Atharva Prashant Pawar (9427) – Comps - A

NLP : Exp 1 : Post Lab DATE:

Q1. Differentiate between nltk & spacy

	Feature	NLTK	Spacy
1.	Functionality	Comprehensive NLP toolkit.	Streamlined for fast text processing.
2.	Performance	Moderate speed	Fast & efficient.
3.	Language Support	Wide variety of languages	Initial focus on English
4.	Ease of use:	More beginner-friendly	Straightforward API.
5.	Models & pipelines	Some pre-trained models available	Pre-trained models for NLP tasks.
6.	Development & Updates	Mature, with regular updates	Actively developed.

DATE:

Q2. Explore tokenization using keras & tokenization using gensim.

=> A. Tokenization using keras:

keras, a high-level deep learning library, provides the 'tokenizer' class to handle text tokenization. It allows you to convert text into sequences of integers, which can be fed into a neural n/w for further processing: !pip install tensorflow keras

eg:
```
from keras.preprocessing.text import Tokenizer
texts = ["Hello Friends", "Good Morning"]
tokenizer = Tokenizer()
tokenizer.fit_on_texts(texts)
sequences = tokenizer.texts_to_sequences(texts)
word_index = tokenizer.word_index
print(f"Sequences: {sequences}; Word Index: {word_index}")
```

B. Tokenization using Gensim:

Gensim is a powerful library for topic modelling & document similarity analysis. For tokenization, Gensim provides a simple utility function called "simple_preprocess". !pip install gensim

eg: Code:
```
from gensim.utils import simple_preprocess
texts = ["Hello Friends", "Good Morning"]

tokenized_texts = [simple_preprocess(text) for text in texts]

print(f"Tokenized Texts: {tokenized_texts}")
```