Experiment No.2
Apply Tokenization on given English and Indian Language Text
Date of Performance:
Date of Submission:

Aim: Apply Tokenization on given English and Indian Language Text

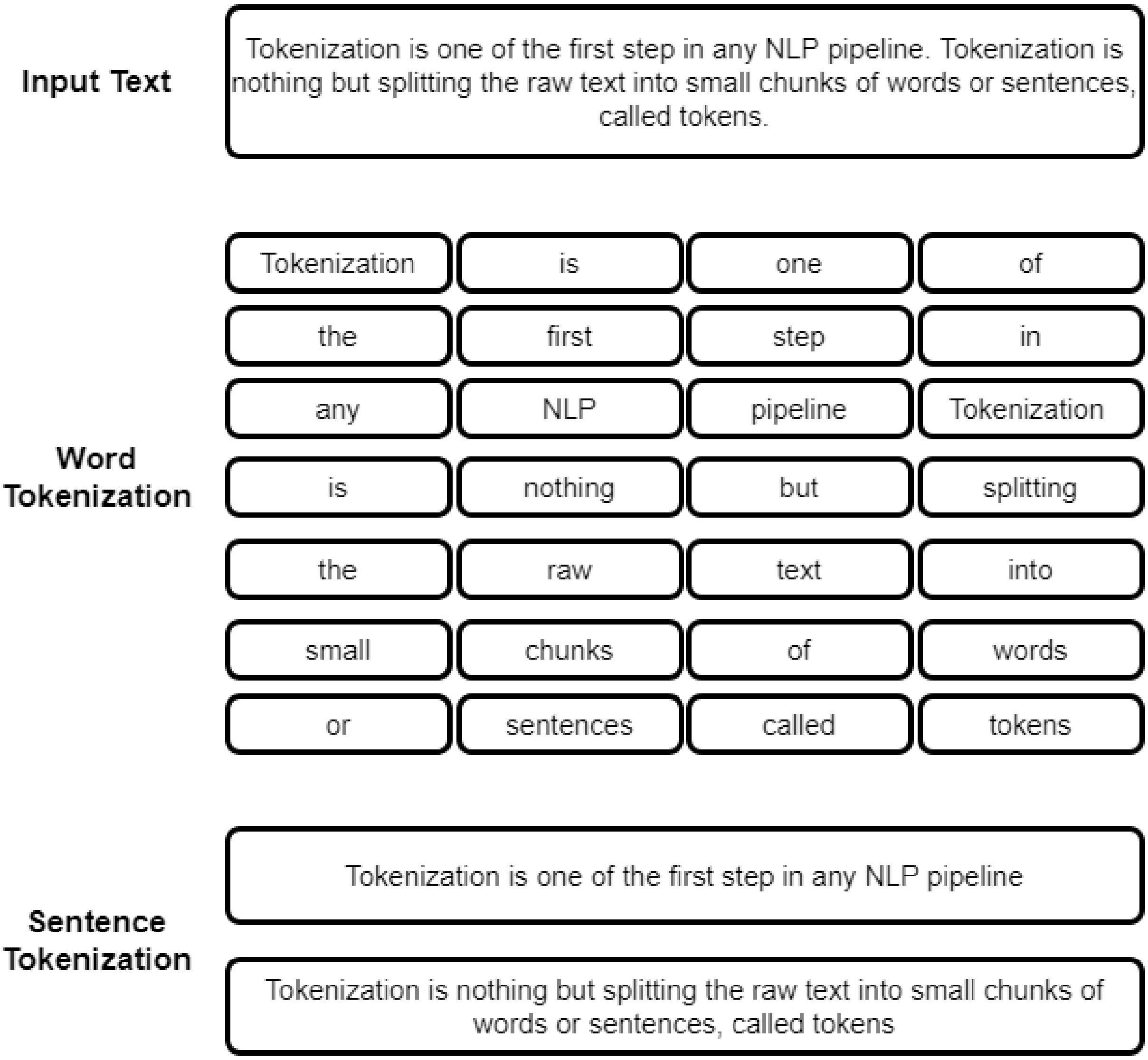
Objective: Able to perform sentence and word tokenization for the given input text for English and Indian Language.

Theory:

Tokenization is one of the first step in any NLP pipeline. Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens. If the text is split into words, then its called as 'Word Tokenization' and if it's split into sentences then its called as 'Sentence Tokenization'. Generally 'space' is used to perform the word tokenization and characters like 'periods, exclamation point and newline char are used for Sentence Tokenization. We have to choose the appropriate method as per the task in hand. While performing the tokenization few characters like spaces, punctuations are ignored and will not be the part of final list of tokens.

Why Tokenization is Required?

Every sentence gets its meaning by the words present in it. So by analyzing the words present in the text we can easily interpret the meaning of the text. Once we have a list of words we can also use statistical tools and methods to get more insights into the text. For example, we can use word count and word frequency to find out important of word in that sentence or document.



Input Text

Tokenization is one of the first step in any NLP pipeline. Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens.

Word Tokenization

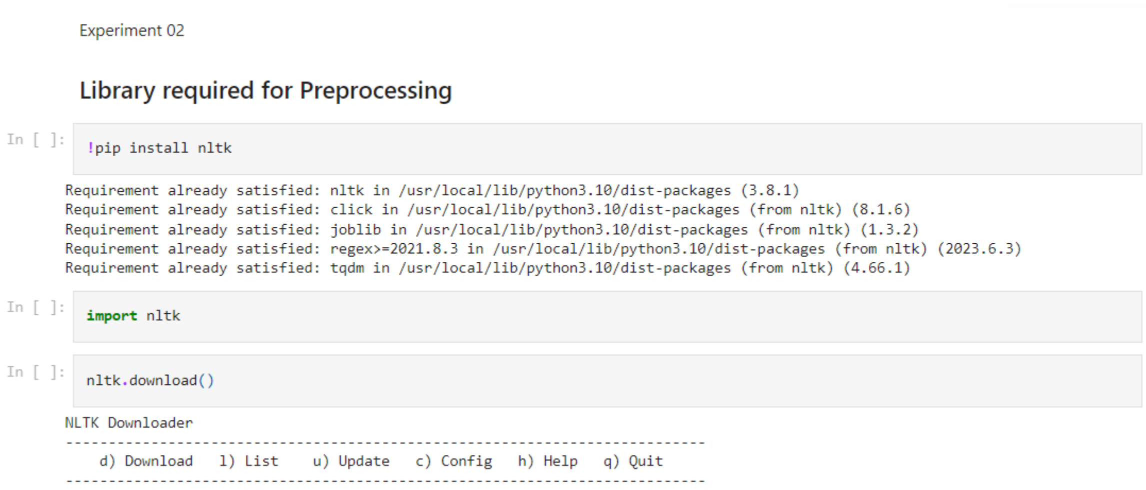
Tokenization	is	one	of
the	first	step	in
any	NLP	pipeline	Tokenization
is	nothing	but	splitting
the	raw	text	into
small	chunks	of	words
or	sentences	called	tokens

Sentence Tokenization

Tokenization is one of the first step in any NLP pipeline

Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens

Output:



Experiment 02

Library required for Preprocessing

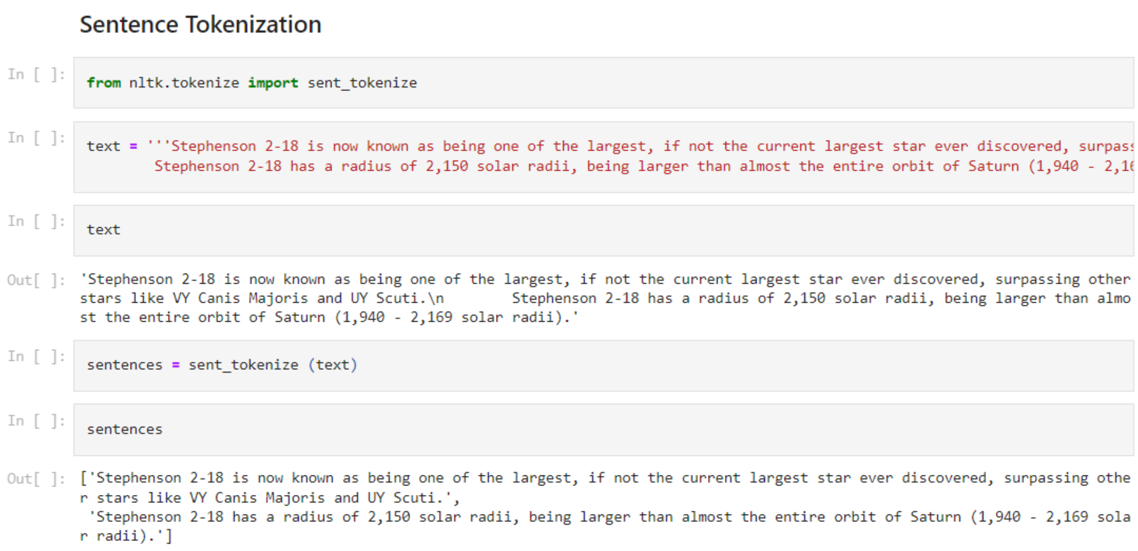
```
In [ ]: !pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.6)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
```

```
In [ ]: import nltk
```

```
In [ ]: nltk.download()
```

```
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
```



Sentence Tokenization

```
In [ ]: from nltk.tokenize import sent_tokenize
```

```
In [ ]: text = '''Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpass
                 Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the entire orbit of Saturn (1,940 - 2,16
```

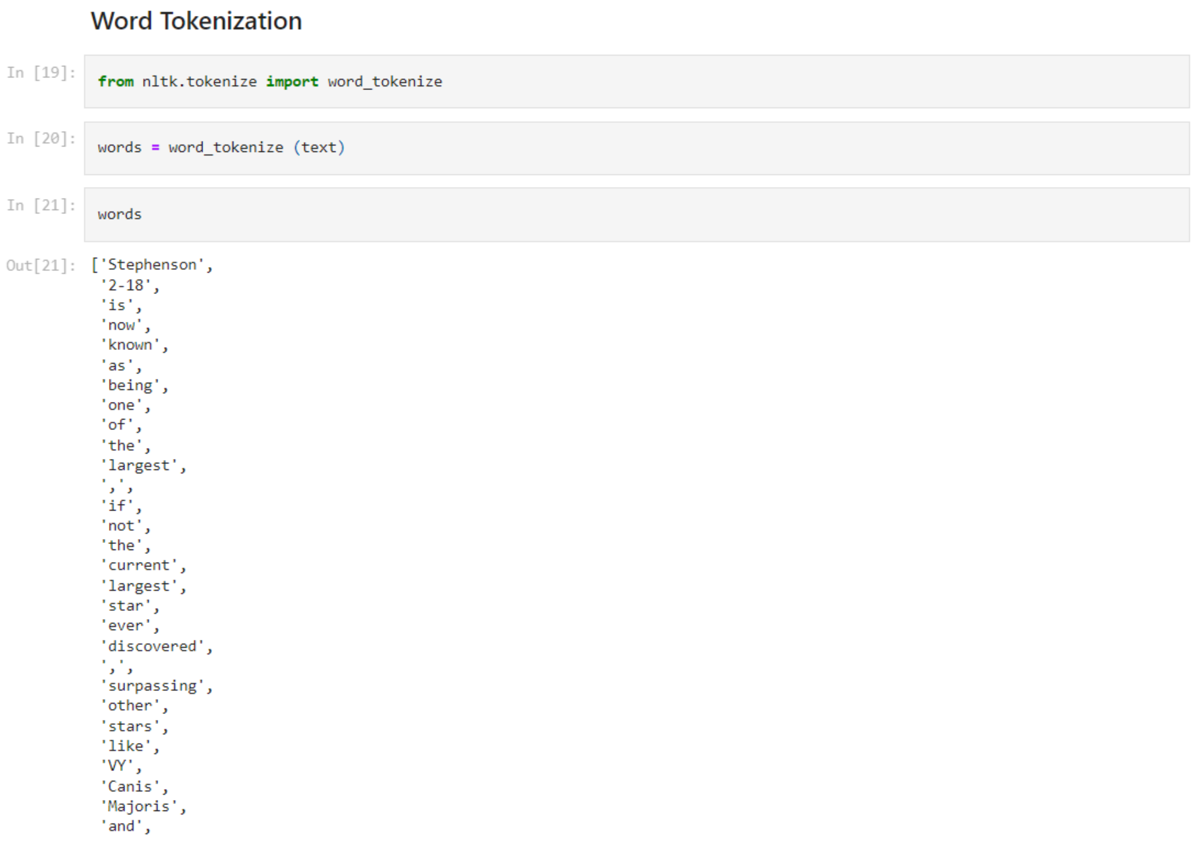
```
In [ ]: text
```

```
Out[ ]: 'Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing other
        stars like VY Canis Majoris and UY Scuti.\n         Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almo
        st the entire orbit of Saturn (1,940 - 2,169 solar radii).'
```

```
In [ ]: sentences = sent_tokenize (text)
```

```
In [ ]: sentences
```

```
Out[ ]: ['Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing othe
        r stars like VY Canis Majoris and UY Scuti.',
         'Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the entire orbit of Saturn (1,940 - 2,169 sola
        r radii).']
```



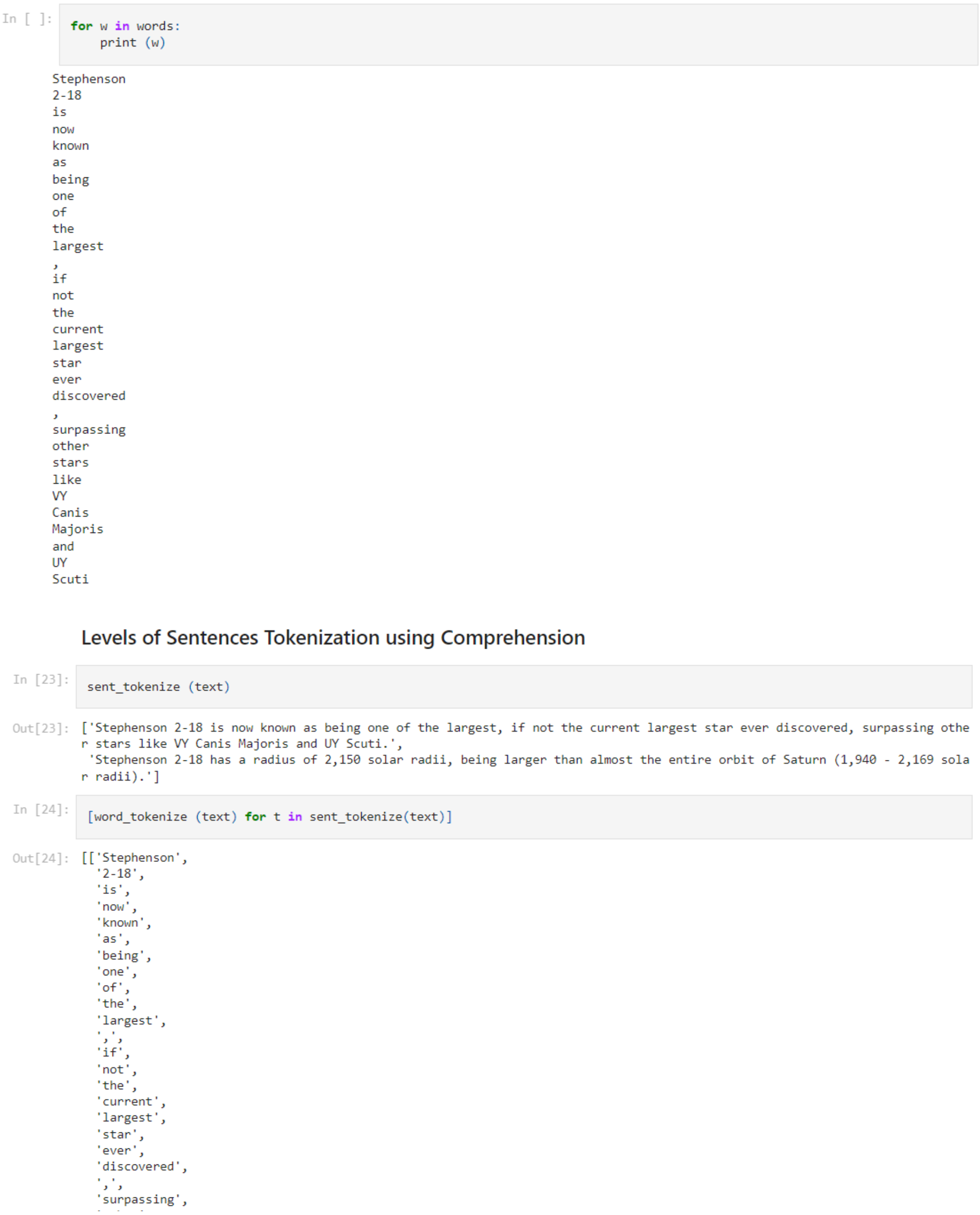
Word Tokenization

```
In [19]: from nltk.tokenize import word_tokenize
```

```
In [20]: words = word_tokenize (text)
```

```
In [21]: words
```

```
Out[21]: ['Stephenson',
          '2-18',
          'is',
          'now',
          'known',
          'as',
          'being',
          'one',
          'of',
          'the',
          'largest',
          ',',
          'if',
          'not',
          'the',
          'current',
          'largest',
          'star',
          'ever',
          'discovered',
          ',',
          'surpassing',
          'other',
          'stars',
          'like',
          'VY',
          'Canis',
          'Majoris',
          'and',
```



```
In [ ]:  for w in words:
             print (w)

         Stephenson
         2-18
         is
         now
         known
         as
         being
         one
         of
         the
         largest
         ,
         if
         not
         the
         current
         largest
         star
         ever
         discovered
         ,
         surpassing
         other
         stars
         like
         VY
         Canis
         Majoris
         and
         UY
         Scuti
```

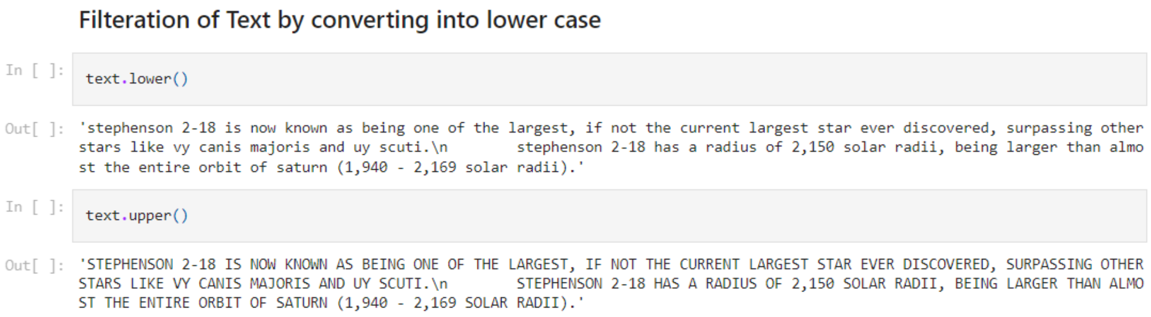
Levels of Sentences Tokenization using Comprehension

```
In [23]:  sent_tokenize (text)

Out[23]:  ['Stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing othe
          r stars like VY Canis Majoris and UY Scuti.',
           'Stephenson 2-18 has a radius of 2,150 solar radii, being larger than almost the entire orbit of Saturn (1,940 - 2,169 sola
          r radii).']

In [24]:  [word_tokenize (text) for t in sent_tokenize(text)]

Out[24]:  [['Stephenson',
            '2-18',
            'is',
            'now',
            'known',
            'as',
            'being',
            'one',
            'of',
            'the',
            'largest',
            ',',
            'if',
            'not',
            'the',
            'current',
            'largest',
            'star',
            'ever',
            'discovered',
            ',',
            'surpassing',
```



Filteration of Text by converting into lower case

```
In [ ]: text.lower()
```

```
Out[ ]: 'stephenson 2-18 is now known as being one of the largest, if not the current largest star ever discovered, surpassing other
        stars like vy canis majoris and uy scuti.\n        stephenson 2-18 has a radius of 2,150 solar radii, being larger than almo
        st the entire orbit of saturn (1,940 - 2,169 solar radii).'
```

```
In [ ]: text.upper()
```

```
Out[ ]: 'STEPHENSON 2-18 IS NOW KNOWN AS BEING ONE OF THE LARGEST, IF NOT THE CURRENT LARGEST STAR EVER DISCOVERED, SURPASSING OTHER
        STARS LIKE VY CANIS MAJORIS AND UY SCUTI.\n        STEPHENSON 2-18 HAS A RADIUS OF 2,150 SOLAR RADII, BEING LARGER THAN ALMO
        ST THE ENTIRE ORBIT OF SATURN (1,940 - 2,169 SOLAR RADII).'
```

Conclusion:

There are a number of tools available for tokenization of Indian language input. Some of the most popular tools include:

iNLTK: iNLTK is a Python library for natural language processing (NLP) in Indian languages. It includes a variety of NLP tools, including a tokenizer for Indian languages.

Mila NMT: Mila NMT is a machine translation toolkit that includes a tokenizer for Indian languages.

Indic NLP Library: The Indic NLP Library is a Python library for NLP in Indian languages. It includes a variety of NLP tools, including a tokenizer for Indian languages.

spaCy: spaCy is a Python library for NLP. It includes a tokenizer for Indian languages, but it is not as comprehensive as the other tools listed above.