Tanay Jhawar
B.E. 4 Roll No.22

EXPERIMENT 2

AIM: Implementation word and sentence tokenization in Python.

SOURCE CODE:

English Language Tokenisation:

```
import re
from nltk.tokenize import word_tokenize,sent_tokenize
from pathlib import Path
text = Path('myfiles.txt').read_text()
tokens = re.findall("[\w']+", text)
print("Word Tokenization without NLTK:\n")
print(tokens)
print("\nWord TOkenization with NLTK:\n")
print(word_tokenize(text))
print("\nSentence Tokenization without NLTK\n")
sentences = text.split('. ')
print(sentences)
print("\nSentence Tokenization with NLTK\n")
print(sent_tokenize(text))
```

Hindi Language Tokenisation:

```
from inltk.inltk import tokenize
hindi_text = """प्राचीन काल में विक्रमादित्य नाम के एक आदर्श राजा हुआ करते थे।
अपने साहस, पराक्रम और शौर्य के लिए  राजा विक्रम मशहूर थे। """
tokenize(hindi_text, "hi")
from indicnlp.tokenize import sentence_tokenize
sentences=sentence_tokenize.sentence_split(hindi_text, lang='hi')
for t in sentences:
   print(t)
```

myfiles.txt

Founded in 2002, SpaceXs mission is to enable humans to become a spacefaring civilization and a multi-planet species by building a self-sustaining city on Mars. In 2008, SpaceXs Falcon 1 became the first privately developed liquid-fuel launch vehicle to orbit the Earth.

OUTPUT:

Fig1. Performing English Language Word and Sentence Tokenization with and without NLTK

Fig2. Performing Hindi Language Word and Sentence Tokenization