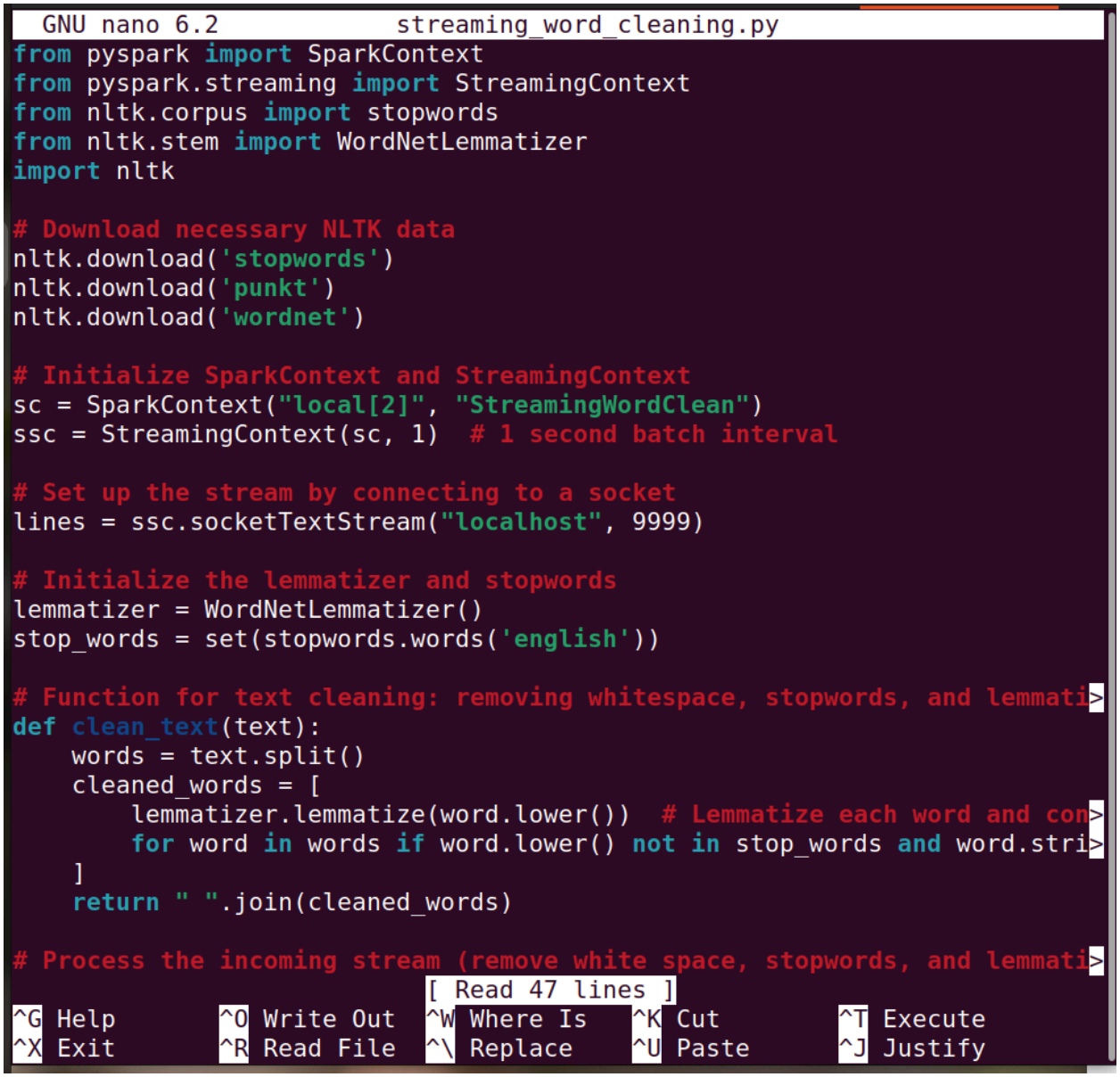
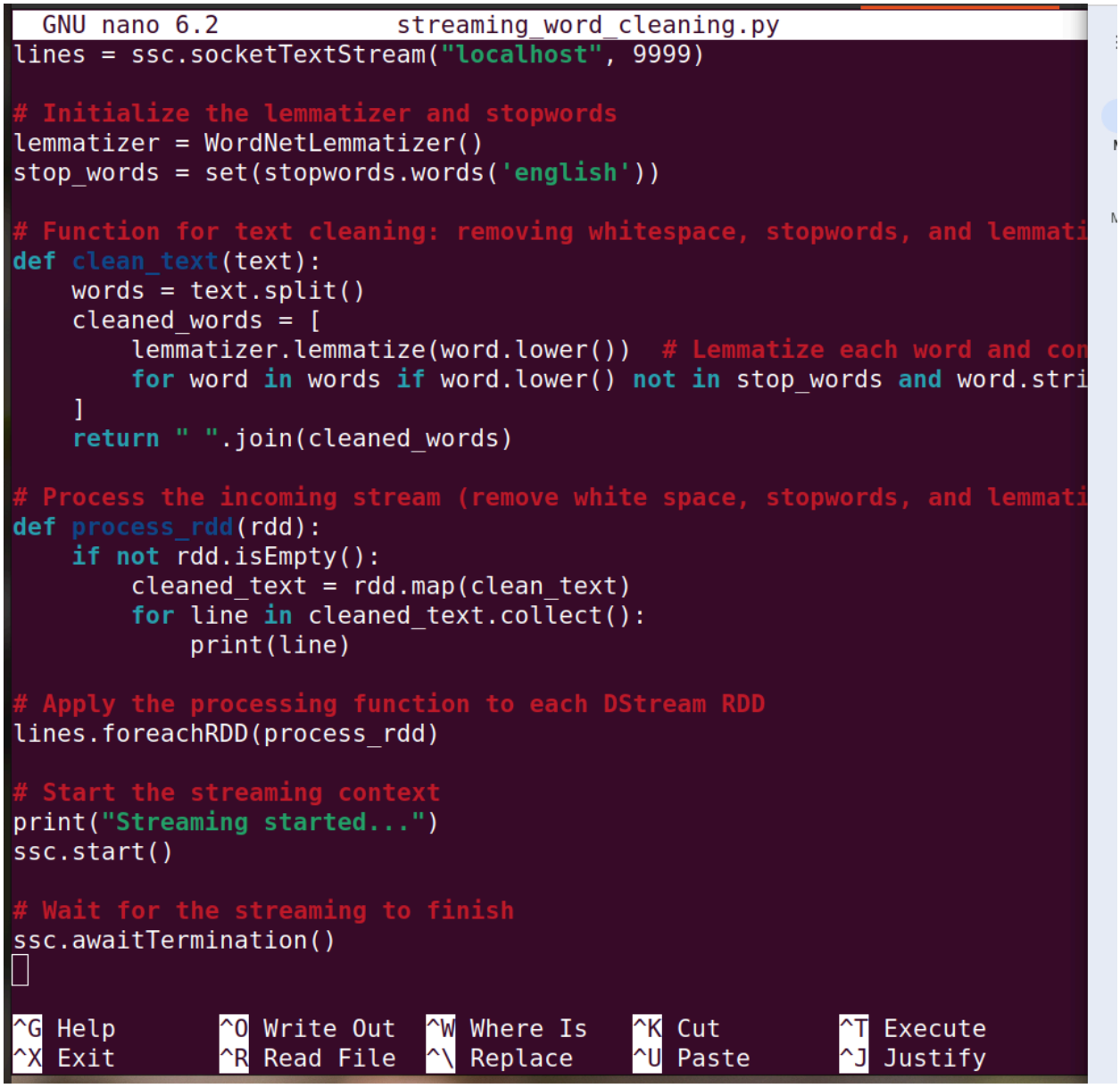

GNU nano 6.2 streaming_word_cleaning.py
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import nltk
# Download necessary NLTK data
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
# Initialize SparkContext and StreamingContext
sc = SparkContext("local[2]", "StreamingWordClean")
ssc = StreamingContext(sc, 1) # 1 second batch interval
# Set up the stream by connecting to a socket
lines = ssc.socketTextStream("localhost", 9999)
# Initialize the lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))
# Function for text cleaning: removing whitespace, stopwords, and lemmati>
def clean_text(text):
words = text.split()
cleaned_words = [
lemmatizer.lemmatize(word.lower()) # Lemmatize each word and con>
for word in words if word.lower() not in stop_words and word.stri>
]
return " ".join(cleaned_words)
Process the incoming stream (remove white space, stopwords, and lemmati>
[Read 47 lines]
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify


GNU nano 6.2 streaming_word_cleaning.py
lines = ssc.socketTextStream("localhost", 9999)
# Initialize the lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))
# Function for text cleaning: removing whitespace, stopwords, and lemmati
def clean_text(text):
words = text.split()
cleaned_words = [
lemmatizer.lemmatize(word.lower()) # Lemmatize each word and con
for word in words if word.lower() not in stop_words and word.stri
]
return " ".join(cleaned_words)
# Process the incoming stream (remove white space, stopwords, and lemmati
def process_rdd(rdd):
if not rdd.isEmpty():
cleaned_text = rdd.map(clean_text)
for line in cleaned_text.collect():
print(line)
# Apply the processing function to each DStream RDD
lines.foreachRDD(process_rdd)
# Start the streaming context
print("Streaming started...")
ssc.start()
# Wait for the streaming to finish
ssc.awaitTermination()
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify