Installation Guide

**Pyspark**

We used the following URL to install PySpark

1. <https://www.youtube.com/watch?v=t63PS3kiTTQ>
2. <http://deelesh.github.io/pyspark-windows.html>

Points to Mention

1. This requires setting environment variables
2. Here Pyspark was installed on the Local System
3. Various other approaches would involve doing it on a virtual machine or on cloud (AWS or Google Cloud)

**Apache Kafka**

The set up steps were followed from the following link

1. <https://kafka.apache.org/0110/documentation/streams/quickstart>

**Points to mention**

1. The above link was extremely helpful as it also allows to check simple message transfers from producer to consumer to test the configuration
2. The commands in Commands.txt is used for Windows. Ou can refer to the Kafka link mentioned above to run on Ubuntu

**References**

1. <https://www.rittmanmead.com/blog/2017/01/getting-started-with-spark-streaming-with-python-and-kafka/>
2. <https://spark.apache.org/docs/latest/streaming-programming-guide.html>
3. <https://hortonworks.com/hadoop-tutorial/introduction-spark-streaming/>
4. <http://i.stanford.edu/~ullman/mmds/ch2n.pdf>
5. <https://www.rittmanmead.com/blog/2017/01/getting-started-with-spark-streaming-with-python-and-kafka/>