-- Load data from HDFS

inputDialogues4 = LOAD 'hdfs:///user/saahil/inputs/episodeIV\_dialogues.txt' USING PigStorage('\t') AS (name:chararray, line:chararray);

inputDialogues5 = LOAD 'hdfs:///user/saahil/inputs/episodeV\_dialogues.txt' USING PigStorage('\t') AS (name:chararray, line:chararray);

inputDialogues6 = LOAD 'hdfs:///user/saahil/inputs/episodeVI\_dialogues.txt' USING PigStorage('\t') AS (name:chararray, line:chararray);

-- Filter out the first 2 lines from each file

ranked4 = RANK inputDialogues4;

OnlyDialogues4 = FILTER ranked4 BY (rank\_inputDialogues4 > 2);

ranked5 = RANK inputDialogues5;

OnlyDialogues5 = FILTER ranked5 BY (rank\_inputDialogues5 > 2);

ranked6 = RANK inputDialogues6;

OnlyDialogues6 = FILTER ranked6 BY (rank\_inputDialogues6 > 2);

-- Merge the three inputs

inputData = UNION OnlyDialogues4, OnlyDialogues5, OnlyDialogues6;

-- Group by name

groupByName = GROUP inputData BY name;

-- Count the number of lines by each character

names = FOREACH groupByName GENERATE $0 as name, COUNT($1) as no\_of\_lines;

namesOrdered = ORDER names BY no\_of\_lines DESC;

-- Remove the outputs folder

rmf hdfs:///user/sona/outputs;

-- Store result in HDFS

STORE namesOrdered INTO 'hdfs:///user/sona/outputs' USING PigStorage('\t');