 'Oh', ',', 'I', 'didn', "'", 't',
'care', 'a', 'straw', 'when', 'I', 'believed', 'in', 'myself', '--', 'and',
'now', 'it', "'", 's', 'an', 'added', 'tie', 'between', 'us', '!', '"', 'He',
'laughed', 'slightly', ',', 'without', 'bitterness', ',', 'and', 'pushed',
'one', 'of', 'the', 'deep', 'arm-chairs', 'forward', '.', '"', 'There:', 'make',
'yourself', 'comfortable', '--', 'and', 'here', 'are', 'the', 'cigars', 'you',
'like', '.', '"', 'He', 'placed', 'them', 'at', 'my', 'elbow', 'and',
'continued', 'to', 'wander', 'up', 'and', 'down', 'the', 'room', ',',
'stopping', 'now', 'and', 'then', 'beneath', 'the', 'picture', '.', '"', 'How',
'it', 'happened', '?', 'I', 'can', 'tell', 'you', 'in', 'five', 'minutes', '--',
'and', 'it', 'didn', "'", 't', 'take', 'much', 'longer', 'to', 'happen', '.',
'.', '.', '.', 'I', 'can', 'remember', 'now', 'how', 'surprised', 'and',
'pleased', 'I', 'was', 'when', 'I', 'got', 'Mrs', '.', 'Stroud', "'", 's',
'note', '.', 'Of', 'course', ',', 'deep', 'down', ',', 'I', 'had', 'always',
'_', 'felt', '_', 'there', 'was', 'no', 'one', 'like', 'him', '--', 'only', 'I',
'had', 'gone', 'with', 'the', 'stream', ',', 'echoed', 'the', 'usual',
'platitudes', 'about', 'him', ',', 'till', 'I', 'half', 'got', 'to', 'think',
'he', 'was', 'a', 'failure', ',', 'one', 'of', 'the', 'kind', 'that', 'are',
'left', 'behind', '.', 'By', 'Jove', ',', 'and', 'he', '_', 'was', '_', 'left',
'behind', '--', 'because', 'he', 'had', 'come', 'to', 'stay', '!', 'The',
'rest', 'of', 'us', 'had', 'to', 'let', 'ourselves', 'be', 'swept', 'along',
'or', 'go', 'under', ',', 'but', 'he', 'was', 'high', 'above', 'the', 'current',
'--', 'on', 'everlasting', 'foundations', ',', 'as', 'you', 'say', '.', '"',
'Well', ',', 'I', 'went', 'off', 'to', 'the', 'house', 'in', 'my', 'most',
'egregious', 'mood', '--', 'rather', 'moved', ',', 'Lord', 'forgive', 'me', ',',
'at', 'the', 'pathos', 'of', 'poor', 'Stroud', "'", 's', 'career', 'of',
```

'failure', 'being', 'crowned', 'by', 'the', 'glory', 'of', 'my', 'painting',
'him', '!', 'Of', 'course', 'I', 'meant', 'to', 'do', 'the', 'picture', 'for',
'nothing', '--', 'I', 'told', 'Mrs', '.', 'Stroud', 'so', 'when', 'she',
'began', 'to', 'stammer', 'something', 'about', 'her', 'poverty', '.', 'I',
'remember', 'getting', 'off', 'a', 'prodigious', 'phrase', 'about', 'the',
'honour', 'being', '_', 'mine', '_', '--', 'oh', ',', 'I', 'was', 'princely',
',', 'my', 'dear', 'Rickham', '!', 'I', 'was', 'posing', 'to', 'myself', 'like',
'one', 'of', 'my', 'own', 'sitters', '.', '"', 'Then', 'I', 'was', 'taken',
'up', 'and', 'left', 'alone', 'with', 'him', '.', 'I', 'had', 'sent', 'all',
'my', 'traps', 'in', 'advance', ',', 'and', 'I', 'had', 'only', 'to', 'set',
'up', 'the', 'easel', 'and', 'get', 'to', 'work', '.', 'He', 'had', 'been',
'dead', 'only', 'twenty-four', 'hours', ',', 'and', 'he', 'died', 'suddenly',
',', 'of', 'heart', 'disease', ',', 'so', 'that', 'there', 'had', 'been', 'no',
'preliminary', 'work', 'of', 'destruction', '--', 'his', 'face', 'was', 'clear',
'and', 'untouched', '.', 'I', 'had', 'met', 'him', 'once', 'or', 'twice', ',',
'years', 'before', ',', 'and', 'thought', 'him', 'insignificant', 'and',
'dingy', '.', 'Now', 'I', 'saw', 'that', 'he', 'was', 'superb', '.', '"', 'I',
'was', 'glad', 'at', 'first', ',', 'with', 'a', 'merely', 'aesthetic',
'satisfaction:', 'glad', 'to', 'have', 'my', 'hand', 'on', 'such', 'a', "'",
'subject', '.', "'", 'Then', 'his', 'strange', 'life-likeness', 'began', 'to',
'affect', 'me', 'queerly', '--', 'as', 'I', 'blocked', 'the', 'head', 'in', 'I',
'felt', 'as', 'if', 'he', 'were', 'watching', 'me', 'do', 'it', '.', 'The',
'sensation', 'was', 'followed', 'by', 'the', 'thought:', 'if', 'he', '_',
'were', '_', 'watching', 'me', ',', 'what', 'would', 'he', 'say', 'to', 'my',
'way', 'of', 'working', '?', 'My', 'strokes', 'began', 'to', 'go', 'a',
'little', 'wild', '--', 'I', 'felt', 'nervous', 'and', 'uncertain', '.', '"',
'Once', ',', 'when', 'I', 'looked', 'up', ',', 'I', 'seemed', 'to', 'see', 'a',
'smile', 'behind', 'his', 'close', 'grayish', 'beard', '--', 'as', 'if', 'he',
'had', 'the', 'secret', ',', 'and', 'were', 'amusing', 'himself', 'by',
'holding', 'it', 'back', 'from', 'me', '.', 'That', 'exasperated', 'me',
'still', 'more', '.', 'The', 'secret', '?', 'Why', ',', 'I', 'had', 'a',
'secret', 'worth', 'twenty', 'of', 'his', '!', 'I', 'dashed', 'at', 'the',
'canvas', 'furiously', ',', 'and', 'tried', 'some', 'of', 'my', 'bravura',
'tricks', '.', 'But', 'they', 'failed', 'me', ',', 'they', 'crumbled', '.', 'I',
'saw', 'that', 'he', 'wasn', "'", 't', 'watching', 'the', 'showy', 'bits', '--',
'I', 'couldn', "'", 't', 'distract', 'his', 'attention;', 'he', 'just', 'kept',
'his', 'eyes', 'on', 'the', 'hard', 'passages', 'between', '.', 'Those', 'were',
'the', 'ones', 'I', 'had', 'always', 'shirked', ',', 'or', 'covered', 'up',
'with', 'some', 'lying', 'paint', '.', 'And', 'how', 'he', 'saw', 'through',
'my', 'lies', '!', '"', 'I', 'looked', 'up', 'again', ',', 'and', 'caught',
'sight', 'of', 'that', 'sketch', 'of', 'the', 'donkey', 'hanging', 'on', 'the',
'wall', 'near', 'his', 'bed', '.', 'His', 'wife', 'told', 'me', 'afterward',
'it', 'was', 'the', 'last', 'thing', 'he', 'had', 'done', '--', 'just', 'a',
'note', 'taken', 'with', 'a', 'shaking', 'hand', ',', 'when', 'he', 'was',
'down', 'in', 'Devonshire', 'recovering', 'from', 'a', 'previous', 'heart',
'attack', '.', 'Just', 'a', 'note', '!', 'But', 'it', 'tells', 'his', 'whole',
'history', '.', 'There', 'are', 'years', 'of', 'patient', 'scornful',
'persistence', 'in', 'every', 'line', '.', 'A', 'man', 'who', 'had', 'swum',

'with', 'the', 'current', 'could', 'never', 'have', 'learned', 'that', 'mighty',
'up-stream', 'stroke', '.', '.', '.', '.', '"', 'I', 'turned', 'back', 'to',
'my', 'work', ',', 'and', 'went', 'on', 'groping', 'and', 'muddling;', 'then',
'I', 'looked', 'at', 'the', 'donkey', 'again', '.', 'I', 'saw', 'that', ',',
'when', 'Stroud', 'laid', 'in', 'the', 'first', 'stroke', ',', 'he', 'knew',
'just', 'what', 'the', 'end', 'would', 'be', '.', 'He', 'had', 'possessed',
'his', 'subject', ',', 'absorbed', 'it', ',', 'recreated', 'it', '.', 'When',
'had', 'I', 'done', 'that', 'with', 'any', 'of', 'my', 'things', '?', 'They',
'hadn', "'", 't', 'been', 'born', 'of', 'me', '--', 'I', 'had', 'just',
'adopted', 'them', '.', '.', '.', '.', '"', 'Hang', 'it', ',', 'Rickham', ',',
'with', 'that', 'face', 'watching', 'me', 'I', 'couldn', "'", 't', 'do',
'another', 'stroke', '.', 'The', 'plain', 'truth', 'was', ',', 'I', 'didn', "'",
't', 'know', 'where', 'to', 'put', 'it', '--', '_', 'I', 'had', 'never',
'known', '_', '.', 'Only', ',', 'with', 'my', 'sitters', 'and', 'my', 'public',
',', 'a', 'showy', 'splash', 'of', 'colour', 'covered', 'up', 'the', 'fact',
'--', 'I', 'just', 'threw', 'paint', 'into', 'their', 'faces', '.', '.', '.',
'.', 'Well', ',', 'paint', 'was', 'the', 'one', 'medium', 'those', 'dead',
'eyes', 'could', 'see', 'through', '--', 'see', 'straight', 'to', 'the',
'tottering', 'foundations', 'underneath', '.', 'Don', "'", 't', 'you', 'know',
'how', ',', 'in', 'talking', 'a', 'foreign', 'language', ',', 'even',
'fluently', ',', 'one', 'says', 'half', 'the', 'time', 'not', 'what', 'one',
'wants', 'to', 'but', 'what', 'one', 'can', '?', 'Well', '--', 'that', 'was',
'the', 'way', 'I', 'painted;', 'and', 'as', 'he', 'lay', 'there', 'and',
'watched', 'me', ',', 'the', 'thing', 'they', 'called', 'my', "'", 'technique',
"'", 'collapsed', 'like', 'a', 'house', 'of', 'cards', '.', 'He', 'didn', "'",
't', 'sneer', ',', 'you', 'understand', ',', 'poor', 'Stroud', '--', 'he',
'just', 'lay', 'there', 'quietly', 'watching', ',', 'and', 'on', 'his', 'lips',
',', 'through', 'the', 'gray', 'beard', ',', 'I', 'seemed', 'to', 'hear', 'the',
'question:', "'", 'Are', 'you', 'sure', 'you', 'know', 'where', 'you', "'",
're', 'coming', 'out', '?', "'", '"', 'If', 'I', 'could', 'have', 'painted',
'that', 'face', ',', 'with', 'that', 'question', 'on', 'it', ',', 'I', 'should',
'have', 'done', 'a', 'great', 'thing', '.', 'The', 'next', 'greatest', 'thing',
'was', 'to', 'see', 'that', 'I', 'couldn', "'", 't', '--', 'and', 'that',
'grace', 'was', 'given', 'me', '.', 'But', ',', 'oh', ',', 'at', 'that',
'minute', ',', 'Rickham', ',', 'was', 'there', 'anything', 'on', 'earth', 'I',
'wouldn', "'", 't', 'have', 'given', 'to', 'have', 'Stroud', 'alive', 'before',
'me', ',', 'and', 'to', 'hear', 'him', 'say:', "'", 'It', "'", 's', 'not',
'too', 'late', '--', 'I', "'", 'll', 'show', 'you', 'how', "'", '?', '"', 'It',
'_', 'was', '_', 'too', 'late', '--', 'it', 'would', 'have', 'been', ',',
'even', 'if', 'he', "'", 'd', 'been', 'alive', '.', 'I', 'packed', 'up', 'my',
'traps', ',', 'and', 'went', 'down', 'and', 'told', 'Mrs', '.', 'Stroud', '.',
'Of', 'course', 'I', 'didn', "'", 't', 'tell', 'her', '_', 'that', '_', '--',
'it', 'would', 'have', 'been', 'Greek', 'to', 'her', '.', 'I', 'simply', 'said',
'I', 'couldn', "'", 't', 'paint', 'him', ',', 'that', 'I', 'was', 'too',
'moved', '.', 'She', 'rather', 'liked', 'the', 'idea', '--', 'she', "'", 's',
'so', 'romantic', '!', 'It', 'was', 'that', 'that', 'made', 'her', 'give', 'me',
'the', 'donkey', '.', 'But', 'she', 'was', 'terribly', 'upset', 'at', 'not',
'getting', 'the', 'portrait', '--', 'she', 'did', 'so', 'want', 'him', "'",

```
'done', "'", 'by', 'some', 'one', 'showy', '!', 'At', 'first', 'I', 'was',
'afraid', 'she', 'wouldn', "'", 't', 'let', 'me', 'off', '--', 'and', 'at',
'my', 'wits', "'", 'end', 'I', 'suggested', 'Grindle', '.', 'Yes', ',', 'it',
'was', 'I', 'who', 'started', 'Grindle:', 'I', 'told', 'Mrs', '.', 'Stroud',
'he', 'was', 'the', "'", 'coming', "'", 'man', ',', 'and', 'she', 'told',
'somebody', 'else', ',', 'and', 'so', 'it', 'got', 'to', 'be', 'true', '.', '.',
'.', '.', 'And', 'he', 'painted', 'Stroud', 'without', 'wincing;', 'and', 'she',
'hung', 'the', 'picture', 'among', 'her', 'husband', "'", 's', 'things', '.',
'.', '.', '.', '"', 'He', 'flung', 'himself', 'down', 'in', 'the', 'arm-chair',
'near', 'mine', ',', 'laid', 'back', 'his', 'head', ',', 'and', 'clasping',
'his', 'arms', 'beneath', 'it', ',', 'looked', 'up', 'at', 'the', 'picture',
'above', 'the', 'chimney-piece', '.', '"', 'I', 'like', 'to', 'fancy', 'that',
'Stroud', 'himself', 'would', 'have', 'given', 'it', 'to', 'me', ',', 'if',
'he', "'", 'd', 'been', 'able', 'to', 'say', 'what', 'he', 'thought', 'that',
'day', '.', '"', 'And', ',', 'in', 'answer', 'to', 'a', 'question', 'I', 'put',
'half-mechanically', '--', '"', 'Begin', 'again', '?', '"', 'he', 'flashed',
'out', '.', '"', 'When', 'the', 'one', 'thing', 'that', 'brings', 'me',
'anywhere', 'near', 'him', 'is', 'that', 'I', 'knew', 'enough', 'to', 'leave',
'off', '?', '"', 'He', 'stood', 'up', 'and', 'laid', 'his', 'hand', 'on', 'my',
'shoulder', 'with', 'a', 'laugh', '.', '"', 'Only', 'the', 'irony', 'of', 'it',
'is', 'that', 'I', '_', 'am', '_', 'still', 'painting', '--', 'since',
'Grindle', "'", 's', 'doing', 'it', 'for', 'me', '!', 'The', 'Strouds', 'stand',
'alone', ',', 'and', 'happen', 'once', '--', 'but', 'there', "'", 's', 'no',
'exterminating', 'our', 'kind', 'of', 'art', '.', '"']
4649
['!', '"', "'", '(', ')', ',', '--', '.', ':', ';', '?', 'A', 'Ah', 'Among',
'And', 'Are', 'Arrt', 'As', 'At', 'Be', 'Begin', 'Burlington', 'But', 'By',
'Carlo', 'Carlo;', 'Chicago', 'Claude', 'Come', 'Croft', 'Destroyed',
'Devonshire', 'Don', 'Dubarry', 'Emperors', 'Florence', 'For', 'Gallery',
'Gideon', 'Gisburn', 'Gisburns', 'Grafton', 'Greek', 'Grindle', 'Grindle:',
'Grindles', 'HAD', 'Had', 'Hang', 'Has', 'He', 'Her', 'Hermia', 'His', 'How',
'I', 'If', 'In', 'It', 'Jack', 'Jove', 'Just', 'Lord', 'Made', 'Miss', 'Money',
'Monte', 'Moon-dancers', 'Mr', 'Mrs', 'My', 'Never', 'No', 'Now', 'Nutley',
'Of', 'Oh', 'On', 'Once', 'Only', 'Or', 'Perhaps', 'Poor', 'Professional',
'Renaissance', 'Rickham', 'Rickham;', 'Riviera', 'Rome', 'Russian', 'Sevres',
'She', 'Stroud', 'Strouds', 'Suddenly', 'That', 'The', 'Then', 'There',
'There:', 'They', 'This', 'Those', 'Though', 'Thwing', 'Thwings', 'To',
'Usually', 'Venetian', 'Victor', 'Was', 'We', 'Well', 'What', 'When', 'Why',
'Yes', 'You', '_', 'a', 'abdication', 'able', 'about', 'about;', 'above',
'abruptly', 'absolute', 'absorbed', 'absurdity', 'academic', 'accuse',
'accustomed', 'across', 'activity', 'add', 'added', 'admirers', 'adopted',
'adulation', 'advance', 'aesthetic', 'affect', 'afraid', 'after', 'afterward',
'again', 'ago', 'ah', 'air', 'alive', 'all', 'almost', 'alone', 'along',
'always', 'am', 'amazement', 'amid', 'among', 'amplest', 'amusing', 'an', 'and',
'another', 'answer', 'answered', 'any', 'anything', 'anywhere', 'apparent',
'apparently', 'appearance', 'appeared', 'appointed', 'are', 'arm', 'arm-chair',
'arm-chairs', 'arms', 'art', 'articles', 'artist', 'as', 'aside', 'asked', 'at',
'atmosphere', 'atom', 'attack', 'attention', 'attention;', 'attitude',
```

'audacities', 'away', 'awful', 'axioms', 'azaleas', 'back', 'background',
'balance', 'balancing', 'balustraded', 'basking', 'bath-rooms', 'be', 'beaming',
'bean-stalk', 'bear', 'beard', 'beauty', 'became', 'because', 'becoming', 'bed',
'been', 'before', 'began', 'begun', 'behind', 'being', 'believed', 'beneath',
'bespoke', 'better', 'better;', 'between', 'big', 'bits', 'bitterness',
'blocked', 'born', 'borne', 'boudoir', 'bravura', 'break', 'breaking',
'breathing', 'bric-a-brac', 'briefly', 'brings', 'bronzes', 'brought', 'brown',
'brush', 'bull', 'business', 'but', 'buying', 'by', 'called', 'came', 'can',
'canvas', 'canvases', 'cards', 'care', 'career', 'caught', 'central', 'chair',
'chap', 'characteristic', 'charming', 'cheap', 'check', 'cheeks', 'chest',
'chimney-piece', 'chucked', 'cigar', 'cigarette', 'cigars', 'circulation',
'circumstance', 'circus-clown', 'claimed', 'clasping', 'clear', 'cleverer',
'close', 'clue', 'coat', 'collapsed', 'colour', 'come', 'comfortable', 'coming',
'companion', 'compared', 'complex', 'confident', 'congesting', 'conjugal',
'constraint', 'consummate', 'contended', 'continued', 'corner', 'corrected',
'could', 'couldn', 'count', 'countenance', 'couple', 'course', 'covered',
'craft', 'cried', 'crossed', 'crowned', 'crumbled', 'cry', 'cured', 'curiosity',
'curious', 'current', 'curtains', 'd', 'dabble', 'damask', 'dark', 'dashed',
'day', 'days', 'dead', 'deadening', 'dear', 'deep', 'deerhound', 'degree',
'delicate', 'demand', 'denied', 'deploring', 'deprecating', 'deprecatingly',
'desire', 'destroyed', 'destruction', 'desultory', 'detail', 'diagnosis', 'did',
'didn', 'died', 'dim', 'dimmest', 'dingy', 'dining-room', 'disarming',
'discovery;', 'discrimination', 'discussion', 'disdain', 'disdained', 'disease',
'disguised', 'display', 'dissatisfied', 'distinguished', 'distract', 'divert',
'do', 'doesn', 'doing', 'domestic', 'don', 'done', 'donkey', 'down', 'dozen',
'dragged', 'drawing-room', 'drawing-rooms', 'drawn', 'dress-closets', 'drew',
'dropped', 'each', 'earth', 'ease', 'easel', 'easy', 'echoed', 'economy',
'effect', 'effects', 'efforts', 'egregious', 'eighteenth-century', 'elbow',
'elegant', 'else', 'embarrassed', 'enabled', 'end', 'endless', 'enjoy',
'enlightenment:', 'enough', 'ensuing', 'equally', 'equanimity', 'escape',
'established', 'etching', 'even', 'event', 'ever', 'everlasting', 'every',
'exasperated', 'except', 'excuse', 'excusing', 'existed', 'expected',
'exquisite', 'exquisitely', 'extenuation', 'exterminating', 'extracting', 'eye',
'eyebrows', 'eyes', 'eyes:', 'face', 'faces', 'fact', 'faded', 'failed',
'failure', 'fair', 'faith', 'false', 'familiar', 'famille-verte', 'fancy',
'fashionable', 'fate', 'feather', 'feet', 'fell', 'fellow', 'felt', 'few',
'fewer', 'finality', 'find', 'fingers', 'first', 'fit', 'fitting', 'five',
'flash', 'flashed', 'florid', 'flowers', 'fluently', 'flung', 'follow',
'followed', 'fond', 'footstep', 'for', 'forced', 'forcing', 'forehead',
'foreign', 'foreseen', 'forgive', 'forgotten', 'form', 'formed', 'forming',
'forward', 'fostered', 'found', 'foundations', 'fragment', 'fragments', 'frame',
'frames', 'frequently', 'friend', 'from', 'full', 'fullest', 'furiously',
'furrowed', 'garlanded', 'garlands', 'gave', 'genial', 'genius', 'gesture',
'get', 'getting', 'give', 'given', 'glad', 'glanced', 'glimpse', 'gloried',
'glory', 'go', 'going', 'gone', 'good', 'good-breeding', 'good-humoured', 'got',
'grace', 'gradually', 'gray', 'grayish', 'great', 'greatest', 'greatness',
'grew', 'groping', 'growing', 'had', 'hadn', 'hair', 'half', 'half-light',
'half-mechanically', 'hall', 'hand', 'hands', 'handsome', 'hanging', 'happen',

'happened', 'hard', 'hardly', 'has', 'have', 'haven', 'having', 'he', 'head',
'hear', 'heard', 'heart', 'height', 'her', 'here', 'here;', 'hermit', 'herself',
'hesitations', 'hide', 'high', 'him', 'him:', 'himself', 'hint', 'his',
'history', 'holding', 'home', 'honour', 'hooded', 'hostess', 'hostess:', 'hot-
house', 'hour', 'hours', 'house', 'how', 'hung', 'husband', 'idea', 'idle',
'idling', 'if', 'immediately', 'in', 'incense', 'indifferent;', 'inevitable',
'inevitably', 'inflexible', 'insensible', 'insignificant', 'instinctively',
'instructive', 'interesting', 'into', 'ironic', 'irony', 'irrelevance',
'irrevocable', 'is', 'it', 'it;', 'its', 'itself', 'jardiniere', 'jealousy',
'just', 'keep', 'kept', 'kind', 'knees', 'knew', 'know', 'know;', 'known',
'laid', 'lair', 'landing', 'language', 'last', 'late', 'later', 'latter',
'laugh', 'laugh:', 'laughed', 'lay', 'leading', 'lean', 'learned', 'least',
'leathery:', 'leave', 'led', 'left', 'leisure', 'lends', 'lent', 'let', 'lies',
'life', 'life-likeness', 'lift', 'lifted', 'light', 'lightly;', 'like', 'liked',
'line', 'lines', 'lingered', 'lips', 'lit', 'little', 'little:', 'live', 'll',
'loathing', 'long', 'longed', 'longer', 'look', 'looked', 'looking', 'lose',
'loss', 'lounging', 'lovely', 'lucky', 'lump', 'luncheon-table', 'luxury',
'lying', 'made', 'make', 'man', 'manage', 'managed', 'mantel-piece', 'marble',
'married', 'may', 'me', 'meant', 'mediocrity', 'medium', 'mentioned', 'mere',
'merely', 'met', 'might', 'mighty', 'millionaire', 'mine', 'mine:', 'minute',
'minutes', 'mirrors', 'modest', 'modesty', 'moment', 'money', 'monumental',
'mood', 'morbidly', 'more', 'most', 'mourn', 'mourned', 'moustache', 'moved',
'much', 'muddling;', 'multiplied', 'murmur', 'muscles', 'must', 'my', 'myself',
'mysterious', 'naive', 'near', 'nearly', 'negatived', 'nervous', 'nervousness;',
'neutral', 'never', 'next', 'no', 'none', 'not', 'note', 'nothing', 'now',
'nymphs', 'oak', 'obituary', 'object', 'objects', 'occurred', 'oddly', 'of',
'off', 'often', 'oh', 'old', 'on', 'once', 'one', 'ones', 'only', 'onto',
'open', 'or', 'other', 'our', 'ourselves', 'out', 'out:', 'outline', 'oval',
'over', 'own', 'packed', 'paid', 'paint', 'painted', 'painted;', 'painter',
'painting', 'painting;', 'pale', 'paled', 'palm-trees;', 'panel', 'panelling',
'pardonable', 'pardoned', 'part', 'passages', 'passing', 'past', 'pastels',
'pathos', 'patient', 'people', 'perceptible', 'perfect', 'persistence',
'persuasively', 'phrase', 'picture', 'pictures', 'pines', 'pink', 'place',
'placed', 'plain', 'platitudes', 'pleased', 'pockets', 'point', 'poised',
'poor', 'portrait', 'posing', 'possessed', 'poverty', 'predicted',
'preliminary', 'presenting', 'prestidigitation', 'pretty', 'previous', 'price',
'pride', 'pride:', 'princely', 'prism', 'problem', 'proclaiming', 'prodigious',
'profusion', 'protest', 'prove', 'public', 'purblind', 'purely', 'pushed',
'put', 'qualities', 'quality', 'queerly', 'question', 'question:', 'quickly',
'quietly', 'quite', 'quote', 'rain', 'raised', 'random', 'rather', 're', 'real',
'really', 'reared', 'reason', 'reassurance', 'recovering', 'recreated',
'reflected', 'reflection', 'regrets', 'relatively', 'remained', 'remember',
'reminded', 'repeating', 'represented', 'reproduction', 'resented', 'resolve',
'resources', 'rest', 'rich', 'rich;', 'ridiculous', 'robbed', 'romantic',
'room', 'rose', 'rs', 'rule', 'run', 's', 'said', 'said:', 'same',
'satisfaction', 'satisfaction:', 'savour', 'saw', 'say', 'say:', 'saying',
'says', 'scorn', 'scornful', 'secret', 'see', 'seemed', 'seen', 'self-
confident', 'send', 'sensation', 'sensitive', 'sent', 'serious', 'set', 'sex',

```
'shade', 'shaking', 'shall', 'she', 'shirked', 'short', 'should', 'shoulder',
'shoulders', 'show', 'showed', 'showy', 'shrug', 'shrugged', 'sight', 'sign',
'silent;', 'silver', 'similar', 'simpleton', 'simplifications', 'simply',
'since', 'single', 'sitter', 'sitters', 'sketch', 'skill', 'slight', 'slightly',
'slowly:', 'small', 'smile', 'smiling', 'sneer', 'so', 'solace', 'some',
'somebody', 'something', 'spacious', 'spaniel', 'speaking-tubes',
'speculations;', 'spite', 'splash', 'square', 'stairs', 'stammer', 'stand',
'standing', 'started', 'stay', 'still', 'stocked', 'stood', 'stopped',
'stopping', 'straddling', 'straight', 'strain', 'straining', 'strange', 'straw',
'stream', 'stroke', 'strokes', 'strolled', 'strongest', 'strongly', 'struck',
'studio', 'stuff', 'subject', 'substantial', 'suburban', 'such', 'suddenly',
'suffered', 'sugar', 'suggested', 'sunburn', 'sunburnt', 'sunlit', 'superb',
'sure', 'surest', 'surface', 'surprise', 'surprised', 'surrounded', 'suspected',
'sweetly', 'sweetness', 'swelling', 'swept', 'swum', 't', 'table', 'take',
'taken', 'talking', 'tea', 'tears', 'technicalities', 'technique', 'tell',
'tells', 'tempting', 'terra-cotta', 'terrace', 'terraces', 'terribly', 'than',
'that', 'the', 'their', 'them', 'then', 'there', 'therefore', 'they', 'thin',
'thing', 'things', 'think', 'this', 'thither', 'those', 'though', 'thought',
'thought:', 'three', 'threshold', 'threw', 'through', 'throwing', 'tie', 'till',
'time', 'timorously', 'tinge', 'tips', 'tired', 'to', 'told', 'tone', 'tones',
'too', 'took', 'tottering', 'touched', 'toward', 'trace', 'trade', 'transmute',
'traps', 'travelled', 'tribute', 'tributes', 'tricks', 'tried', 'trouser-
presses', 'true', 'truth', 'turned', 'twenty', 'twenty-four', 'twice',
'twirling', 'unaccountable', 'uncertain', 'under', 'underlay', 'underneath',
'understand', 'unexpected', 'untouched', 'unusual', 'up', 'up-stream', 'up;',
'upon', 'upset', 'upstairs', 'us', 'used', 'usual', 'value', 'varnishing',
'vases', 've', 'veins', 'velveteen', 'very', 'villa', 'vindicated',
'virtuosity', 'vista', 'vocation', 'voice', 'wall', 'wander', 'want', 'wanted',
'wants', 'was', 'wasn', 'watched', 'watching', 'water-colour', 'waves', 'way',
'weekly', 'weeks', 'welcome', 'went', 'were', 'what', 'when', 'whenever',
'where', 'which', 'while', 'white', 'white-panelled', 'who', 'whole', 'whom',
'why', 'wide', 'widow', 'wife', 'wild', 'wincing', 'wincing;', 'window-
curtains', 'wish', 'with', 'without', 'wits', 'woman', 'women', 'women:', 'won',
'wonder', 'wondered', 'word', 'work', 'working', 'worth', 'would', 'wouldn',
'year', 'years', 'yellow', 'yet', 'you', 'younger', 'your', 'yourself']
1159
```

[66]:
```
vocab = {token:integer for integer,token in enumerate(all_words)} # Give Id for
  ↪each unique tokens.
print(vocab)
```

```
{'!': 0, '"': 1, "'": 2, '(': 3, ')': 4, ',': 5, '--': 6, '.': 7, ':': 8, ';':
9, '?': 10, 'A': 11, 'Ah': 12, 'Among': 13, 'And': 14, 'Are': 15, 'Arrt': 16,
'As': 17, 'At': 18, 'Be': 19, 'Begin': 20, 'Burlington': 21, 'But': 22, 'By':
23, 'Carlo': 24, 'Carlo;': 25, 'Chicago': 26, 'Claude': 27, 'Come': 28, 'Croft':
29, 'Destroyed': 30, 'Devonshire': 31, 'Don': 32, 'Dubarry': 33, 'Emperors': 34,
'Florence': 35, 'For': 36, 'Gallery': 37, 'Gideon': 38, 'Gisburn': 39,
'Gisburns': 40, 'Grafton': 41, 'Greek': 42, 'Grindle': 43, 'Grindle:': 44,
```

'Grindles': 45, 'HAD': 46, 'Had': 47, 'Hang': 48, 'Has': 49, 'He': 50, 'Her':
51, 'Hermia': 52, 'His': 53, 'How': 54, 'I': 55, 'If': 56, 'In': 57, 'It': 58,
'Jack': 59, 'Jove': 60, 'Just': 61, 'Lord': 62, 'Made': 63, 'Miss': 64, 'Money':
65, 'Monte': 66, 'Moon-dancers': 67, 'Mr': 68, 'Mrs': 69, 'My': 70, 'Never': 71,
'No': 72, 'Now': 73, 'Nutley': 74, 'Of': 75, 'Oh': 76, 'On': 77, 'Once': 78,
'Only': 79, 'Or': 80, 'Perhaps': 81, 'Poor': 82, 'Professional': 83,
'Renaissance': 84, 'Rickham': 85, 'Rickham;': 86, 'Riviera': 87, 'Rome': 88,
'Russian': 89, 'Sevres': 90, 'She': 91, 'Stroud': 92, 'Strouds': 93, 'Suddenly':
94, 'That': 95, 'The': 96, 'Then': 97, 'There': 98, 'There:': 99, 'They': 100,
'This': 101, 'Those': 102, 'Though': 103, 'Thwing': 104, 'Thwings': 105, 'To':
106, 'Usually': 107, 'Venetian': 108, 'Victor': 109, 'Was': 110, 'We': 111,
'Well': 112, 'What': 113, 'When': 114, 'Why': 115, 'Yes': 116, 'You': 117, '_':
118, 'a': 119, 'abdication': 120, 'able': 121, 'about': 122, 'about;': 123,
'above': 124, 'abruptly': 125, 'absolute': 126, 'absorbed': 127, 'absurdity':
128, 'academic': 129, 'accuse': 130, 'accustomed': 131, 'across': 132,
'activity': 133, 'add': 134, 'added': 135, 'admirers': 136, 'adopted': 137,
'adulation': 138, 'advance': 139, 'aesthetic': 140, 'affect': 141, 'afraid':
142, 'after': 143, 'afterward': 144, 'again': 145, 'ago': 146, 'ah': 147, 'air':
148, 'alive': 149, 'all': 150, 'almost': 151, 'alone': 152, 'along': 153,
'always': 154, 'am': 155, 'amazement': 156, 'amid': 157, 'among': 158,
'amplest': 159, 'amusing': 160, 'an': 161, 'and': 162, 'another': 163, 'answer':
164, 'answered': 165, 'any': 166, 'anything': 167, 'anywhere': 168, 'apparent':
169, 'apparently': 170, 'appearance': 171, 'appeared': 172, 'appointed': 173,
'are': 174, 'arm': 175, 'arm-chair': 176, 'arm-chairs': 177, 'arms': 178, 'art':
179, 'articles': 180, 'artist': 181, 'as': 182, 'aside': 183, 'asked': 184,
'at': 185, 'atmosphere': 186, 'atom': 187, 'attack': 188, 'attention': 189,
'attention;': 190, 'attitude': 191, 'audacities': 192, 'away': 193, 'awful':
194, 'axioms': 195, 'azaleas': 196, 'back': 197, 'background': 198, 'balance':
199, 'balancing': 200, 'balustraded': 201, 'basking': 202, 'bath-rooms': 203,
'be': 204, 'beaming': 205, 'bean-stalk': 206, 'bear': 207, 'beard': 208,
'beauty': 209, 'became': 210, 'because': 211, 'becoming': 212, 'bed': 213,
'been': 214, 'before': 215, 'began': 216, 'begun': 217, 'behind': 218, 'being':
219, 'believed': 220, 'beneath': 221, 'bespoke': 222, 'better': 223, 'better;':
224, 'between': 225, 'big': 226, 'bits': 227, 'bitterness': 228, 'blocked': 229,
'born': 230, 'borne': 231, 'boudoir': 232, 'bravura': 233, 'break': 234,
'breaking': 235, 'breathing': 236, 'bric-a-brac': 237, 'briefly': 238, 'brings':
239, 'bronzes': 240, 'brought': 241, 'brown': 242, 'brush': 243, 'bull': 244,
'business': 245, 'but': 246, 'buying': 247, 'by': 248, 'called': 249, 'came':
250, 'can': 251, 'canvas': 252, 'canvases': 253, 'cards': 254, 'care': 255,
'career': 256, 'caught': 257, 'central': 258, 'chair': 259, 'chap': 260,
'characteristic': 261, 'charming': 262, 'cheap': 263, 'check': 264, 'cheeks':
265, 'chest': 266, 'chimney-piece': 267, 'chucked': 268, 'cigar': 269,
'cigarette': 270, 'cigars': 271, 'circulation': 272, 'circumstance': 273,
'circus-clown': 274, 'claimed': 275, 'clasping': 276, 'clear': 277, 'cleverer':
278, 'close': 279, 'clue': 280, 'coat': 281, 'collapsed': 282, 'colour': 283,
'come': 284, 'comfortable': 285, 'coming': 286, 'companion': 287, 'compared':
288, 'complex': 289, 'confident': 290, 'congesting': 291, 'conjugal': 292,
'constraint': 293, 'consummate': 294, 'contended': 295, 'continued': 296,

'corner': 297, 'corrected': 298, 'could': 299, 'couldn': 300, 'count': 301,
'countenance': 302, 'couple': 303, 'course': 304, 'covered': 305, 'craft': 306,
'cried': 307, 'crossed': 308, 'crowned': 309, 'crumbled': 310, 'cry': 311,
'cured': 312, 'curiosity': 313, 'curious': 314, 'current': 315, 'curtains': 316,
'd': 317, 'dabble': 318, 'damask': 319, 'dark': 320, 'dashed': 321, 'day': 322,
'days': 323, 'dead': 324, 'deadening': 325, 'dear': 326, 'deep': 327,
'deerhound': 328, 'degree': 329, 'delicate': 330, 'demand': 331, 'denied': 332,
'deploring': 333, 'deprecating': 334, 'deprecatingly': 335, 'desire': 336,
'destroyed': 337, 'destruction': 338, 'desultory': 339, 'detail': 340,
'diagnosis': 341, 'did': 342, 'didn': 343, 'died': 344, 'dim': 345, 'dimmest':
346, 'dingy': 347, 'dining-room': 348, 'disarming': 349, 'discovery;': 350,
'discrimination': 351, 'discussion': 352, 'disdain': 353, 'disdained': 354,
'disease': 355, 'disguised': 356, 'display': 357, 'dissatisfied': 358,
'distinguished': 359, 'distract': 360, 'divert': 361, 'do': 362, 'doesn': 363,
'doing': 364, 'domestic': 365, 'don': 366, 'done': 367, 'donkey': 368, 'down':
369, 'dozen': 370, 'dragged': 371, 'drawing-room': 372, 'drawing-rooms': 373,
'drawn': 374, 'dress-closets': 375, 'drew': 376, 'dropped': 377, 'each': 378,
'earth': 379, 'ease': 380, 'easel': 381, 'easy': 382, 'echoed': 383, 'economy':
384, 'effect': 385, 'effects': 386, 'efforts': 387, 'egregious': 388,
'eighteenth-century': 389, 'elbow': 390, 'elegant': 391, 'else': 392,
'embarrassed': 393, 'enabled': 394, 'end': 395, 'endless': 396, 'enjoy': 397,
'enlightenment:': 398, 'enough': 399, 'ensuing': 400, 'equally': 401,
'equanimity': 402, 'escape': 403, 'established': 404, 'etching': 405, 'even':
406, 'event': 407, 'ever': 408, 'everlasting': 409, 'every': 410, 'exasperated':
411, 'except': 412, 'excuse': 413, 'excusing': 414, 'existed': 415, 'expected':
416, 'exquisite': 417, 'exquisitely': 418, 'extenuation': 419, 'exterminating':
420, 'extracting': 421, 'eye': 422, 'eyebrows': 423, 'eyes': 424, 'eyes:': 425,
'face': 426, 'faces': 427, 'fact': 428, 'faded': 429, 'failed': 430, 'failure':
431, 'fair': 432, 'faith': 433, 'false': 434, 'familiar': 435, 'famille-verte':
436, 'fancy': 437, 'fashionable': 438, 'fate': 439, 'feather': 440, 'feet': 441,
'fell': 442, 'fellow': 443, 'felt': 444, 'few': 445, 'fewer': 446, 'finality':
447, 'find': 448, 'fingers': 449, 'first': 450, 'fit': 451, 'fitting': 452,
'five': 453, 'flash': 454, 'flashed': 455, 'florid': 456, 'flowers': 457,
'fluently': 458, 'flung': 459, 'follow': 460, 'followed': 461, 'fond': 462,
'footstep': 463, 'for': 464, 'forced': 465, 'forcing': 466, 'forehead': 467,
'foreign': 468, 'foreseen': 469, 'forgive': 470, 'forgotten': 471, 'form': 472,
'formed': 473, 'forming': 474, 'forward': 475, 'fostered': 476, 'found': 477,
'foundations': 478, 'fragment': 479, 'fragments': 480, 'frame': 481, 'frames':
482, 'frequently': 483, 'friend': 484, 'from': 485, 'full': 486, 'fullest': 487,
'furiously': 488, 'furrowed': 489, 'garlanded': 490, 'garlands': 491, 'gave':
492, 'genial': 493, 'genius': 494, 'gesture': 495, 'get': 496, 'getting': 497,
'give': 498, 'given': 499, 'glad': 500, 'glanced': 501, 'glimpse': 502,
'gloried': 503, 'glory': 504, 'go': 505, 'going': 506, 'gone': 507, 'good': 508,
'good-breeding': 509, 'good-humoured': 510, 'got': 511, 'grace': 512,
'gradually': 513, 'gray': 514, 'grayish': 515, 'great': 516, 'greatest': 517,
'greatness': 518, 'grew': 519, 'groping': 520, 'growing': 521, 'had': 522,
'hadn': 523, 'hair': 524, 'half': 525, 'half-light': 526, 'half-mechanically':
527, 'hall': 528, 'hand': 529, 'hands': 530, 'handsome': 531, 'hanging': 532,

'happen': 533, 'happened': 534, 'hard': 535, 'hardly': 536, 'has': 537, 'have': 538, 'haven': 539, 'having': 540, 'he': 541, 'head': 542, 'hear': 543, 'heard': 544, 'heart': 545, 'height': 546, 'her': 547, 'here': 548, 'here;': 549, 'hermit': 550, 'herself': 551, 'hesitations': 552, 'hide': 553, 'high': 554, 'him': 555, 'him:': 556, 'himself': 557, 'hint': 558, 'his': 559, 'history': 560, 'holding': 561, 'home': 562, 'honour': 563, 'hooded': 564, 'hostess': 565, 'hostess:': 566, 'hot-house': 567, 'hour': 568, 'hours': 569, 'house': 570, 'how': 571, 'hung': 572, 'husband': 573, 'idea': 574, 'idle': 575, 'idling': 576, 'if': 577, 'immediately': 578, 'in': 579, 'incense': 580, 'indifferent;': 581, 'inevitable': 582, 'inevitably': 583, 'inflexible': 584, 'insensible': 585, 'insignificant': 586, 'instinctively': 587, 'instructive': 588, 'interesting': 589, 'into': 590, 'ironic': 591, 'irony': 592, 'irrelevance': 593, 'irrevocable': 594, 'is': 595, 'it': 596, 'it;': 597, 'its': 598, 'itself': 599, 'jardiniere': 600, 'jealousy': 601, 'just': 602, 'keep': 603, 'kept': 604, 'kind': 605, 'knees': 606, 'knew': 607, 'know': 608, 'know;': 609, 'known': 610, 'laid': 611, 'lair': 612, 'landing': 613, 'language': 614, 'last': 615, 'late': 616, 'later': 617, 'latter': 618, 'laugh': 619, 'laugh:': 620, 'laughed': 621, 'lay': 622, 'leading': 623, 'lean': 624, 'learned': 625, 'least': 626, 'leathery:': 627, 'leave': 628, 'led': 629, 'left': 630, 'leisure': 631, 'lends': 632, 'lent': 633, 'let': 634, 'lies': 635, 'life': 636, 'life-likeness': 637, 'lift': 638, 'lifted': 639, 'light': 640, 'lightly;': 641, 'like': 642, 'liked': 643, 'line': 644, 'lines': 645, 'lingered': 646, 'lips': 647, 'lit': 648, 'little': 649, 'little:': 650, 'live': 651, 'll': 652, 'loathing': 653, 'long': 654, 'longed': 655, 'longer': 656, 'look': 657, 'looked': 658, 'looking': 659, 'lose': 660, 'loss': 661, 'lounging': 662, 'lovely': 663, 'lucky': 664, 'lump': 665, 'luncheon-table': 666, 'luxury': 667, 'lying': 668, 'made': 669, 'make': 670, 'man': 671, 'manage': 672, 'managed': 673, 'mantel-piece': 674, 'marble': 675, 'married': 676, 'may': 677, 'me': 678, 'meant': 679, 'mediocrity': 680, 'medium': 681, 'mentioned': 682, 'mere': 683, 'merely': 684, 'met': 685, 'might': 686, 'mighty': 687, 'millionaire': 688, 'mine': 689, 'mine:': 690, 'minute': 691, 'minutes': 692, 'mirrors': 693, 'modest': 694, 'modesty': 695, 'moment': 696, 'money': 697, 'monumental': 698, 'mood': 699, 'morbidly': 700, 'more': 701, 'most': 702, 'mourn': 703, 'mourned': 704, 'moustache': 705, 'moved': 706, 'much': 707, 'muddling;': 708, 'multiplied': 709, 'murmur': 710, 'muscles': 711, 'must': 712, 'my': 713, 'myself': 714, 'mysterious': 715, 'naive': 716, 'near': 717, 'nearly': 718, 'negatived': 719, 'nervous': 720, 'nervousness;': 721, 'neutral': 722, 'never': 723, 'next': 724, 'no': 725, 'none': 726, 'not': 727, 'note': 728, 'nothing': 729, 'now': 730, 'nymphs': 731, 'oak': 732, 'obituary': 733, 'object': 734, 'objects': 735, 'occurred': 736, 'oddly': 737, 'of': 738, 'off': 739, 'often': 740, 'oh': 741, 'old': 742, 'on': 743, 'once': 744, 'one': 745, 'ones': 746, 'only': 747, 'onto': 748, 'open': 749, 'or': 750, 'other': 751, 'our': 752, 'ourselves': 753, 'out': 754, 'out:': 755, 'outline': 756, 'oval': 757, 'over': 758, 'own': 759, 'packed': 760, 'paid': 761, 'paint': 762, 'painted': 763, 'painted;': 764, 'painter': 765, 'painting': 766, 'painting;': 767, 'pale': 768, 'paled': 769, 'palm-trees;': 770, 'panel': 771, 'panelling': 772, 'pardonable': 773, 'pardoned': 774, 'part': 775, 'passages': 776, 'passing': 777, 'past': 778, 'pastels': 779, 'pathos': 780, 'patient': 781, 'people': 782, 'perceptible':

783, 'perfect': 784, 'persistence': 785, 'persuasively': 786, 'phrase': 787,
'picture': 788, 'pictures': 789, 'pines': 790, 'pink': 791, 'place': 792,
'placed': 793, 'plain': 794, 'platitudes': 795, 'pleased': 796, 'pockets': 797,
'point': 798, 'poised': 799, 'poor': 800, 'portrait': 801, 'posing': 802,
'possessed': 803, 'poverty': 804, 'predicted': 805, 'preliminary': 806,
'presenting': 807, 'prestidigitation': 808, 'pretty': 809, 'previous': 810,
'price': 811, 'pride': 812, 'pride:': 813, 'princely': 814, 'prism': 815,
'problem': 816, 'proclaiming': 817, 'prodigious': 818, 'profusion': 819,
'protest': 820, 'prove': 821, 'public': 822, 'purblind': 823, 'purely': 824,
'pushed': 825, 'put': 826, 'qualities': 827, 'quality': 828, 'queerly': 829,
'question': 830, 'question:': 831, 'quickly': 832, 'quietly': 833, 'quite': 834,
'quote': 835, 'rain': 836, 'raised': 837, 'random': 838, 'rather': 839, 're':
840, 'real': 841, 'really': 842, 'reared': 843, 'reason': 844, 'reassurance':
845, 'recovering': 846, 'recreated': 847, 'reflected': 848, 'reflection': 849,
'regrets': 850, 'relatively': 851, 'remained': 852, 'remember': 853, 'reminded':
854, 'repeating': 855, 'represented': 856, 'reproduction': 857, 'resented': 858,
'resolve': 859, 'resources': 860, 'rest': 861, 'rich': 862, 'rich;': 863,
'ridiculous': 864, 'robbed': 865, 'romantic': 866, 'room': 867, 'rose': 868,
'rs': 869, 'rule': 870, 'run': 871, 's': 872, 'said': 873, 'said:': 874, 'same':
875, 'satisfaction': 876, 'satisfaction:': 877, 'savour': 878, 'saw': 879,
'say': 880, 'say:': 881, 'saying': 882, 'says': 883, 'scorn': 884, 'scornful':
885, 'secret': 886, 'see': 887, 'seemed': 888, 'seen': 889, 'self-confident':
890, 'send': 891, 'sensation': 892, 'sensitive': 893, 'sent': 894, 'serious':
895, 'set': 896, 'sex': 897, 'shade': 898, 'shaking': 899, 'shall': 900, 'she':
901, 'shirked': 902, 'short': 903, 'should': 904, 'shoulder': 905, 'shoulders':
906, 'show': 907, 'showed': 908, 'showy': 909, 'shrug': 910, 'shrugged': 911,
'sight': 912, 'sign': 913, 'silent;': 914, 'silver': 915, 'similar': 916,
'simpleton': 917, 'simplifications': 918, 'simply': 919, 'since': 920, 'single':
921, 'sitter': 922, 'sitters': 923, 'sketch': 924, 'skill': 925, 'slight': 926,
'slightly': 927, 'slowly:': 928, 'small': 929, 'smile': 930, 'smiling': 931,
'sneer': 932, 'so': 933, 'solace': 934, 'some': 935, 'somebody': 936,
'something': 937, 'spacious': 938, 'spaniel': 939, 'speaking-tubes': 940,
'speculations;': 941, 'spite': 942, 'splash': 943, 'square': 944, 'stairs': 945,
'stammer': 946, 'stand': 947, 'standing': 948, 'started': 949, 'stay': 950,
'still': 951, 'stocked': 952, 'stood': 953, 'stopped': 954, 'stopping': 955,
'straddling': 956, 'straight': 957, 'strain': 958, 'straining': 959, 'strange':
960, 'straw': 961, 'stream': 962, 'stroke': 963, 'strokes': 964, 'strolled':
965, 'strongest': 966, 'strongly': 967, 'struck': 968, 'studio': 969, 'stuff':
970, 'subject': 971, 'substantial': 972, 'suburban': 973, 'such': 974,
'suddenly': 975, 'suffered': 976, 'sugar': 977, 'suggested': 978, 'sunburn':
979, 'sunburnt': 980, 'sunlit': 981, 'superb': 982, 'sure': 983, 'surest': 984,
'surface': 985, 'surprise': 986, 'surprised': 987, 'surrounded': 988,
'suspected': 989, 'sweetly': 990, 'sweetness': 991, 'swelling': 992, 'swept':
993, 'swum': 994, 't': 995, 'table': 996, 'take': 997, 'taken': 998, 'talking':
999, 'tea': 1000, 'tears': 1001, 'technicalities': 1002, 'technique': 1003,
'tell': 1004, 'tells': 1005, 'tempting': 1006, 'terra-cotta': 1007, 'terrace':
1008, 'terraces': 1009, 'terribly': 1010, 'than': 1011, 'that': 1012, 'the':
1013, 'their': 1014, 'them': 1015, 'then': 1016, 'there': 1017, 'therefore':

```
1018, 'they': 1019, 'thin': 1020, 'thing': 1021, 'things': 1022, 'think': 1023,
'this': 1024, 'thither': 1025, 'those': 1026, 'though': 1027, 'thought': 1028,
'thought:': 1029, 'three': 1030, 'threshold': 1031, 'threw': 1032, 'through':
1033, 'throwing': 1034, 'tie': 1035, 'till': 1036, 'time': 1037, 'timorously':
1038, 'tinge': 1039, 'tips': 1040, 'tired': 1041, 'to': 1042, 'told': 1043,
'tone': 1044, 'tones': 1045, 'too': 1046, 'took': 1047, 'tottering': 1048,
'touched': 1049, 'toward': 1050, 'trace': 1051, 'trade': 1052, 'transmute':
1053, 'traps': 1054, 'travelled': 1055, 'tribute': 1056, 'tributes': 1057,
'tricks': 1058, 'tried': 1059, 'trouser-presses': 1060, 'true': 1061, 'truth':
1062, 'turned': 1063, 'twenty': 1064, 'twenty-four': 1065, 'twice': 1066,
'twirling': 1067, 'unaccountable': 1068, 'uncertain': 1069, 'under': 1070,
'underlay': 1071, 'underneath': 1072, 'understand': 1073, 'unexpected': 1074,
'untouched': 1075, 'unusual': 1076, 'up': 1077, 'up-stream': 1078, 'up;': 1079,
'upon': 1080, 'upset': 1081, 'upstairs': 1082, 'us': 1083, 'used': 1084,
'usual': 1085, 'value': 1086, 'varnishing': 1087, 'vases': 1088, 've': 1089,
'veins': 1090, 'velveteen': 1091, 'very': 1092, 'villa': 1093, 'vindicated':
1094, 'virtuosity': 1095, 'vista': 1096, 'vocation': 1097, 'voice': 1098,
'wall': 1099, 'wander': 1100, 'want': 1101, 'wanted': 1102, 'wants': 1103,
'was': 1104, 'wasn': 1105, 'watched': 1106, 'watching': 1107, 'water-colour':
1108, 'waves': 1109, 'way': 1110, 'weekly': 1111, 'weeks': 1112, 'welcome':
1113, 'went': 1114, 'were': 1115, 'what': 1116, 'when': 1117, 'whenever': 1118,
'where': 1119, 'which': 1120, 'while': 1121, 'white': 1122, 'white-panelled':
1123, 'who': 1124, 'whole': 1125, 'whom': 1126, 'why': 1127, 'wide': 1128,
'widow': 1129, 'wife': 1130, 'wild': 1131, 'wincing': 1132, 'wincing;': 1133,
'window-curtains': 1134, 'wish': 1135, 'with': 1136, 'without': 1137, 'wits':
1138, 'woman': 1139, 'women': 1140, 'women:': 1141, 'won': 1142, 'wonder': 1143,
'wondered': 1144, 'word': 1145, 'work': 1146, 'working': 1147, 'worth': 1148,
'would': 1149, 'wouldn': 1150, 'year': 1151, 'years': 1152, 'yellow': 1153,
'yet': 1154, 'you': 1155, 'younger': 1156, 'your': 1157, 'yourself': 1158}
```

- Below are the first 50 entries in this vocabulary:

[69]:
```
#Showing First 50 unique tokens with their id


for i, item in enumerate(vocab.items()):
    print(item)
    if i >= 50: break
```

```
('!', 0)
('"', 1)
("'", 2)
('(', 3)
(')', 4)
(',', 5)
('--', 6)
('.', 7)
(':', 8)
```

```
(';', 9)
('?', 10)
('A', 11)
('Ah', 12)
('Among', 13)
('And', 14)
('Are', 15)
('Arrt', 16)
('As', 17)
('At', 18)
('Be', 19)
('Begin', 20)
('Burlington', 21)
('But', 22)
('By', 23)
('Carlo', 24)
('Carlo;', 25)
('Chicago', 26)
('Claude', 27)
('Come', 28)
('Croft', 29)
('Destroyed', 30)
('Devonshire', 31)
('Don', 32)
('Dubarry', 33)
('Emperors', 34)
('Florence', 35)
('For', 36)
('Gallery', 37)
('Gideon', 38)
('Gisburn', 39)
('Gisburns', 40)
('Grafton', 41)
('Greek', 42)
('Grindle', 43)
('Grindle:', 44)
('Grindles', 45)
('HAD', 46)
('Had', 47)
('Hang', 48)
('Has', 49)
('He', 50)
```

- Below, we illustrate the tokenization of a short sample text using a small vocabulary:
- Putting it now all together into a tokenizer class

[72]:
```
class SimpleTokenizerV1:
    def __init__(self, vocab):
        self.str_to_int = vocab
        self.int_to_str = {i:s for s,i in vocab.items()}

    def encode(self, text):
        preprocessed = re.split(r'([,.?_!"()\']|--|\s)', text)
        preprocessed = [item.strip() for item in preprocessed if item.strip()]
        ids = [self.str_to_int[s] for s in preprocessed]
        return ids

    def decode(self, ids):
        text = " ".join([self.int_to_str[i] for i in ids])
        # Replace spaces before the specified punctuations
        text = re.sub(r'\s+([,.?!"()\'])', r'\1', text)
        return text
```

- The `encode` function turns text into token IDs
- The `decode` function turns token IDs back into text

- We can use the tokenizer to encode (that is, tokenize) texts into integers
- These integers can then be embedded (later) as input of/for the LLM

[33]:
```
tokenizer = SimpleTokenizerV1(vocab)

text = """It's the last he painted, you know," Mrs. Gisburn said with␣
 ↪pardonable pride."""
ids = tokenizer.encode(text)
print(ids)
```

```
[1, 58, 2, 872, 1013, 615, 541, 763, 5, 1155, 608, 5, 1, 69, 7, 39, 873, 1136,
773, 812, 7]
```

- We can decode the integers back into text

[34]:
```
tokenizer.decode(ids)
```

[34]:
```
'" It\' s the last he painted, you know," Mrs. Gisburn said with pardonable
pride.'
```

[35]:
```
tokenizer.decode(tokenizer.encode(text))
```

[35]:
```
'" It\' s the last he painted, you know," Mrs. Gisburn said with pardonable
pride.'
```

1.4 2.4 Adding special context tokens

- It's useful to add some "special" tokens for unknown words and to denote the end of a text

- Some tokenizers use special tokens to help the LLM with additional context
- Some of these special tokens are
  - `[BOS]` (beginning of sequence) marks the beginning of text
  - `[EOS]` (end of sequence) marks where the text ends (this is usually used to concatenate multiple unrelated texts, e.g., two different Wikipedia articles or two different books, and so on)
  - `[PAD]` (padding) if we train LLMs with a batch size greater than 1 (we may include multiple texts with different lengths; with the padding token we pad the shorter texts to the longest length so that all texts have an equal length)
- `[UNK]` to represent works that are not included in the vocabulary
- Note that GPT-2 does not need any of these tokens mentioned above but only uses an `<|endoftext|>` token to reduce complexity
- The `<|endoftext|>` is analogous to the `[EOS]` token mentioned above
- GPT also uses the `<|endoftext|>` for padding (since we typically use a mask when training on batched inputs, we would not attend padded tokens anyways, so it does not matter what these tokens are)
- GPT-2 does not use an `<UNK>` token for out-of-vocabulary words; instead, GPT-2 uses a byte-pair encoding (BPE) tokenizer, which breaks down words into subword units which we will discuss in a later section
- We use the `<|endoftext|>` tokens between two independent sources of text:
- Let's see what happens if we tokenize the following text:

[36]:
```
tokenizer = SimpleTokenizerV1(vocab)

text = "Hello, do you like tea. Is this-- a test?"

tokenizer.encode(text)
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
Cell In[36], line 5
      1 tokenizer = SimpleTokenizerV1(vocab)
      3 text = "Hello, do you like tea. Is this-- a test?"
----> 5 tokenizer.encode(text)

Cell In[32], line 9, in SimpleTokenizerV1.encode(self, text)
      7 preprocessed = re.split(r'([,.?_!"()\']|--|\s)', text)
      8 preprocessed = [item.strip() for item in preprocessed if item.strip()]
----> 9 ids = [self.str_to_int[s] for s in preprocessed]
     10 return ids

Cell In[32], line 9, in <listcomp>(.0)
      7 preprocessed = re.split(r'([,.?_!"()\']|--|\s)', text)
```

```
      8 preprocessed = [item.strip() for item in preprocessed if item.strip()]
----> 9 ids = [self.str_to_int[s] for s in preprocessed]
     10 return ids

KeyError: 'Hello'
```

- The above produces an error because the word "Hello" is not contained in the vocabulary
- To deal with such cases, we can add special tokens like `"<|unk|>"` to the vocabulary to represent unknown words
- Since we are already extending the vocabulary, let's add another token called `"<|endoftext|>"` which is used in GPT-2 training to denote the end of a text (and it's also used between concatenated text, like if our training datasets consists of multiple articles, books, etc.)

[ ]:
```
preprocessed = re.split(r'([,.?_!"()\']|--|\s)', raw_text)
preprocessed = [item.strip() for item in preprocessed if item.strip()]

all_tokens = sorted(list(set(preprocessed)))
all_tokens.extend(["<|endoftext|>", "<|unk|>"])

vocab = {token:integer for integer,token in enumerate(all_tokens)}
```

[ ]:
```
len(vocab.items())
```

[ ]:
```
for i, item in enumerate(list(vocab.items())[-5:]):
    print(item)
```

- We also need to adjust the tokenizer accordingly so that it knows when and how to use the new `<unk>` token

[ ]:
```
class SimpleTokenizerV2:
    def __init__(self, vocab):
        self.str_to_int = vocab
        self.int_to_str = { i:s for s,i in vocab.items()}

    def encode(self, text):
        preprocessed = re.split(r'([,.?_!"()\']|--|\s)', text)
        preprocessed = [item.strip() for item in preprocessed if item.strip()]
        preprocessed = [item if item in self.str_to_int
                        else "<|unk|>" for item in preprocessed]

        ids = [self.str_to_int[s] for s in preprocessed]
        return ids

    def decode(self, ids):
        text = " ".join([self.int_to_str[i] for i in ids])
        # Replace spaces before the specified punctuations
```

```
        text = re.sub(r'\s+([,.?!"()\']))', r'\1', text)
        return text
```

Let's try to tokenize text with the modified tokenizer:

[ ]:
```
tokenizer = SimpleTokenizerV2(vocab)

text1 = "Hello, do you like tea?"
text2 = "In the sunlit terraces of the palace."

text = " <|endoftext|> ".join((text1, text2))

print(text)
```

[ ]:
```
tokenizer.encode(text)
```

[ ]:
```
tokenizer.decode(tokenizer.encode(text))
```

1.5 2.5 BytePair encoding

- GPT-2 used BytePair encoding (BPE) as its tokenizer
- it allows the model to break down words that aren't in its predefined vocabulary into smaller subword units or even individual characters, enabling it to handle out-of-vocabulary words
- For instance, if GPT-2's vocabulary doesn't have the word "unfamiliarword," it might tokenize it as ["unfam", "iliar", "word"] or some other subword breakdown, depending on its trained BPE merges
- The original BPE tokenizer can be found here: https://github.com/openai/gpt-2/blob/master/src/encoder.py
- In this chapter, we are using the BPE tokenizer from OpenAI's open-source tiktoken library, which implements its core algorithms in Rust to improve computational performance
- I created a notebook in the ./bytepair_encoder that compares these two implementations side-by-side (tiktoken was about 5x faster on the sample text)

[ ]:
```
# pip install tiktoken
```

[ ]:
```
import importlib
import tiktoken

print("tiktoken version:", importlib.metadata.version("tiktoken"))
```

[ ]:
```
tokenizer = tiktoken.get_encoding("gpt2")
```

[ ]:
```
text = "Hello, do you like tea? <|endoftext|> In the sunlit terraces of␣
 ↪someunknownPlace."

integers = tokenizer.encode(text, allowed_special={"<|endoftext|>"})
```

```
print(integers)
```

[ ]:
```
strings = tokenizer.decode(integers)

print(strings)
```

- BPE tokenizers break down unknown words into subwords and individual characters:

1.6 2.6 Data sampling with a sliding window

- We train LLMs to generate one word at a time, so we want to prepare the training data accordingly where the next word in a sequence represents the target to predict:

[ ]:
```
with open("the-verdict.txt", "r", encoding="utf-8") as f:
    raw_text = f.read()

enc_text = tokenizer.encode(raw_text)
print(len(enc_text))
```

- For each text chunk, we want the inputs and targets
- Since we want the model to predict the next word, the targets are the inputs shifted by one position to the right

[ ]:
```
enc_sample = enc_text[50:]
```

[ ]:
```
context_size = 4

x = enc_sample[:context_size]
y = enc_sample[1:context_size+1]

print(f"x: {x}")
print(f"y:      {y}")
```

- One by one, the prediction would look like as follows:

[ ]:
```
for i in range(1, context_size+1):
    context = enc_sample[:i]
    desired = enc_sample[i]

    print(context, "---->", desired)
```

[ ]:
```
for i in range(1, context_size+1):
    context = enc_sample[:i]
    desired = enc_sample[i]

    print(tokenizer.decode(context), "---->", tokenizer.decode([desired]))
```

- We will take care of the next-word prediction in a later chapter after we covered the attention mechanism

- For now, we implement a simple data loader that iterates over the input dataset and returns the inputs and targets shifted by one
- Install and import PyTorch (see Appendix A for installation tips)

[ ]:
```
import torch
print("PyTorch version:", torch.__version__)
```

- We use a sliding window approach where we slide the window one word at a time (this is also known as `stride=1`):
- Create dataset and dataloader that extract chunks from the input text dataset

[ ]:
```
from torch.utils.data import Dataset, DataLoader


class GPTDatasetV1(Dataset):
    def __init__(self, txt, tokenizer, max_length, stride):
        self.tokenizer = tokenizer
        self.input_ids = []
        self.target_ids = []

        # Tokenize the entire text
        token_ids = tokenizer.encode(txt, allowed_special={'<|endoftext|>'})

        # Use a sliding window to chunk the book into overlapping sequences of␣
 ↪max_length
        for i in range(0, len(token_ids) - max_length, stride):
            input_chunk = token_ids[i:i + max_length]
            target_chunk = token_ids[i + 1: i + max_length + 1]
            self.input_ids.append(torch.tensor(input_chunk))
            self.target_ids.append(torch.tensor(target_chunk))

    def __len__(self):
        return len(self.input_ids)

    def __getitem__(self, idx):
        return self.input_ids[idx], self.target_ids[idx]
```

[ ]:
```
def create_dataloader_v1(txt, batch_size=4, max_length=256, stride=128,␣
 ↪shuffle=True, drop_last=True):

    # Initialize the tokenizer
    tokenizer = tiktoken.get_encoding("gpt2")

    # Create dataset
    dataset = GPTDatasetV1(txt, tokenizer, max_length, stride)

    # Create dataloader
```

```
    dataloader = DataLoader(
        dataset, batch_size=batch_size, shuffle=shuffle, drop_last=drop_last)

    return dataloader
```

- Let's test the dataloader with a batch size of 1 for an LLM with a context size of 4:

[ ]:
```
with open("the-verdict.txt", "r", encoding="utf-8") as f:
    raw_text = f.read()
```

[ ]:
```
dataloader = create_dataloader_v1(raw_text, batch_size=1, max_length=4,␣
 ↪stride=1, shuffle=False)

data_iter = iter(dataloader)
first_batch = next(data_iter)
print(first_batch)
```

[ ]:
```
second_batch = next(data_iter)
print(second_batch)
```

- An example using stride equal to the context length (here: 4) as shown below:
- We can also create batched outputs
- Note that we increase the stride here so that we don't have overlaps between the batches, since more overlap could lead to increased overfitting

[ ]:
```
dataloader = create_dataloader_v1(raw_text, batch_size=8, max_length=4,␣
 ↪stride=4, shuffle=False)

data_iter = iter(dataloader)
inputs, targets = next(data_iter)
print("Inputs:\n", inputs)
print("\nTargets:\n", targets)
```

1.7 2.7 Creating token embeddings

- The data is already almost ready for an LLM
- But lastly let us embed the tokens in a continuous vector representation using an embedding layer
- Usually, these embedding layers are part of the LLM itself and are updated (trained) during model training
- Suppose we have the following four input examples with input ids 5, 1, 3, and 2 (after tokenization):

[ ]:
```
input_ids = torch.tensor([2, 3, 5, 1])
```

- For the sake of simplicity, suppose we have a small vocabulary of only 6 words and we want to create embeddings of size 3:

[ ]:
```
vocab_size = 6
output_dim = 3

torch.manual_seed(123)
embedding_layer = torch.nn.Embedding(vocab_size, output_dim)
```

- This would result in a 6x3 weight matrix:

[ ]:
```
print(embedding_layer.weight)
```

- For those who are familiar with one-hot encoding, the embedding layer approach above is essentially just a more efficient way of implementing one-hot encoding followed by matrix multiplication in a fully-connected layer, which is described in the supplementary code in ./embedding_vs_matmul
- Because the embedding layer is just a more efficient implementation that is equivalent to the one-hot encoding and matrix-multiplication approach it can be seen as a neural network layer that can be optimized via backpropagation
- To convert a token with id 3 into a 3-dimensional vector, we do the following:

[ ]:
```
print(embedding_layer(torch.tensor([3])))
```

- Note that the above is the 4th row in the `embedding_layer` weight matrix
- To embed all four `input_ids` values above, we do

[ ]:
```
print(embedding_layer(input_ids))
```

[ ]:
```
- An embedding layer is essentially a look-up operation:
```

- **You may be interested in the bonus content comparing embedding layers with regular linear layers: `../02_bonus_efficient-multihead-attention`**

1.8 2.8 Encoding word positions

- Embedding layer convert IDs into identical vector representations regardless of where they are located in the input sequence:
- Positional embeddings are combined with the token embedding vector to form the input embeddings for a large language model:
- The BytePair encoder has a vocabulary size of 50,257:
- Suppose we want to encode the input tokens into a 256-dimensional vector representation:

[ ]:
```
vocab_size = 50257
output_dim = 256

token_embedding_layer = torch.nn.Embedding(vocab_size, output_dim)
```

- If we sample data from the dataloader, we embed the tokens in each batch into a 256-dimensional vector

- If we have a batch size of 8 with 4 tokens each, this results in a 8 x 4 x 256 tensor:

```
[ ]: max_length = 4
dataloader = create_dataloader_v1(raw_text, batch_size=8,␣
 ↪max_length=max_length, stride=max_length, shuffle=False)
data_iter = iter(dataloader)
inputs, targets = next(data_iter)
```

```
[ ]: print("Token IDs:\n", inputs)
print("\nInputs shape:\n", inputs.shape)
```

```
[ ]: token_embeddings = token_embedding_layer(inputs)
print(token_embeddings.shape)
```

- GPT-2 uses absolute position embeddings, so we just create another embedding layer:

```
[ ]: block_size = max_length
pos_embedding_layer = torch.nn.Embedding(block_size, output_dim)
```

```
[ ]: pos_embeddings = pos_embedding_layer(torch.arange(max_length))
print(pos_embeddings.shape)
```

- To create the input embeddings used in an LLM, we simply add the token and the positional embeddings:

```
[ ]: input_embeddings = token_embeddings + pos_embeddings
print(input_embeddings.shape)
```

- In the initial phase of the input processing workflow, the input text is segmented into separate tokens
- Following this segmentation, these tokens are transformed into token IDs based on a predefined vocabulary:

2 Summary and takeaways

See the `./dataloader.ipynb` code notebook, which is a concise version of the data loader that we implemented in this chapter and will need for training the GPT model in upcoming chapters.

See `./exercise-solutions.ipynb` for the exercise solutions.