# Introduction to tokenization

Puteaux, Fall/Winter 2020-2021

```
##############################################
##                                          ##
##  Natural Language Processing in Python   ##
##                                          ##
##############################################
```

```
§1 Introduction to Natural Language Processing in Python
```

```
§1.1 Regular expressions & word tokenization
```

## 1 Introduction to tokenization

### 1.1 What is tokenization?

- It turns a string or document into tokens (smaller chunks).
- It's one step in preparing a text for NLP.
- It has many different theories and rules.
- Users can create their own rules using regular expressions.
- There are some examples:
  - *breaking out words or sentences*
  - *separating punctuation*
  - *separating all hashtags in a tweet*

### 1.2 What is the NLTK library?

- NLTK: Natural Language Toolkit

### 1.3 Code of the NLTK library:

[1]:
```
from nltk.tokenize import word_tokenize

word_tokenize("Hi there!")
```

[1]:
```
['Hi', 'there', '!']
```

### 1.4 Why tokenize?

- Easier to map part of speech.
- To match common words.
- To remove unwanted tokens.
- E.g.,

```
>>> word_tokenize("I don't like Sam's shoes.")
['I', 'do', "n't", 'like', 'Sam', "'s", 'shoes', '.']
```

### 1.5 What are the other NLTK tokenizers?

- `sent_tokenize`: tokenize a document into sentences.
- `regexp_tokenize`: tokenize a string or document based on a regular expression pattern.
- `TweetTokenizer`: special class just for tweet tokenization, allowing separate hashtags, mentions, and lots of exclamation points, such as '!!!'.

### 1.6 Code of regex practice (the difference between `re.search()` and `re.match()`):

[2]:
```
import re

re.match('abc', 'abcde')
```

[2]:
```
<re.Match object; span=(0, 3), match='abc'>
```

[3]:
```
re.search('abc', 'abcde')
```

[3]:
```
<re.Match object; span=(0, 3), match='abc'>
```

[4]:
```
re.match('cd', 'abcde')
```

[5]:
```
re.search('cd', 'abcde')
```

[5]:
```
<re.Match object; span=(2, 4), match='cd'>
```

### 1.7 Practice exercises for introduction to tokenization:

▶ **Data pre-loading:**

[6]:
```
scene_one = open("ref2. Monty Python and the Holy Grail.txt").read()
```

▶ **NLTK word tokenization with practice:**

[7]:
```
# Import necessary modules
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
```

```
# Split scene_one into sentences: sentences
sentences = sent_tokenize(scene_one)

# Use word_tokenize to tokenize the fourth sentence: tokenized_sent
tokenized_sent = word_tokenize(sentences[3])

# Make a set of unique tokens in the entire scene: unique_tokens
unique_tokens = set(word_tokenize(scene_one))

# Print the unique tokens result
print(unique_tokens)
```

```
{'Ulk', 'uhh', 'CRASH', 'women', 'purely', 'protect', 'As', 'God', 'Zoot',
'stress', 'bottoms', 'guarded', 'scimitar', 'threw', 'beside', 'full',
"'Course", 'scholar', 'Use', 'bridges', 'bed', 'Bors', 'two-thirds', 'mortally',
'bangin', 'Ni', 'occasion', 'va.', 'Right', 'angels', 'already', 'i', 'main',
'breadth', 'jump', 'tragic', 'lying', 'foul', 'seems', 'spank', 'excuse',
'ratified', 'scratch', 'B', 'daft', 'animal', 'Practice', 'medical', 'Thy',
'Hold', 'gone', 'na', 'imperialist', 'Great', 'Bravest', 'bastard', 'noise',
'Quickly', 'as', 'nibble', 'anyone', 'Back', 'scrape', 'cheesy', 'committed',
'On', 'through', 'on', 'drilllll', 'problem', 'song', 'pack', 'Gable',
'Brother', 'house', 'death', 'Lie', 'mer', 'wounding', 'gave', 'up', 'tired',
'huge', 'tie', 'sample', 'understand', 'ones', 'Shrubberies', 'meeting', 'clad',
'can', 'autonomous', 'room', 'worry', 'legs', 'hacked', 'Even', 'sign', 'Order',
'looney', 'ROBIN', 'electric', 'if', 'son', 'donaeis', 'now', 'howl', 'samite',
'purpose', 'Thsss', 'haste', 'executive', 'hospitality', 'ordinary', 'entrance',
'sloths', 'perpetuates', 'over', 'Say', 'boil', 'feathers', 'cover', 'Huy',
'stood', 'seen', 'went', 'outdoors', 'GUARD', 'dungeon', 'time-a', 'this',
'elderberries', 'rhymes', 'its', 'binding', 'art', 'easily', 'OFFICER', 'tell',
'Hiyaah', 'and', 'CHARACTER', 'a', 'Exactly', 'out-clever', 'employed',
'seemed', 'bottom', 'bond', 'LEFT', 'busy', 'spirit', 'obviously', 'kill',
'quack', 'winter', 'bunny', 'guard', 'awfully', 'door-opening', 'man',
'shelter', 'Unfortunately', 'scales', 'going', 'averting', 'aaaaaah',
'learning', 'shrubber', 'gon', 'scarper', 'CROWD', '14', 'Halt', 'rock',
'penalty', 'unarmed', 'Castle', 'talk', 'fled', 'forced', 'smelt', 'u', 'mile',
'person', 'names', 'ask', 'kills', '…', 'Saint', 'lambs', 'pond', 'centuries',
'speak', 'grail-shaped', 'brush', 'aloft', "n't", 'travellers', 'delirious',
'riding', '13', 'confuse', 'holy', 'dictatorship', 'aside', 'worked', 'sawwwww',
'count', 'Holy', 'Beast', 'started', "'Ecky-ecky-ecky-ecky-pikang-zoop-boing-
goodem-zoo-owli-zhiv", 'lives', 'BROTHER', 'So', 'runes', 'awaaaaay', 'By',
'for', 'giggle', 'Aaaaaah', 'able', '9', 'Bedwere', 'king', 'dunno', 'PARTY',
'wipers', 'draw', 'Woa', 'so-called', 'Aauuuves', 'Firstly', 'Packing', 'foe',
'Bedevere', 'near', 'sword', 'RIGHT', 'reads', 'me', 'Clark', "'Dennis",
'sniff', 'fellows', 'times', 'once', 'Hic', 'kingdom', 'Your', 'la', 'ignore',
'7', 'Anyway', 'domine', 'agree', 'smashing', 'stew', 'If', "C'est",
'punishment', 'thank', 'three', 'guests', 'enough', 'since', 'please',
'Aramaic', 'vain', 'Launcelot', "'re", 'grail', 'commune', 'King', 'nor',
```

```
'dare', 'snows', 'hopeless', 'Mercea', 'b', 'Most', 'also', 'warmer', 'bats',
'Cherries', 'hundred-and-fifty', 'previous', 'cadeau', 'interested', 'shit',
'father', 'Bring', 'ounce', 'illustrious', 'regulations', 'nine', 'Winter',
'derives', 'Hiyya', 'understanding', 'completely', 'favor', 'place', 'Book',
'At', 'I', 'formed', 'bicker', 'period', 'pounds', 'ARTHUR', 'guards', 'left',
'court', 'somewhere', 'expect', 'Oh', 'anything', 'Court', 'expensive',
'classes', 'Running', 'rocks', 'resumes', 'guest', 'outdated', 'earth', 'head',
'sixteen', 'eisrequiem', 'snuff', 'fatal', 'training', 'minute', 'chops',
'soft', 'real', 'sorry', 'Dramatically', 'heroic', 'GUEST', 'Armaments', 'Lord',
'sod', 'because', 'Is', 'lads', 'strewn', 'ere', 'aaugh', 'Nador', 'wan', 'We',
'supports', 'mayest', 'pound', 'tiny', 'line', 'nick', 'SCENE', 'SENTRY',
'either', 'HEAD', 'false', 'carving', 'north', 'hello', 'entering', 'music',
'Run', 'HERBERT', 'MINSTREL', 'owns', 'cope', 'automatically', 'summon',
'safety', 'meant', 'Alright', 'HISTORIAN', 'groveling', 'behaviour', 'Hey',
"'sorry", 'waste', 'use', 'masses', 'Concorde', 'eis', '2', 'his', 'aunties',
'lovely', 'wield', 'Auuuuuuuugh', 'plain', 'Rheged', 'retreat', 'hear',
'ridden', 'frighten', 'sovereign', 'breakfast', 'work', 'oral', 'MONKS', 'Wood',
'liege', 'weapon', 'sometimes', 'dirty', 'Hya', 'economic', 'tonight', 'BORS',
'laurels', 'zoosh', 'donkey-bottom', 'from', '11', 'remember', 'give', 'warned',
'empty', '15', 'Why', 'Over', 'This', 'spanked', 'An', 'swords', 'testicles',
'Mother', 'FRENCH', 'society', 'anyway', 'name', 'needs', 'although',
'footwork', 'ALL', 'marrying', 'buggering', 'teeth', 'command', 'twong',
"'Aauuuuugh", 'thump', 'Eee', 'he', 'each', 'defeat', 'CRONE', 'Defeat',
'being', 'effect', 'my', 'built', 'arrange', 'bite', 'Bad', 'proved', 'against',
'these', 'k-nnnnniggets', 'himself', "'", 'Tale', 'relics', 'something',
'smack', 'mangled', 'daring', 'bring', 'Said', 'known', 'looks', 'watery',
'course', 'appease', 'unladen', 'Whoa', 'types-a', 'Winston', 'cereals', 'boys',
'are', 'risk', 'lady', 'test', 'fooling', 'SUN', 'follow', 'ill.', 'BLACK',
'gravy', 'pimples', 'All', 'length', 'marry', 'forward', 'valor', 'favorite',
'Olfin', 'biggest', 'nice', 'Off', 'some', 'police', 'heard', 'nearer', 'lies',
'end', 'Farewell', 'velocity', 'inferior', 'Shall', 'Lady', 'Cornwall',
'defeator', 'life', 'Honestly', 'chest', 'Thee', 'like', 'questions',
'Caerbannog', 'warm', 'Crapper', 'Torment', 'while', 'Hello', 'mumble', 'stone',
'differences', 'Aagh', 'Ninepence', 'NARRATOR', 'woosh', 'Ooh', 'Pin', 'Quoi',
'gay', 'fruit', 'Bread', 'weighs', 'That', 'strength', 'Very', 'major',
'asking', 'rather', 'A', 'WOMAN', 'recover', 'even', 'Battle', 'repressed', '!',
'word', 'resting', 'off', 'approaching', 'European', 'fallen', 'But', 'Hooray',
'strand', 'Be', 'rodent', 'Patsy', 'could', '3', 'Un', 'Lake', 'chance', 'Waa',
'vests', 'thwonk', 'accent', 'Tell', 'They', "'To", 'same', 'temptation',
'fight', 'question', 'accomplished', 'simple', 'make', 'Providence', 'Surely',
'took', 'daughter', 'Aaah', 'enemies', 'Knights', 'With', 'says', 'would',
'arm', 'with', 'newt', 'imprisoned', 'PERSON', 'sink', 'ways', 'call', 'basis',
'hiyaah', 'snore', 'bitching', 'whose', 'legally', 'More', 'counting', 'Oooo',
'doors', 'doing', 'Uh', 'feast', 'saw', 'bid', 'wide', 'SECOND', "d'you",
'stand', 'them', 'see', 'SHRUBBER', 'chastity', 'Erm', 'Ewing', 'awaaay',
'Thou', 'requiem', 'someone', 'breath', 'face', 'trade', 'used', '--', 'sworn',
'air-speed', 'jokes', 'OTHER', 'Anybody', 'Since', 'she', 'last', 'farcical',
'utterly', 'tail', 'ladies', 'cough', 'apart', 'Good', 'KNIGHTS', 'Action',
```

```
'single-handed', 'frozen', 'carry', 'collective', 'Sorry', 'Mmm', 'open',
'dying', 'grips', 'retold', 'remain', 'brought', 'shows', 'supposed', '1',
'eight', 'twin', 'between', 'wonderful', '18', 'eet', 'Man', 'Pie', 'blessing',
'wood', "'aaaah", 'throughout', 'Am', 'witch', 'own', 'Once', 'bleeder', 'sad',
'Apples', 'here', 'making', 'largest', 'yet', 'Every', 'capital', 'chorus',
'entered', 'getting', 'Pure', 'orangutans', 'consulted', 'everyone', 'lie',
'Badon', 'Hoo', 'Blue', 'undressing', 'ugly', 'biscuits', 'Herbert', 'Huyah',
'certainly', 'join', 'named', 'Aaaaaaaah', 'ROGER', 'spam', "'ve", 'fly',
'show', 'squeak', 'Father', 'ferocity', 'Yes', 'nasty', 'bold', 'bless',
'reached', 'Today', 'so', 'get', 'look', 'pansy', 'wind', 'merger', 'Aaauggh',
"'ll", 'am', 'brave', 'Bristol', 'into', 'unsingable', 'Then', 'Yay', 'does',
'Lancelot', 'old', 'we', 'rich', 'dangerous', 'cry', 'Actually', 'VILLAGER',
'returns', 'sacrifice', 'wicked', 'what', 'tea', 'cop', 'LUCKY', 'totally',
'of', 'twang', 'Listen', 'Our', 'Iiiives', 'under', 'fire', 'rewr', 'eyes',
'tear', 'blondes', 'mightiest', 'ha', 'splat', 'burned', 'keep', '24',
'Consult', 'kind', 'their', 'Does', 'suggesting', 'carrying', 'His', 'build',
'minutes', 'Chicken', 'plover', 'set', 'bravely', 'move', 'another', 'our',
'officer', 'Help', 'icy', ',', 'Galahad', 'Sir', 'Three', 'absolutely',
'keeper', 'bowels', 'git', 'Attila', 'knows', 'maintain', 'DENNIS', 'creak',
'brain', 'valleys', 'hand', 'behold', 'brunettes', 'water', 'rejoicing',
'spoken', '..', "'em", 'Supposing', 'tropical', 'shrubberies', 'done',
'majority', 'Burn', 'Dis-mount', 'carried', 'Ohh', 'upon', 'that', 'direction',
'Pendragon', 'cut', 'kings', 'HEADS', 'plan', 'finds', 'north-east',
'exploiting', 'happens', 'Try', 'been', 'union', 'shall', 'wants', 'Midget',
'Skip', 'Brave', 'tackle', 'silly', 'hamster', 'say', 'Build', 'Have', 'Bones',
'k-nnniggets', 'problems', 'Will', 'might', 'TIM', 'sing', 'scots', 'KING',
'pure', 'keen', '19', 'Britons', 'swamp', 'PIGLET', 'Could', 'depressing',
'impersonate', 'dressing', 'Iesu', 'aquatic', 'Huh', 'not', "'Morning",
'kneecaps', 'bows', 'swallows', '22', 'rabbit', 'married', 'heart', 'bugger-
folk', 'Now', 'until', 'doctors', 'Himself', 'stops', 'Looks', 'rest',
'Nothing', 'big', 'round', 'Until', 'signifying', 'scared', 'beat', 'argue',
'internal', 'note', 'arms', 'offensive', 'where', 'burn', 'power', 'Ha',
'nightfall', 'nervous', 'continue', 'handsome', 'Lead', 'evil', 'nothing',
'thing', 'THE', 'Oooohoohohooo', ':', 'performance', 'floats', 'surprise',
'dynamite', 'WITCH', 'pause', 'sun', 'Summer', "e'er", 'And', 'Ay', 'exciting',
'change', 'auntie', 'along', 'ptoo', 'than', "'Til", 'awaits', 'terribly',
'legendary', 'hall', 'formidable', 'humble', 'Aaaugh', 'dictating', 'peril',
'Mud', 'Together', 'Ives', 'ooh', 'gouged', 'autocracy', 'running', 'avenged',
'pull', 'BRIDE', 'sort', 'havin', 'tinder', 'when', 'Meanwhile', 'Too',
'Divine', 'bet', 'armed', 'buggered', 'grip', 'unplugged', 'sneaking', 'U',
'suspenseful', 'routines', 'setting', 'tart', 'Camelot', 'MAN', 'together',
'Hang', 'convinced', 'Saxons', 'separate', 'radio', 'feel', 'coconuts',
'gallantly', 'Thpppppt', 'always', 'SOLDIER', "'Ni", 'dancing', 'stupid', '20',
'given', 'must', 'hee', 'ever', 'Haw', 'glad', 'worst', "'round", 'ho',
'Grenade', 'towards', 'guided', 'weather', 'eh', 'men', 'object', 'GUARDS',
'it', 'science', 'take', 'Fine', 'filth', 'uuggggggh', 'eat', 'sight',
'taunting', 'sweet', 'far', 'Are', 'working', 'mud', 'repressing', 'walking',
'relax', 'personally', 'easy', 'is', 'Ah', 'were', 'See', 'honored', 'chord',
```

```
'CUSTOMER', 'lad', 'in', 'Spring', 'goes', 'wrong', 'third', 'by', 'ceremony',
'lost', 'Hallo', 'wait', 'removed', 'yelling', 'violence', 'awhile', 'almost',
'Other', 'no', 'all', '[', 'She', 'nostrils', 'lonely', 'decided', 'French',
"'uuggggggh", 'covered', 'ours', 'tough', 'anarcho-syndicalist', 'Table',
'chickened', "'cause", 'herring', 'acting', 'sank', 'Autumn', 'Loimbard',
'dogma', 'enchanter', 'find', 'class', 'night', '.', 'husk', 'words',
'minstrels', 'chickening', 'gentle', 'search', 'passed', 'quests', 'Away',
'bridgekeeper', 'clear', 'somebody', 'Here', 'long', 'RANDOM', 'called',
'Cider', 'PRINCESS', 'let', "'it", 'amazes', 'stayed', 'many', 'Heh', 'None',
'need', 'attend', 'shalt', 'lord', 'To', 'Round', 'e', 'afraid', "'is",
'saying', 'cost', 'rescue', 'good', 'invincible', 'Please', 'Old', 'clop',
'alight', 'tree', 'whether', 'examine', 'sons', 'Supreme', 'or', 'beacon',
'Ridden', 'closest', 'twenty', 'starling', 'Look', 'mooooooo', 'What', 'you',
'woods', 'one', 'strategy', 'valiant', 'king-a', 'feet', 'ANIMATOR',
'strangers', 'Twenty-one', 'mercy', 'liver', 'vouchsafed', 'perilous',
"'Ooooooh", 'Speak', 'tit', 'Tower', 'watch', 'clunk', 'reared', 'mangy',
'Allo', 'swallow', 'bravest', 'very', 'compared', 'stretched', 'haaa',
'pissing', 'Peng', 'stab', "'Man", 'workers', 'example', 'y', 'haw',
'identical', 'dear', 'side', 'fourth', 'terrible', 'ride', 'your', 'without',
'intermission', 'fought', 'suffered', 'dappy', 'sell', 'Get', 'temptress',
'Clear', 'stuffed', 'yellow', 'miserable', 'knocked', '4', 'impeccable', 'Uuh',
'having', 'mac', 'order', 'escape', 'Thppt', "'First", 'DEAD', 'France', 'case',
'answer', 'It', 'writing', 'coconut', 'vital', '#', 'dressed', 'Amen', 'climes',
'mean', 'Nine', 'Those', 'twenty-four', 'Stop', 'behind', 'naughty', 'warning',
'outside', 'vicious', 'curtains', "'anging", 'return', 'splash', 'Monsieur',
'Oui', 'woman', 'quick', 'bum', 'BRIDGEKEEPER', 'CARTOON', 'think', 'crash',
'PATSY', 'Antioch', 'castle', 'really', 'bother', 'Make', "'Ere", 'Explain',
"'forgive", 'sequin', 'to', 'most', 'properly', 'too', 'such', 'indeed', 'May',
'types', 'c', 'One', 'Thppppt', 'ninepence', 'living', 'bloody', 'killer',
'verses', 'Schools', 'bird', 'Arthur', 'medieval', 'Bon', 'maybe', 'Black',
'met', 'helpful', 'mandate', 'Tim', 'mayhem', 'pray', 'bit', 'pen', 'chosen',
'gra', 'GREEN', 'CRAPPER', 'gurgle', 'killed', 'assault', 'never', 'o', 'jam',
'handle', 'at', 'Dragon', 'looked', 'underwear', 'Who', 'telling', 'dramatic',
'profane', 'who', 'Robinson', 'yourself', 'wayy', 'cross', 'whoever', 'other',
'witness', 'lair', 'Princess', 'Britain', 'girl', 'wet', 'least', 'forty-three',
'thonk', 'dub', 'Gorge', 'turned', 'worried', 's', 'purest', 'Assyria', 'come',
'table', 'system', 'temperate', 'Christ', 'Aauuggghhh', 'coming', 'Eh', 'For',
'VILLAGERS', 'body', 'happy', 'wishes', 'split', 'color', 'silence', 'creature',
'wart', '6', 'Anarcho-syndicalism', 'vary', 'lapin', 'Forgive', 'alarm', 'door',
'Aaaah', 'wise', 'Neee-wom', 'SIR', 'How', 'today', 'Quiet', 'finest', 'friend',
'carved', 'ARMY', 'limbs', '8', 'every', 'Gawain', 'slightly', 'Greetings',
'whom', 'sire', 'servant', 'business', 'Lucky', 'Excalibur', 'Aggh', 'Ow', 'Of',
'thou', 'found', 'Grail', 'danger', 'Hmm', 'creep', 'wo', 'send', 'walk',
'land', 'score', 'mashed', 'hospital', 'un', 'dorsal', 'ready', 'lived',
'Everything', 'knights', 'had', 'badger', 'hang', '21', 'bladders', 'down',
'Dennis', 'dragging', 'Ages', 'scott', 'less', 'CAMERAMAN', 'string', 'baaaa',
'apologise', 'laden', 'held', 'Prepare', 'Which', 'Did', 'dad', 'further',
'weight', 'town', 'just', 'Doctor', 'longer', 'idiom', 'lose', 'about', 'spake',
```

```
'crone', 'yours', 'INSPECTOR', 'knew', 'leaps', 'sheep', 'frontal', 'heads',
'crying', 'looking', 'commands', 'changed', 'pweeng', 'discovers', 'castanets',
'Y', 'has', 'adversary', 'Far', 'boom', 'keepers', 'five', 'duck', 'smashed',
'die', 'pig-dogs', 'Bravely', 'any', 'clllank', 'flesh', 'raped', 'spooky',
'ehh', 'explain', 'everything', 'bridge', 'knees-bent', 'van', 'may', 'dress',
'better', 'inside', 'best', 'Bloody', 'war', 'required', 'WINSTON', 'suffice',
'shrubbery', 'high-pitched', 'Shh', 'south', 'Follow', 'non-migratory',
'general', 'individually', 'ca', 'path', 'Hand', 'flight', 'Um', 'alive', 'Do',
'laughing', 'English', 'ran', 'outwit', 'Shrubber', 'elbows', 'discovered',
'cart', 'push', 'high', 'present', 'Ayy', 'thud', 'diaphragm', 'second', 'clap',
'bois', 'Knight', 'Put', 'bad', 'Really', 'Dappy', 'zone', 'Hoa', 'Thursday',
'vote', 'peasant', 'bathing', 'Angnor', 'hmm', 'the', 'sigh', 'er', 'Churches',
'presence', 'why', 'GUESTS', 'headed', 'persons', 'said', 'Prince', 'seem',
'Wait', 'suit', 'listen', 'reasonable', 'CART-MASTER', 'assist', 'saved',
'nose', 'N', 'armor', 'heh', 'Perhaps', 'beds', 'carries', 'thanks', 'Picture',
'go', 'Between', 'Found', 'quarrel', 'l', 'straight', 'soiled', 'Steady', '17',
'g', 'creeper', 'emperor', 'seek', 'drink', 'dance', "'Here", 'pass', 'bones',
'large', 'band', 'Thank', 'manner', 'shivering', 'snap', 'thirty-seven',
'wound', 'particularly', 'successful', 'fifty', 'Aauuugh', 'home', 'eats',
'influential', 'Chapter', 'Aaauugh', 'depart', 'carp', 'Hiyah', '16', 'taunt',
'England', 'mooo', 'food', 'Would', 'Keep', 'Joseph', 'know', 'Umhm', 'worthy',
'Let', 'Ask', 'Or', 'Pull', 'banana-shaped', 'prevent', 'mate', 'Wayy', 'heeh',
'praised', 'thy', 'blood', 'young', 'spanking', 'an', 'Riiight', 'much',
'Victory', 'cave', 'felt', 'Nu', 'conclusion', 'dark', 'crossed', "'T", 'baby',
'mystic', 'harmless', 'strongest', 'visually', 'triumphs', 'history', 'unclog',
'j', 'couple', 'was', 'PRISONER', 'language', 'suddenly', 'leap', 'My', 'Uhh',
'Uugh', 'makes', 'excepting', 'ai', 'us', 'quiet', 'hidden', 'settles',
'African', 'MIDGET', 'slash', 'him', 'Enchanter', 'Robin', 'suppose', 'ratios',
'got', 'still', 'nice-a', 'fart', 'gained', 'wounded', 'The', 'window-dresser',
'knock', 'Where', 'lunged', 'forget', 'trough', 'k-niggets', 'lobbest',
'whispering', 'those', 'Swamp', 'private', 'lot', 'force', 'else', 'ju',
'vache', 'fold', 'scene', 'stay', 'back', 'Stay', 'comin', 'trusty', 'charged',
'kicked', 'cartoon', 'STUNNER', 'progress', 'Not', 'biters', 'asks', 'did',
'four', 'Eternal', 'Rather', 'particular', 'Fiends', 'nineteen-and-a-half',
'Mind', 'two-level', 'thought', 'bits', 'wiper', 'halves', 'LOVELY', 'Almighty',
'10', 'bride', 'higher', 'pig', 'When', 'horse', 'folk', 'Four', 'forth',
'roar', 'Iiiiives', 'idea', 'voluntarily', 'Welcome', 'strange', 'ungallant',
'yel', 'the-not-quite-so-brave-as-Sir-Lancelot', ';', 'Idiom', 'Aaaaaaaaah',
'fair', 'dull', 'pram', 'moment', 'well', 'Aaagh', '5', 'sonny', 'Leaving',
'refuse', 'traveller', 'right', 'Psalms', '(', 'sacred', 'Ahh', 'quite',
'Camaaaaaargue', 'sister', 'PRINCE', 'earthquakes', 'fwump', 'dona', 'proceed',
'Augh', 'small', 'actually', 'ye', 'ENCHANTER', 'Just', 'there', 'only', "'d",
'beautiful', 'Forward', 'bang', 'bed-wetting', 'oooh', 'demand', 'pestilence',
'Yapping', 'social', "'aaggggh", 'give-away', 'grovel', 'Fetchez', 'MIDDLE',
'model', 'ZOOT', 'number', 'necessary', 'WIFE', 'help', 'time', 'islands',
'pay', 'True', 'passing', 'oo', 'do', 'kneeling', 'shimmering', 'special',
'blow', '12', 'Piglet', 'Recently', 'middle', 'leave', 'then', 'act', 'told',
'accompanied', "'m", 'out', 'FATHER', 'nobody', 'hoo', 'outrageous',
```

```
'Chickennn', 'Oooh', 'Behold', 'CHARACTERS', 'supreme', 'sir', 'bonk', 'In',
'martin', 'courage', 'luck', 'mad', 'worse', 'fell', 'task', 'day', 'throwing',
'horn', 'sense', 'thine', 'soon', 'GALAHAD', 'Like', 'scenes', "'S", 'witches',
'chanting', 'pulp', 'bleed', 'unhealthy', 'doubt', 'sex', 'leads', 'distress',
'siren', 'Five', 'moistened', 'answers', 'Come', 'Dingo', 'joyful', 'Excuse',
'late', 'using', 'taken', 'mine', 'logically', 'Frank', 'request',
'approacheth', 'There', 'oh', 'n', 'try', 'remembered', 'Heee', 'disheartened',
'uh', 'week', 'whop', 'um', 'live', 'animator', 'chu', 'glory', 'tracts', 'be',
'etc', 'ponds', 'Chop', 'immediately', 'clack', 'bringing', 'Guards',
'anywhere', 'KNIGHT', 'flint', 'year', 'Gallahad', 'de', 'Roger', 'lobbed',
'Hee', 'fine', 'attack', 'became', 'carve', 'master', 'Umm', 'Quick',
'ruffians', 'Remove', ']', 'Charge', 'sure', 'enjoying', 'Hurry', 'rrrr',
'Yeaah', 'LAUNCELOT', 'duty', 'way', 'magne', 'self-perpetuating', 'advancing',
'parts', 'Hyy', 'Peril', 'more', 'cruel', 'CONCORDE', 'DINGO', 'away', 'hills',
'treat', 'Yeaaah', 'made', 'scribble', 'wave', 'He', 'Arimathea',
'distributing', 'have', ')', 'trumpets', 'sent', 'migrate', "'ni", 'grin',
'her', 'clue', 'Splendid', 'tops', 'bi-weekly', 'after', 'throat', 'Quite',
'auuuuuuuugh', 'bastards', 'Must', 'decision', 'Ho', 'inherent', 'Not-appearing-
in-this-film', 'Open', 'preserving', 'w', 'fortune', 'Beyond', 'trouble',
'Maynard', 'liar', 'Go', 'put', 'opera', 'conclusions', 'Ector', 'AMAZING',
'Never', 'oui', 'Chaste', 'but', 'dine', 'will', 'broken', 'raised', 'rope',
'You', 'shut', 'sponge', 'Nay', 'run', 'tale', 'Tall', 'enter', 'GOD', 'around',
'singing', 'affairs', 'government', 'Well', 'guiding', 'ham', 'OF', 'wings',
'deeds', 'taking', 'anchovies', 'leg', 'bint', 'design', 'deal', 'hat',
'Silence', 'certain', "'Erbert", 'again', 'sharp', 'wooden', 'stop', 'hast',
'tap-dancing', 'Thpppt', 'next', 'freedom', 'mistake', 'OLD', 'Stand', 'whinny',
'indefatigable', 'wedding', 'packing', 'great', 'Hm', 'Therefore', "'old",
'Alice', 'MAYNARD', 'W', '23', 'BEDEVERE', 'arrows', 'matter', 'wedlock',
'VOICE', 'allowed', 'died', 'birds', 'p', 'yeah', 'tiny-brained', 'beyond',
'two', "'s", "'shrubberies", 'clang', 'nearly', 'Silly', 'Hah', 'burst', 'yes',
'Throw', 'considerable', 'cast', 'first', 'which', 'Yup', 'they', 'No', 'bosom',
'Walk', 'pussy', 'Anthrax', 'O', 'flights', 'Jesus', 'forest', 'therefore',
'new', 'knight', 'eccentric', 'feint', 'point', 'afoot', 'wherein', 'buy',
'GIRLS', "'til", 'NI', 'illegitimate-faced', 'dead', 'rode', 'Death', 'foot',
'Aaaaugh', 'later', 'headoff', 'Aah', 'little', '?', 'Guy', 'things', 'grenade',
'Hill', 'how', 'ethereal', 'Yeah', 'ni', 'blanket', 'Mine', 'Bridge', 'glass',
'Seek', 'horrendous', 'pointy', 'seldom', 'Two', 'people', 'bad-tempered',
'should', 'welcome', 'clank', 'aptly', 'hell', 'Shut', 'DIRECTOR', 'lucky',
'uuup', 'basic', 'Uther', 'quest', 'country', 'mother', 'Agh', 'varletesses',
'streak', 'become', 'Cut', 'kick', 'Message', 'bells', 'turns', 'want', 'miss'}
```

▶ **Package pre-loading:**

[8]:
```
import re
```

▶ **Regex (`re.search()`) practice:**

[9]:
```
# Search for the first occurrence of "coconuts" in scene_one: match
match = re.search("coconuts", scene_one)

# Print the start and end indexes of match
print(match.start(), match.end())
```

```
580 588
```

[10]:
```
# Write a regular expression to search for anything in square brackets: pattern1
pattern1 = r"\[.*\]"

# Use re.search to find the first text in square brackets
print(re.search(pattern1, scene_one))
```

```
<re.Match object; span=(9, 32), match='[wind] [clop clop clop]'>
```

[11]:
```
# Find the script notation at the beginning of the fourth sentence and print it
pattern2 = r"[\w\s#]+:"
print(re.match(pattern2, sentences[3]))
```

```
<re.Match object; span=(0, 7), match='ARTHUR:'>
```