1. با استفاده از عبارات منظم متن انگلیسی زیر را به توکن ها تبدیل کنید.

جهت استفاده از عبارات منظم ماژول regex با دستور re اضافه شده است. سپس متن خود را در متغیر my_string گذاشته و دستور های عبارات منظم خود را در متغیر pattern نوشته شده است و در نهایت تمامی توکن ها از متن استخراج میشوند.

- عبارات منظم بر اساس متد استاندارد نوشته شده در داخل کتاب رفرنس نوشته شده است که تا جای ممکن درست ترین حروف را برای توکن کردن در نظر بگیرد. هدف هر عبارت روبه روی آن کامنت شده است.
- پرچم re.VERBOSE به جهت حذف whitespace و افزایش خوانایی توکن ها به کار رفته است.

کد:

assignment1.py

```python
import re
my_string = """After World War II, the British greatly reduced the use of the full stop and other punctuation points af
But before that, many Britons were more scrupulous at maintaining the French form. In French, the period only follows a
Over the years, however, the lack of convention in some style guides has made it difficult to determine which two-word
Minimization of punctuation in typewritten material became economically desirable in the 1960s and 1970s for the many u
Widespread use of electronic communication through mobile phones and the Internet during the 1990s allowed for a marked
"""
PATTERN = r"""
    (?:[A-Z]\.)+              # Abbreviations, e.g. U.S.A., S.O.E.
    | \w+(?:-\w+)*            # Words with optional internal hyphens
    | \$?\d+(?:\.\d+)?%?      # Currency and percentages, e.g. $12.40, 82%
    | \.\.\.                  # Ellipsis
    | [.,;"'?():_`-]          # Separate punctuation tokens
"""
# Find all tokens
tokens = re.findall(PATTERN, my_string, re.VERBOSE)

print(tokens)
```

خروجی:

```
"E:\Alzahra\NLP Course\Homework1\Scripts\python.exe" "E:\Alzahra\NLP Course\Homework1\assignment1.py"
['After', 'World', 'War', 'II', ',', 'the', 'British', 'greatly', 'reduced', 'the', 'use', 'of', 'the', 'full', 'stop', 'and', 'other', 'punctuation', 'points', 'after', 'abbreviations', 'in', 'at

Process finished with exit code 0
```

After', 'World', 'War', 'II', ',', 'the', 'British', 'greatly', 'reduced', 'the', 'use', 'of', 'the', 'full', 'stop', ']
'and', 'other', 'punctuation', 'points', 'after', 'abbreviations', 'in', 'at', 'least', 'semi-formal', 'writing', ',',
'while', 'the', 'Americans', 'more', 'readily', 'kept', 'such', 'use', 'until', 'more', 'recently', ',', 'and', 'still',
'maintain', 'it', 'more', 'than', 'Britons', '.', 'The', 'classic', 'example', ',', 'considered', 'by', 'their',
'American', 'counterparts', 'quite', 'curious', ',', 'was', 'the', 'maintenance', 'of', 'the', 'internal',
'comma', 'in', 'a', 'British', 'organisation', 'of', 'secret', 'agents', 'called', 'the', '"', 'Special',
'Operations', ',', 'Executive', '"', ',', '"', 'S.O.', ',', 'E', '"', ',', 'which', 'is', 'not', 'found', 'in', 'histories',
'written', 'after', 'about', '1960', '.', 'But', 'before', 'that', ',', 'many', 'Britons', 'were', 'more',
'scrupulous', 'at', 'maintaining', 'the', 'French', 'form', '.', 'In', 'French', ',', 'the', 'period', 'only',
'follows', 'an', 'abbreviation', 'if', 'the', 'last', 'letter', 'in', 'the', 'abbreviation', 'is', 'not', 'the', 'last',
'letter', 'of', 'its', 'antecedent', ':', '"', 'M.', '"', 'is', 'the', 'abbreviation', 'for', '"', 'monsieur', '"', 'while', '"',
'Mme', '"', 'is', 'that', 'for', '"', 'madame', '"', '.', 'Like', 'many', 'other', 'cross-channel', 'linguistic',

'acquisitions', ',', 'many', 'Britons', 'readily', 'took', 'this', 'up', 'and', 'followed', 'this', 'rule', 'themselves', ',', 'while', 'the', 'Americans', 'took', 'a', 'simpler', 'rule', 'and', 'applied', 'it', 'rigorously', '.', 'Over', 'the', 'years', ',', 'however', ',', 'the', 'lack', 'of', 'convention', 'in', 'some', 'style', 'guides', 'has', 'made', 'it', 'difficult', 'to', 'determine', 'which', 'two-word', 'abbreviations', 'should', 'be', 'abbreviated', 'with', 'periods', 'and', 'which', 'should', 'not', '.', 'The', 'U.S.', 'media', 'tend', 'to', 'use', 'periods', 'in', 'two-word', 'abbreviations', 'like', 'United', 'States', '(', 'U.S.', ')', ',', 'but', 'not', 'personal', 'computer', '(', 'PC', ')', 'or', 'television', '(', 'TV', ')', '.', 'Many', 'British', 'publications', 'have', 'gradually', 'done', 'away', 'with', 'the', 'use', 'of', 'periods', 'in', 'abbreviations', '.', 'Minimization', 'of', 'punctuation', 'in', 'typewritten', 'material', 'became', 'economically', 'desirable', 'in', 'the', '1960s', 'and', '1970s', 'for', 'the', 'many', 'users', 'of', 'carbon-film', 'ribbons', 'since', 'a', 'period', 'or', 'comma', 'consumed', 'the', 'same', 'length', 'of', 'non-reusable', 'expensive', 'ribbon', 'as', 'did', 'a', 'capital', 'letter', '.', 'Widespread', 'use', 'of', 'electronic', 'communication', 'through', 'mobile', 'phones', 'and', 'the', 'Internet', 'during', 'the', '1990s', 'allowed', 'for', 'a', 'marked', 'rise', 'in', 'colloquial', 'abbreviation', '.', 'This', 'was', 'due', 'largely', 'to', 'increasing', 'popularity', 'of', 'textual', 'communication', 'services', 'such', 'as', 'instant', '-', 'and', 'text', 'messaging', '.', 'SMS', ',', 'for', 'instance', ',', 'supports', 'message', 'lengths', 'of', '160', 'characters', 'at', 'most', '(', 'using', 'the', 'GSM', '03', '.', '38', 'character', 'set', ')', '.', 'This', 'brevity', 'gave', 'rise', 'to', 'an', 'informal', 'abbreviation', 'scheme', 'sometimes', 'called', 'Textese', ',', 'with', 'which', '10', 'or', 'more', 'of', 'the', 'words', 'in', 'a', 'typical', 'SMS', 'message', 'are', 'abbreviated', '.', 'More', 'recently', 'Twitter', ',', 'a', 'popular', 'social', 'networking', 'service', ',', 'began', 'driving', 'abbreviation', 'use', 'with', '140', 'character', 'message', 'limits', '.']

2. با استفاده از عبارات منظم و الگوریتم Porter (الگوریتم کاملتر را از اینترنت استخراج کنید) ریشه کلمات (Stemming) را بدست آورید.

جهت نوشتن قوانین الگوریتم porter از فایل ضمیمه شده برای پیدا کردن قوانین استفاده شده. تابع apply_porter_stemming جهت چک کردن مرحله به مرحله تکست و عمل stemming بر روی متن مورد نظر. در قانون اول که شامل سه بخش میشود، در بخش اول کلمات جمع به ساده و در بخش دوم افعال به شکل ساده تغییر کردند و در بخش سوم حرف y اضافه در انتهای فعل را به i تغییر داده شده. در قانون دوم بیشتر suffixها را به کلمات ساده تبدیل شدند و عوض شدند در قانون سوم نیز به همین ترتیب یا به کلمات ساده تر تبدیل شدند یا حرف اضافه از آنها حذف شد. در قانون چهارم حرف اضافی ly که به عنوان متمم گاها درکنار کلمات قرار میگیرد حذف شده.

- جهت به کار گیری تابع نوشته شده، ابتدا توکن های ساخته شده در مرحله قبل را به تکست تبدیل شدند و بعد تابع را پیاده سازی کرده و پرینت میگیریم.
- رفرنس: https://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf

```
# Join tokens to form the text for stemming
text = " ".join(tokens)

# Apply stemming
stemmed_text = apply_porter_stemming(text)

print(stemmed_text)
```

assignment1

```
def apply_porter_stemming(text: Any) -> str
```

- Stemming function

```python
def apply_porter_stemming(text):
    # Rule 1a: Plurals
    text = re.sub(r'(sses)\b', r'ss', text)
    text = re.sub(r'(ies)\b', r'i', text)
    text = re.sub(r'([^s])s\b', r'\1', text)

    # Rule 1b: Past tense and gerund forms
    text = re.sub(r'(ed|ing)\b', '', text)

    # Rule 1c: Change 'y' to 'i' when preceded by a vowel
    text = re.sub(r'([aeiou])y\b', r'\1i', text)

    # Rule 2: Double suffixes
    text = re.sub(r'ization\b', 'ize', text)
    text = re.sub(r'ational\b', 'ate', text)
    text = re.sub(r'ation\b', 'ate', text)
    text = re.sub(r'ator\b', 'ate', text)
    text = re.sub(r'alism\b', 'al', text)
    text = re.sub(r'iveness\b', 'ive', text)
    text = re.sub(r'fulness\b', 'ful', text)
    text = re.sub(r'ousness\b', 'ous', text)
    text = re.sub(r'aliti\b', 'al', text)
    text = re.sub(r'iviti\b', 'ive', text)
    text = re.sub(r'biliti\b', 'ble', text)

    # Rule 3:
    text = re.sub(r'icate\b', 'ic', text)
    text = re.sub(r'ative\b', '', text)
    text = re.sub(r'alize\b', 'al', text)
    text = re.sub(r'iciti\b', 'ic', text)
    text = re.sub(r'ical\b', 'ic', text)
    text = re.sub(r'ful\b', '', text)
    text = re.sub(r'ness\b', '', text)

    # Rule 4: adverb suffix -ly
    text = re.sub(r'ly\b', '', text)

    return text
```

3. با استفاده از عبارات منظم متن را Nomalize کنید.

برای normalization چندین مرحله از روی رفرنس ذکر شده انجام شد.

- Lowercase
- Punctuation removal
- Remove extra whitespace
- Common contractions
- stop words (and, the, is, in, of, are)

```python
def normalization(text):
    # Step1: Lowercase
    text = text.lower()

    # Step2: Remove punctuation
    text = re.sub(r'[^\w\s.-]', '', text)

    # Step3: Remove extra whitespace
    text = text = re.sub(r'\s+', ' ', text).strip()

    # Step4: Handle common contractions
    contractions = {
        "don't": "do not",
        "doesn't": "does not",
        "can't": "cannot",
        "won't": "will not",
        "i'm": "i am",
        "you're": "you are",
        "we're": "we are",
        "they're": "they are",
        "isn't": "is not",
        "aren't": "are not",
        "wasn't": "was not",
        "weren't": "were not",
        "it's": "it is",
        "that's": "that is",
        "there's": "there is",
    }

    for contraction, expanded_form in contractions.items():
        text = re.sub(r'\b' + contraction + r'\b', expanded_form, text)
```

```python

    # Step5: Remove stop words
    tokens = text.split()
    stop_words = set(["and", "the", "is", "in", "of", "on", "an", "a"])
    tokens = [word for word in tokens if word not in stop_words]

    return text

```

در نهایت متن به شکل زیر قرار گرفت:

after world war ii the british great reduc the use of the full stop and other punctuate point after abbreviate in at least semi-formal writ while the american more readi kept such use until more recent and still maintain it more than briton . the classic example consider by their american counterpart quite curiou wa the maintenance of the internal comma in a british organisate of secret agent call the special operate executive s.o. e which i not found in histori written after about 1960 . but before that many briton were more scrupulou at maintain the french form . in french the period on follow an abbreviate if the last letter in the abbreviate i not the last letter of it antecedent m. i the abbreviate for monsieur while mme i that for madame . like many other cross-channel linguistic acquisition many briton readi took thi up and follow thi rule themselve while the american took a simpler rule and appli it rigorous . over the year however the lack of convention in some style guide ha made it difficult to determine which two-word abbreviate should be abbreviat with period and which should not . the u.s. media tend to use period in two-word abbreviate like unit state u.s. but not personal computer pc or television tv . many british public have gradual done awai with the use of period in abbreviate . minimize of punctuate in typewritten material became economical desirable in the 1960 and 1970 for the many user of carbon-film ribbon since a period or comma consum the same length of non-reusable expensive ribbon a did a capital letter . widespread use of electronic communic through mobile phone and the internet dur the 1990 allow for a mark rise in colloquial abbreviate . thi wa due large to increas popularity of textual communic service such a instant - and text messag . sms for instance support message length of 160 character at most us the gsm 03 . 38 character set . thi brevity gave rise to an informal abbreviate scheme sometime call textese with which 10 or more of the word in a typic sms message are abbreviat . more recent twitter a popular social network service began driv abbreviate use with 140 character message limit .

در این متن چندین ایراد بزرگ وجود دارد که به درستی کلمات ساده سازی نشده اند.

- چندین s، ed که به جهت تغییر اسم جمع حذف قرار بود بشوند باعث شدند کل کلمه اشتباه در نظر گرفته بشود.
- در lowercase کردن تمامی کلمات باعث میشود گاها برخی از اسامی اشتباه استنباط شوند.
- کلماتی که دارای – بودند معنای خود را از دست دادند.
- تبدیل کردن ing باعث شده است که کلمه کامل از بین برود.