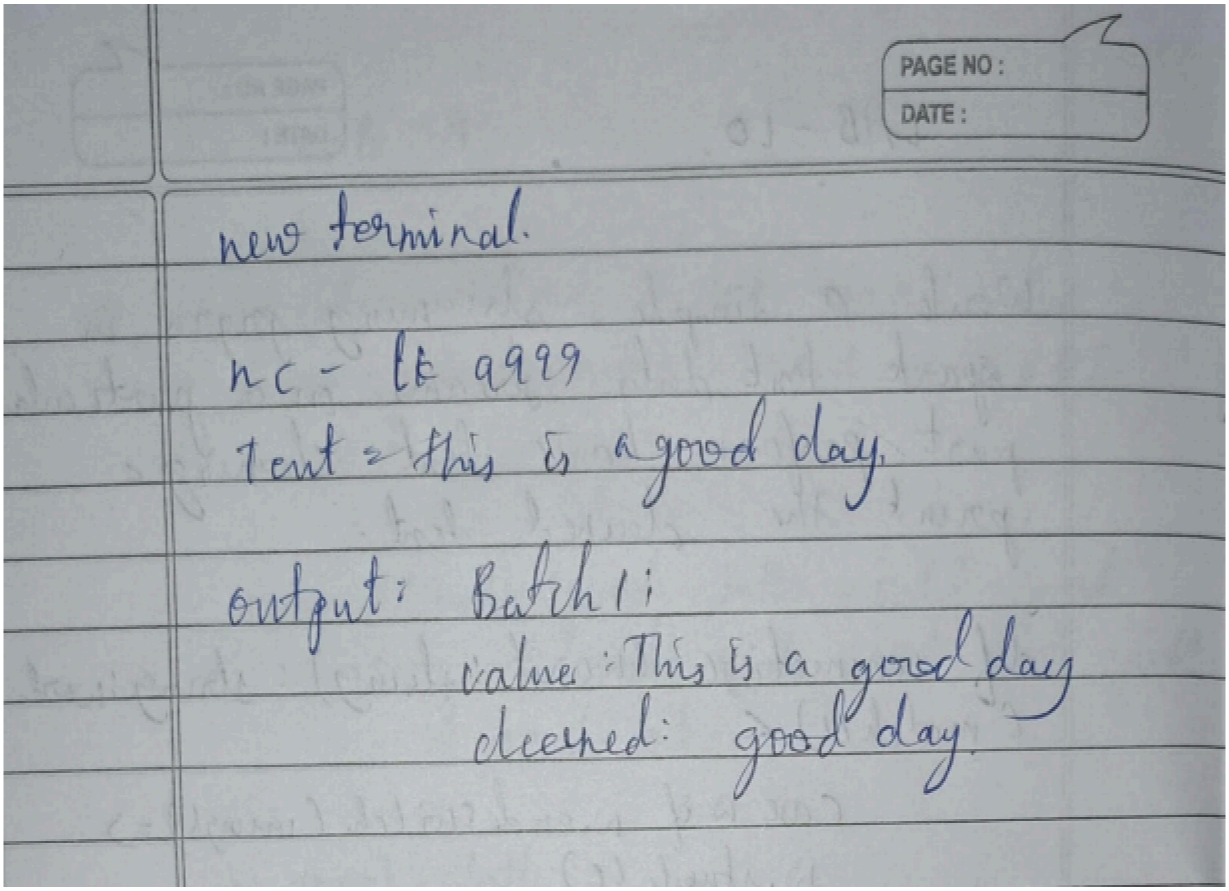
Lab10 Streaming Program in Spark

Code:

```
from pyspark import SparkConf, SparkContext
from pyspark.streaming import StreamingContext
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import nltk
import re

# Download required NLTK data files (only once)
nltk.download('stopwords')
nltk.download('wordnet')

# Initialize stop words and lemmatizer
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

# Function to clean and process text
def clean_text(line):
    # Lowercase
    line = line.lower()
    # Remove punctuation and digits
    line = re.sub(r'[^a-z\s]', '', line)
    # Remove extra spaces and split into words
    words = line.strip().split()
    # Remove stop words and apply lemmatization
    cleaned = [lemmatizer.lemmatize(word) for word in words if word not in stop_words]
    return cleaned

# Spark Configuration
conf = SparkConf().setAppName("TextCleaningStream").setMaster("local[2]")
sc = SparkContext(conf=conf)
ssc = StreamingContext(sc, 5)  # 5 second batch interval

# Create DStream from localhost on port 9999
lines = ssc.socketTextStream("localhost", 9999)

# Clean the text using flatMap
cleaned_words = lines.flatMap(clean_text)

# Print the cleaned words
cleaned_words.pprint()
```

```
# Start Streaming
ssc.start()
ssc.awaitTermination()
```

Output Screenshot:



PAGE NO :

DATE :

new terminal.

nc - lk 9999

Text = this is a good day.

output: Batch 1:
value : This is a good day
cleaned: good day.