EXPERIMENT-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. (Open Ended Question).

Code with Screenshots:

```
import org.apache.spark.SparkConf

import org.apache.spark.streaming.{Seconds, StreamingContext}

import org.apache.spark.streaming.dstream.DStream


object StreamingTextCleaner {


  val stopWords = Set(

    "a", "an", "the", "is", "are", "was", "were", "this", "that",

    "on", "in", "and", "or", "of", "to", "for", "with", "as", "by"

  )


  def simpleLemmatize(word: String): String = {

    if (word.endsWith("ing")) word.dropRight(3)

    else if (word.endsWith("ed")) word.dropRight(2)

    else word

  }


  def main(args: Array[String]): Unit = {

    if (args.length < 2) {
```

```
  System.err.println("Usage: StreamingTextCleaner <hostname> <port>")
  System.exit(1)
}

val hostname = args(0)
val port = args(1).toInt

val conf = new SparkConf().setAppName("StreamingTextCleaner").setMaster("local[*]")
val ssc = new StreamingContext(conf, Seconds(5))

val lines = ssc.socketTextStream(hostname, port)

val cleanedLines: DStream[String] = lines.map(_.toLowerCase)
  .map(_.trim)
  .map(line => line.replaceAll("[^a-zA-Z\\s]", ""))
  .flatMap(_.split("\\s+"))
  .filter(word => word.nonEmpty && !stopWords.contains(word))
  .map(simpleLemmatize)
  .filter(_.nonEmpty)
  .reduceByWindow(
    (a: String, b: String) => a + " " + b,
    Seconds(5),
    Seconds(5)
```

```
    )

    cleanedLines.print()

    ssc.start()
    ssc.awaitTermination()
  }
}
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.2.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.5.0)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from nltk) (4.67.1)
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
+------+
|word  |
+------+
|hello |
|hate  |
|hate  |
|love  |
|dont  |
|want  |
|cant  |
|put   |
|nobody|
|else  |
+------+
```