root@e8e749c6171a:/# vim episodeIV\_dialogues.txt

root@e8e749c6171a:/# vim episodeV\_dialogues.txt

root@e8e749c6171a:/# vim episodeVI\_dialogues.txt

root@e8e749c6171a:/# hdfs dfs -mkdir /user/root/bhu

root@e8e749c6171a:/# hdfs dfs -put ./episodeIV\_dialogues.txt /user/root/bhu

root@e8e749c6171a:/# hdfs dfs -put ./episodeV\_dialogues.txt /user/root/bhu

root@e8e749c6171a:/# hdfs dfs -put ./episodeVI\_dialogues.txt /user/root/bhu

A screen shot of a computer

Description automatically generated

-- Load the data from HDFS

file1 = LOAD 'hdfs:///user/root/bhu/episodeIV\_dialogues.txt' AS (line:chararray);

file2 = LOAD 'hdfs:///user/root/bhu/episodeV\_dialogues.txt' AS (line:chararray);

file3 = LOAD 'hdfs:///user/root/bhu/episodeVI\_dialogues.txt' AS (line:chararray);

-- Concatenate the three files

allFiles = UNION file1, file2, file3;

-- Tokenize the lines into words

words = FOREACH allFiles GENERATE FLATTEN(TOKENIZE(line)) AS word;

-- Define characters and filter lines spoken by characters

characters = FILTER words BY word MATCHES '^(character1|character2|character3)$';

-- Group by character and count the lines

characterLines = GROUP characters BY word;

characterLineCount = FOREACH characterLines GENERATE group AS character, COUNT(characters) AS line\_count;

-- Output the results

DUMP characterLineCount;

STORE characterLineCount INTO 'dialoguesOutput' USING PigStorage('\t');

A screen shot of a computer

Description automatically generated

A computer screen with white text

Description automatically generated

A screen shot of a computer

Description automatically generated

A black screen with white text

Description automatically generated