INSTRUCTION or README FILE

1. **About the Project**

   This was a twitter sentimental analysis project using the data from Kaggle dataset. This is actually a natural language processing project using pyspark and spark MLlib. The project was implemented on Hadoop cluster and stored the data set on HDFS storage cluster.

   The original data is on this website https://www.kaggle.com/datasets/kazanova/sentiment140, However it was modified a bit by dropping unwanted columns. The whole dataset was saved on Haddop as "trainB", but it was divided into train dataset as trainB1 and test dataset as testB1. Clear details are in the project itself.

2. **Instructions on how to run the project**

   The project can be run by taking the whole Jupiter notebook and open it on Hadoop cluster under Jupiter Hub or copy the contents into a pyspark Jupiter notebook in the cluster. Then run it as a normal notebook. Make sure that there is "WARN" not "ERROR" under ……, otherwise it will not run.