Tokenize text using NLTK in python

To run the below python program, (NLTK) natural language toolkit has to be installed in your system.
The NLTK module is a massive tool kit, aimed at helping you with the entire Natural Language Processing (NLP) methodology.
In order to install NLTK run the following commands in your terminal.

The above installation will take quite some time due to the massive amount of tokenizers, chunkers, other algorithms, and all of the corpora to be downloaded.

Some terms that will be frequently used are 😗 **Corpus** – Body of text, singular. Corpora is the plural of this.

- **Lexicon** – Words and their meanings.
- **Token** – Each “entity” that is a part of whatever was split up based on rules. For examples, each word is a token when a sentence is “tokenized” into words. Each sentence can also be a token, if you tokenized the sentences out of a paragraph.

So basically tokenizing involves splitting sentences and words from the body of the text.

```
from  nltk.tokenize  import  sent_tokenize, word_tokenize

text  =  "Natural language processing (NLP) is a field "  +  \

    "of computer science, artificial intelligence "  +  \

    "and computational linguistics concerned with "  +  \

    "the interactions between computers and human "  +  \

    "(natural) languages, and, in particular, "  +  \

    "concerned with programming computers to "  +  \

    "fruitfully process large natural language "  +  \

    "corpora. Challenges in natural language "  +  \

    "processing frequently involve natural "  +  \

    "language understanding, natural language"  +  \

    "generation frequently from formal, machine"  +  \

    "-readable logical forms), connecting language "  +  \
    "and machine perception, managing human-"  +  \

    "computer dialog systems, or some combination "  +  \

    "thereof."

print``(sent_tokenize(text))

print``` (word_tokenize(text)) ``
```

OUTPUT

```
[‘Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora.’, ‘Challenges in natural language processing frequently involve natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), connecting language and machine perception, managing human-computer dialog systems, or some combination thereof.’]
[‘Natural’, ‘language’, ‘processing’, ‘(’, ‘NLP’, ‘)’, ‘is’, ‘a’, ‘field’, ‘of’, ‘computer’, ‘science’, ‘,’, ‘artificial’, ‘intelligence’, ‘and’, ‘computational’, ‘linguistics’, ‘concerned’, ‘with’, ‘the’, ‘interactions’, ‘between’, ‘computers’, ‘and’, ‘human’, ‘(’, ‘natural’, ‘)’, ‘languages’, ‘,’, ‘and’, ‘,’, ‘in’, ‘particular’, ‘,’, ‘concerned’, ‘with’, ‘programming’, ‘computers’, ‘to’, ‘fruitfully’, ‘process’, ‘large’, ‘natural’, ‘language’, ‘corpora’, ‘.’, ‘Challenges’, ‘in’, ‘natural’, ‘language’, ‘processing’, ‘frequently’, ‘involve’, ‘natural’, ‘language’, ‘understanding’, ‘,’, ‘natural’, ‘language’, ‘generation’, ‘(’, ‘frequently’, ‘from’, ‘formal’, ‘,’, ‘machine-readable’, ‘logical’, ‘forms’, ‘)’, ‘,’, ‘connecting’, ‘language’, ‘and’, ‘machine’, ‘perception’, ‘,’, ‘managing’, ‘human-computer’, ‘dialog’, ‘systems’, ‘,’, ‘or’, ‘some’, ‘combination’, ‘thereof’, ‘.’]
```

So there, we have created tokens, which are sentences initially and words later.