Lab 10 Write a Simple Streaming program in Spark to receive text data streams on a particular port, perform basic text cleaning (like white space removal, stop words removal, lemmatization, etc.), and print the cleaned text on the screen (o/p).

-> Install dependencies & download NLTK data:

text_cleaning_stream.py

```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import re
sc = SparkContext("local[2]", "TextCleaningStream")
ssc = StreamingContext(sc, 5)
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()
def clean_text(line):
    line = line.lower()
    cline = re.sub(r"[^a-z\s]", "", line)
    tokens = [word for word in line.split() if word not in stop_words]
    lemmatized = [lemmatizer.lemmatize(word) for word in tokens]
    return " ".join(lemmatized)
lines = ssc.socketTextStream("localhost", 9999)
cleaned = lines.map(clean_text)
cleaned.pprint()
ssc.start()
ssc.awaitTermination()
```

(entry.getValue());

o/p = The quick brown fox jumps over the lazy dog.

-> quick brown fox jump lazy dog.