Spark streaming - Multiple choice questions - Examples

Answer to the following questions. There is only one right answer for each question.

1. (2 points) Consider the following Spark Streaming applications.

(Application A)

```
from pyspark.streaming import StreamingContext

# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)

# Define a DStream associated with the TPC socket localhost:9999
# Apply window and map the input strings to integers
inputWindowDStream = ssc.socketTextStream("localhost", 9999)\
                .window(20, 10)\
                .map(lambda value: int(value))

# Sum values
sumWindowDStream = inputWindowDStream\
                .reduce(lambda v1, v2: v1 + v2)

# Apply a filter
resDStream = sumWindowDStream\
                .filter(lambda value: value > 5)

# Print the result on the standard output
resDStream.pprint()

# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

(Application B)

```
from pyspark.streaming import StreamingContext

# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)

# Define a DStream associated with the TPC socket localhost:999
# Map the input strings to integers
inputDStream = ssc.socketTextStream("localhost", 9999)\
            .map(lambda value: int(value))

# Sum values
sumDStream = inputDStream\
            .reduce(lambda v1, v2: v1 + v2)
```

```
# Define windows
sumWindowDStream = sumDStream\
            .window(20,10)

# Apply a filter
resDStream = sumWindowDStream\
            .filter(lambda value: value > 5)

# Print the result on the standard output
resDStream.pprint()

# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

(Application C)

```
from pyspark.streaming import StreamingContext

# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)

# Define a DStream associated with the TPC socket localhost:9999
# Map the input strings to integers
inputDStream = ssc.socketTextStream("localhost", 9999)\
            .map(lambda value: int(value))

# Define windows
inputWindowDStream = inputDStream\
            .window(20, 10)

# Sum values
sumWindowDStream = inputWindowDStream\
            .reduce(lambda v1, v2: v1 + v2)

# Apply a filter
resDStream = sumWindowDStream\
            .filter(lambda value: value > 5)

# Print the result on the standard output
resDStream.pprint()

# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

Which one of the following statements is true?Applications A, B, And C are equivalent in terms of returned result, i.e., given the same input they return the same result.

b) Applications A and B are equivalent in terms of returned result, i.e., given the same input they return the same result, while C is not equivalent to the other two applications.

c) Applications A and C are equivalent in terms of returned result, i.e., given the same input they return the same result, while B is not equivalent to the other two applications.

d) Applications B and C are equivalent in terms of returned result, i.e., given the same input they return the same result, while A is not equivalent to the other two applications.

2. (2 points) Consider the following Spark Streaming applications.

(Application A)

```
from pyspark.streaming import StreamingContext

# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)

# Define a DStream associated with the TPC socket localhost:9999
# Apply window and map the input strings to integers
inputWindowDStream = ssc.socketTextStream("localhost", 9999)\
                .window(20, 10)\
                .map(lambda value: int(value))

# Sum values
sumWindowDStream = inputWindowDStream\
                .reduce(lambda v1, v2: v1 + v2)

# Apply a filter
resDStream = sumWindowDStream\
                .filter(lambda value: value > 5)

# Print the result on the standard output
resDStream.pprint()

# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

(Application B)

```
from pyspark.streaming import StreamingContext
```

```
# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)


# Define a DStream associated with the TPC socket localhost:9999
# Map the input strings to integers
inputDStream = ssc.socketTextStream("localhost", 9999)\
                .map(lambda value: int(value))

# Sum values
sumDStream = inputDStream\
                .reduce(lambda v1, v2: v1 + v2)

# Define windows
sumWindowDStream = sumDStream\
                .window(20, 10)

# Apply a filter
resDStream = sumWindowDStream\
                .filter(lambda value: value > 5)

# Print the result on the standard output
resDStream.pprint()

# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

(Application C)

```
from pyspark.streaming import StreamingContext

# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)

# Define a DStream associated with the TPC socket localhost:9999
# Map the input strings to integers
inputDStream = ssc.socketTextStream("localhost", 9999)\
                .map(lambda value: int(value))

# Sum values
sumDStream = inputDStream\
                .reduce(lambda v1, v2: v1 + v2)

# Apply a filter
sumFilterDStream = sumDStream\
                .filter(lambda value: value > 5)

# Define windows
resDStream = sumFilterDStream\
                .window(20, 10)
```

```
# Print the result on the standard output
resDStream.pprint(

# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

Which one of the following statements is true?Applications A, B, And C are equivalent in terms of returned result, i.e., given the same input they return the same result.

b) Applications A and B are equivalent in terms of returned result, i.e., given the same input they return the same result, while C is not equivalent to the other two applications.

c) Applications A and C are equivalent in terms of returned result, i.e., given the same input they return the same result, while B is not equivalent to the other two applications.

d) Applications B and C are equivalent in terms of returned result, i.e., given the same input they return the same result, while A is not equivalent to the other two applications.

3. (2 points) Consider the following Spark Streaming application.

```
from pyspark.streaming import StreamingContext

# Create a Spark Streaming Context object
ssc = StreamingContext(sc, 10)

# Define a DStream associated with the TPC socket localhost:9999
# Map the input strings to integers
inputDStream = ssc.socketTextStream("localhost", 9999)\
                .map(lambda value: int(value))

# Sum values
sumDStream = inputDStream\
                .reduce(lambda v1, v2: v1 + v2)

# Define windows
resDStream = sumDStream
                .window(20 ,10)



# Print the result on the standard output
resDStream.pprint()
```

```
# Start the computation
ssc.start()
ssc.awaitTerminationOrTimeout(360)
ssc.stop(stopSparkContext=False)
```

Consider the following input data
Time: 1s -> “2“
Time: 3s -> “2”
Time: 5s -> “1”
Time: 12s -> “4”
Time: 14s -> “2”

Which one of the following statements is true?

a) The application, after 20 seconds, prints on the standard output the value 11.

b) The application, after 20 seconds, prints on the standard output the values 5 and 6.

c) The application, after 20 seconds, prints on the standard output the value 6.

d) The application, after 20 seconds, prints on the standard output the value 5.