# Assignment 8.1

# Write a word count program to implement word count using Pig.

A = load '/home/ec2-user/news/news\_1488746642.txt' using PigStorage() as (text:chararray);

words = FOREACH A GENERATE FLATTEN(TOKENIZE(REPLACE(LOWER(TRIM(text)),'[\\p{Punct},\\p{Cntrl}]',''))) AS word;

words\_grouped = GROUP words by word;

## to get count of each word

======================

cnt\_words = foreach words\_grouped generate group, COUNT(words) as cnt;

## total number of words

========================

word\_cnts = foreach cnt\_words generate cnt;

word\_cnts\_grp = group word\_cnts all;

total\_num\_words = foreach word\_cnts\_grp generate SUM(word\_cnts);

## to get total number of distinct words

=======================

distinct\_words = DISTINCT words;

grp = group distinct\_words all;

total\_dist = foreach grp generate COUNT(distinct\_words);