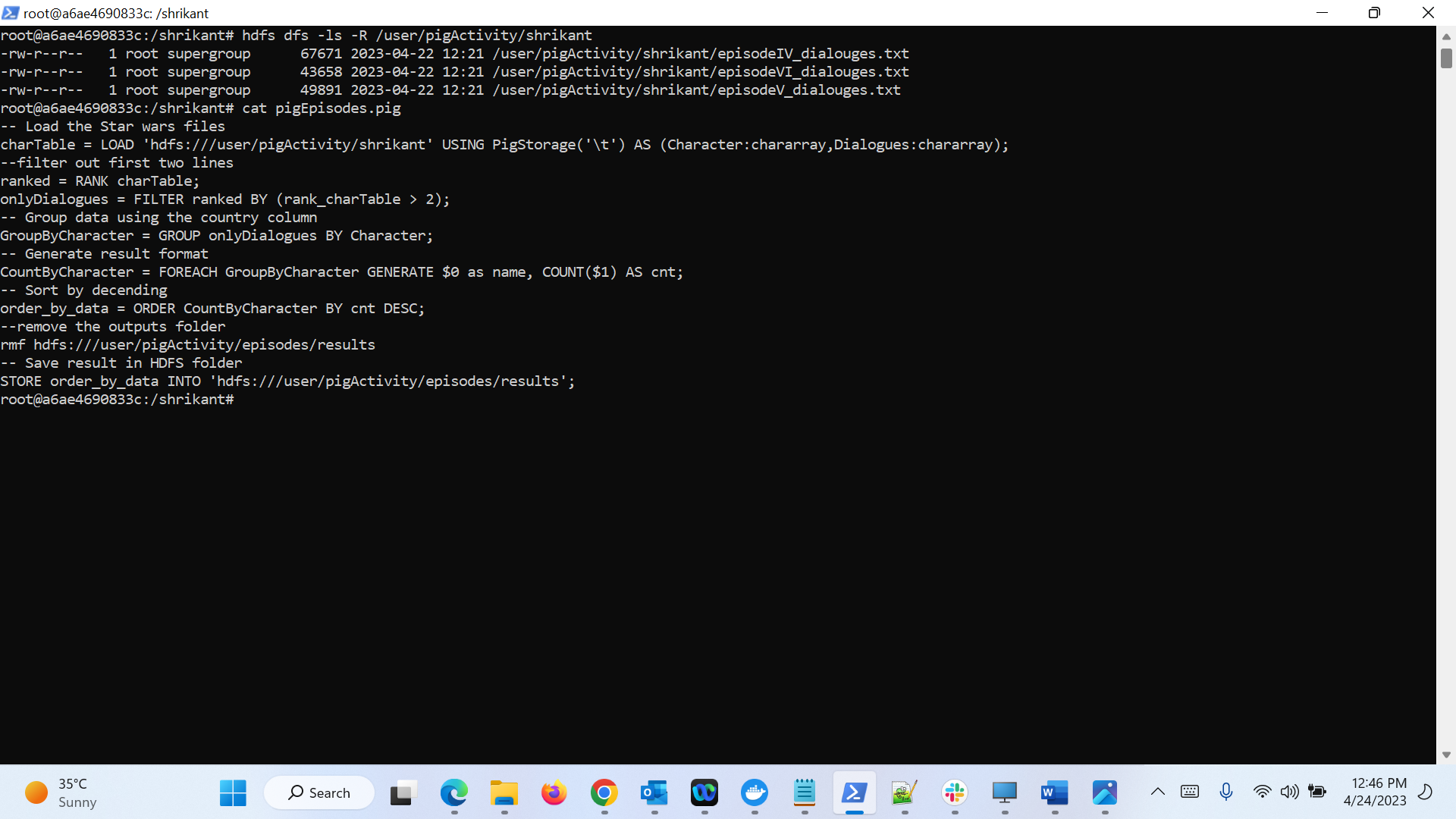
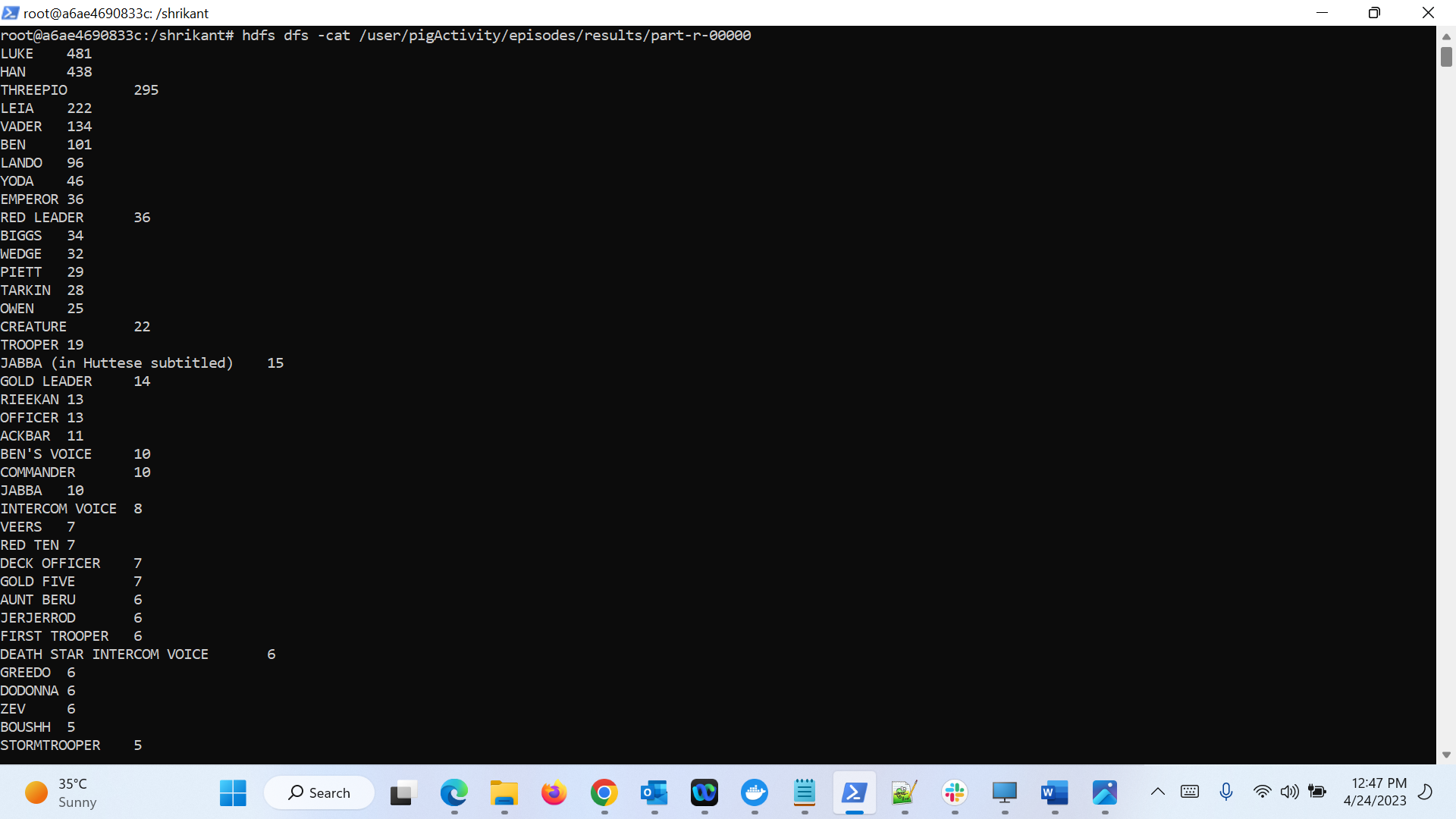
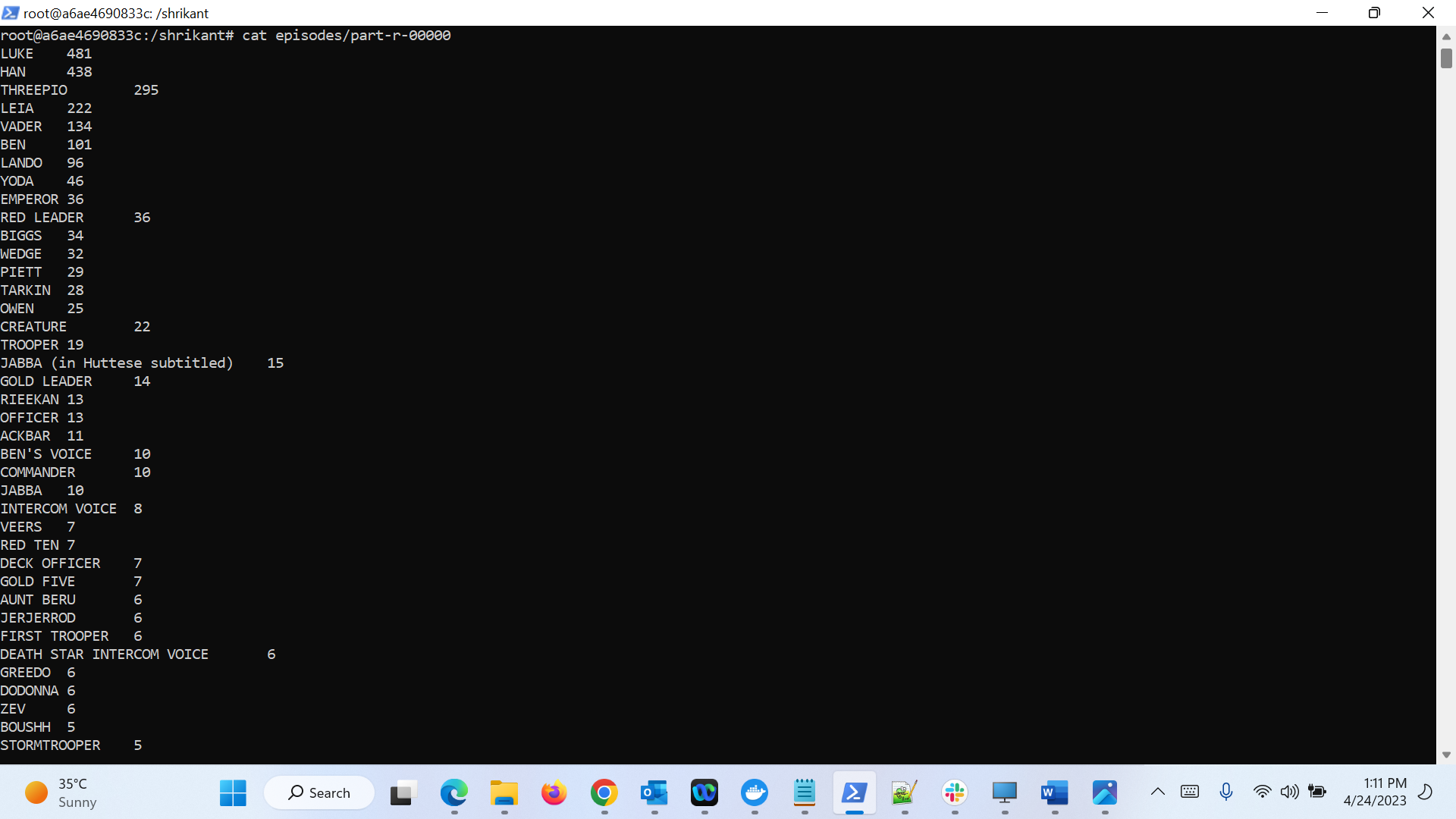
Pig file



Result of episodes\_dialouges in HDFS



Result of episodeIV\_dialouges in Local 

-- Load data from HDFS

ep4 = LOAD 'hdfs:///user/pigActivity/episodeIV\_dialouges.txt' USING PigStorage('\t') AS (characher:chararray, line:chararray);

ep5 = LOAD 'hdfs:///user/pigActivity/episodeV\_dialouges.txt' USING PigStorage('\t') AS (characher:chararray, line:chararray);

ep6 = LOAD 'hdfs:///user/pigActivity/episodeVI\_dialouges.txt' USING PigStorage('\t') AS (characher:chararray, line:chararray);

-- Filter out the first 2 lines

ep4\_ranked = RANK ep4;

ep4\_OnlyDialouges = FILTER ep4\_ranked BY (rank\_ep4 > 2);

ep5\_ranked = RANK ep5;

ep5\_OnlyDialouges = FILTER ep5\_ranked BY (rank\_ep5 > 2);

ep6\_ranked = RANK ep6;

ep6\_OnlyDialouges = FILTER ep6\_ranked BY (rank\_ep6 > 2);

-- Combine the 3 episodes

dialogues = UNION ep4\_OnlyDialouges, ep5\_OnlyDialouges, ep6\_OnlyDialouges;

-- Group by characher

groupByName = GROUP dialogues BY characher;

-- Count the number of lines by each character

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

charachersOrdered = ORDER charachers BY no\_of\_lines DESC;

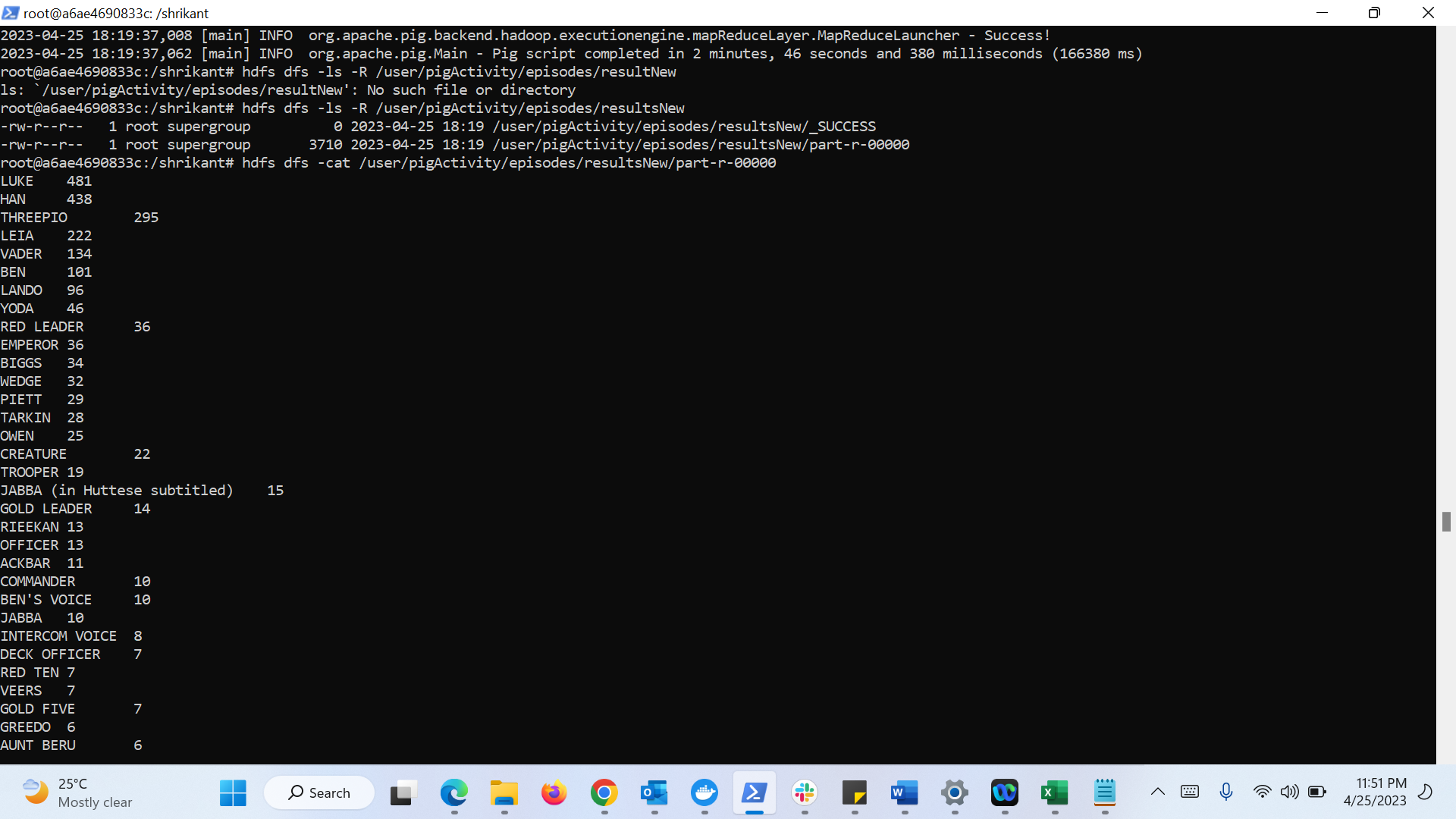
-- Remove the outputs folder

rmf hdfs:///user/ pigActivity/episodes/resultsNew

-- Store result in HDFS

STORE charachersOrdered INTO 'hdfs:///user/ pigActivity/episodes/resultsNew' USING PigStorage('\t');

HDFS



local 