Ex. No: 08

Date :

**Finding TF - IDF**

**Aim:**

To find the Term Frequency and Inverse Document Frequency (TF-IDF) of the book dataset using Pig Latin script.

**Procedure:**

1. Install the latest version of Pig (pig 17.0).
2. Download the book dataset or create your own dataset.
3. Create a file book.pig in home and write the scripts for the following actions.
4. Load the dataset to a relation by using LOAD command.
5. Split the datas into tokens by TOKENIZE command.
6. Then compute the document frequency and inverse document frequency.
7. Combine the document frequency and inverse document frequency.
8. Store the output in output directory by STORE command.

**Script:**

Save this commands in book.pig file.

*-- Load the book dataset*

A = LOAD 'book\_dataset.txt' USING TextLoader AS (line:chararray);

*-- Tokenize the documents into words*

B = FOREACH A GENERATE FLATTEN(TOKENIZE(REPLACE(LOWER(line),'[^a-zA-Z0-9\\s]',''))) AS word, line;

*-- Compute the term frequency (TF) for each word in each document*

WC = GROUP B BY (word,line);

WC\_TF = FOREACH WC GENERATE group.word AS word, group.line AS line, COUNT(B) AS tf;

*-- Compute the document frequency (DF) for each word*

W\_DF = FOREACH (GROUP WC\_TF BY word) GENERATE group AS word, COUNT(WC\_TF) AS df;

*-- Compute the number of unique documents*

NUM\_DOC = FOREACH (GROUP B ALL) GENERATE COUNT(B) AS NUMDOC;

*-- Compute the inverse document frequency (IDF) for each word*

word\_idf = FOREACH W\_DF GENERATE word, LOG((double)(NUM\_DOC.NUMDOC) / (double)df) AS idf;

*-- Join TF and IDF to calculate TF-IDF for each word in each document*

W\_TF\_IDF = JOIN WC\_TF BY word LEFT OUTER, word\_idf BY word;

TF\_IDF = FOREACH W\_TF\_IDF GENERATE WC\_TF::line AS A, WC\_TF::word AS word, WC\_TF::tf\*word\_idf::idf AS tf\_idf;

*-- Group TF-IDF values by document and store the result*

GROUPED\_TF\_IDF = GROUP TF\_IDF BY A;

FINAL\_TF\_IDF = FOREACH GROUPED\_TF\_IDF GENERATE group AS A, TF\_IDF;

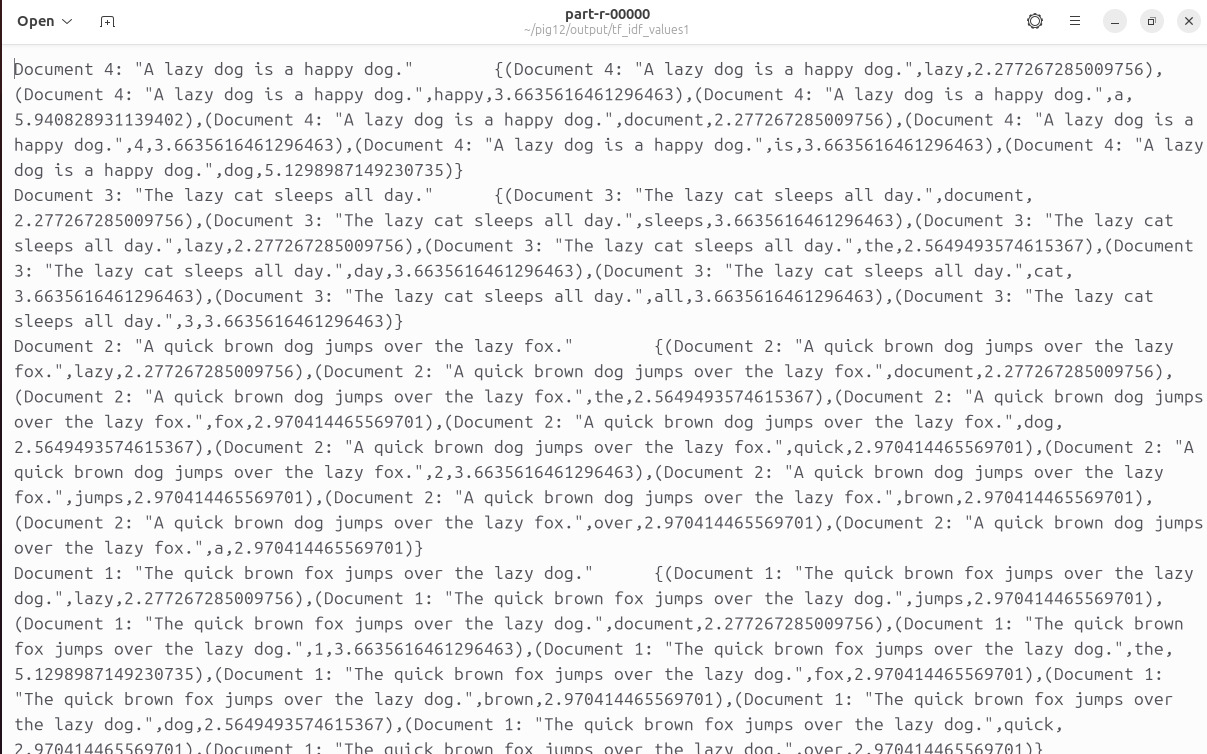
*-- Store the result*

STORE FINAL\_TF\_IDF INTO 'output/bookoutput' USING PigStorage();

**Execution:**

grunt > exec Book.pig

**Output:**



**Result :**

Thus the script for finding TF-IDF is written in Pig Latin and executed successfully.