



Министерство науки и высшего образования Российской Федерации  
Калужский филиал федерального государственного автономного  
образовательного учреждения высшего образования  
«Московский государственный технический университет  
имени Н.Э. Баумана  
(национальный исследовательский университет)»  
(КФ МГТУ им. Н.Э. Баумана)

---

ФАКУЛЬТЕТ ИУК Информатика и управление

КАФЕДРА ИУК4 Программное обеспечение ЭВМ, информационные технологии

## ДОМАШНЯЯ РАБОТА 2

«»

по дисциплине: «Технологии обработки больших данных»

Выполнил: студент группы ИУК4-72Б

\_\_\_\_\_  
(Подпись)

Губин Е.В.

\_\_\_\_\_  
(И.О. Фамилия)

Проверил:

\_\_\_\_\_  
(Подпись)

Голубева С.Е.

\_\_\_\_\_  
(И.О. Фамилия)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:

- Оценка:

Калуга, 2025

**Цель:** формирование практических навыков реализации pig-скриптов для обработки больших данных.

**Задачи:**

1. Получить навыки обработки больших данных, используя Pig Latin.
2. Изучить принцип работы Pig Latin.
3. Изучить синтаксис Pig Latin.
4. Уметь писать запросы, комбинируя несколько источников данных.

**Вариант №9**

**Задание 1:**

Выполнить задание из лабораторной работы №2, используя язык Pig Latin:

Построить индекс файла. Для каждого слова в файле результат должен содержать номера всех строк, в которых появляется данное слово. Индекс должен быть регистро-независимым. Результат должен быть сохранен в файле в виде:  
((word1 (1 42 58)), (word2 (34 55 776 3456), ...)).

**Задание 2:**

База данных твитов состоит из двух файлов:

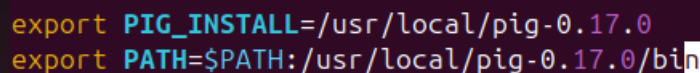
Файл tweets.csv имеет формат: tweet\_id, tweet, login

Файл users.csv имеет формат: login, user\_name, state

Необходимо создать эти файлы и выполнить задание по варианту, используя Pig Latin:

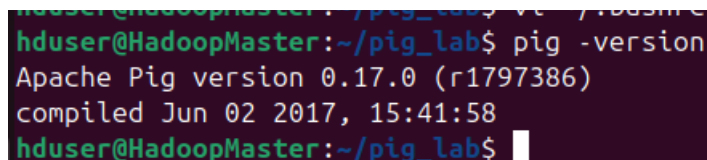
Найти всех пользователей, написавших менее 3 твитов. Подсчитать общее количество твитов, написанных этими пользователями. Вычислить долю, которую составляют эти твиты от общего количества твитов в базе.

**Ход выполнения работы:**



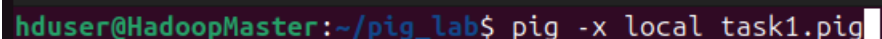
```
export PIG_INSTALL=/usr/local/pig-0.17.0
export PATH=$PATH:/usr/local/pig-0.17.0/bin
```

Рисунок 1 Переменные среды для Pig Latin



```
hduser@HadoopMaster:~/pig_lab$ pig -version
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
hduser@HadoopMaster:~/pig_lab$
```

Рисунок 2 Проверка версии pig'



```
hduser@HadoopMaster:~/pig_lab$ pig -x local task1.pig
```

Рисунок 3 Пример запуска скрипта

```
{(2556),(2553),(4459),(1073),(3466),(1074),(1060),(1116),(1059),(3537),(1056),(3455),(3517),(3542),(3517),(1105),
(3451),(3545),(1132),(1076),(3444),(3443),(1042),(1076),(3564),(1135),(3510),(1081),(3568),(3484),(1138),(1141),(3429),
(3488),(1142),(1145),(3575),(3506),(3576),(3506),(3424),(1027),(1149),(3423),(3421),(3491),(3409),(1022),(3409),(1021),(
3404),(3403),(3598),(3603),(1096),(1093),(1163),(1012),(3497),(3608),(3609),(1095),(1009),(1008),(3396),(2236),(28),(339
4),(3393),(1005),(4851),(3393),(1003),(1001),(3392),(3389),(999),(28),(3389),(996),(29),(3622),(3622),(3385),(992),(3631
),(1180),(2177),(3633),(3376),(3375),(3636),(3369),(2248),(2175),(1191),(1192),(3366),(1193),(1196),(2250),(2174),(979),
(2173),(38),(1202),(42),(4825),(3351),(1208),(3652),(2254),(2258),(2259),(3343),(968),(967),(3332),(966),(965),(4817),(3
324),(3665),(3323),(45),(3668),(3323),(960),(958),(2167),(3672),(3673),(1224),(953),(949),(51),(3318),(2270),(1228),(331
1),(3310),(3310),(4800),(3684),(2156),(3306),(3685),(3302),(2154),(3301),(3688),(3688),(57),(3689),(2275),(3298),(2151),
(935),(924),(3700),(3289),(3287),(922),(2143),(1252),(3705),(3707),(3283),(918),(3280),(1256),(3279),(917),(3714),(2142),
(3274),(3271),(3270),(906),(3267),(3727),(3727),(3264),(4776),(2189),(3259),(3258),(1272),(4775),(1281),(3736),(4774),(
895),(3736),(3254),(1284),(893),(1288),(2289),(3251),(1289),(4773),(3248),(886),(4772),(2290),(69),(2290),(71),(73),(375
8),(73),(5),(1294),(1295),(3239),(4770),(1295),(1297),(3236),(877),(3764),(1301),(4768),(3234),(873),(77),(870),(4768),(
870),(869),(3773),(4766),(3228),(1309),(1309),(3780),(856),(3224),(2298),(3784),(856),(1310),(3791),(2303),(852),(852),(
3795),(1318),(3796),(847),(2308),(3215),(2127),(4755),(2309),(1323),(843),(2309),(3211),(842),(91),(841),(3208),(4749),(
3208),(4748),(4747),(2315),(839),(1327),(3206),(3205),(3205),(3814),(830),(1109),(3196),(3193),(3817),(94),(3818),
(3819),(3187),(3825),(4732),(823),(3180),(2318),(823),(3176),(3834),(2115),(819),(819),(818),(816),(4724),(4724),
(3841),(99),(4724),(3841),(4723),(3161),(1357),(813),(813),(1362),(4719),(4718),(105),(810),(3852),(2324),(1375),(4715),
(4715),(3143),(111),(3522),(802),(3143),(3142),(3141),(1382),(2107),(3137),(1391),(112),(3865),(2332),(2107),(1392),(11
5),(2333),(116),(2106),(798),(3132),(117),(1399),(2227),(3131),(4702),(792),(791),(3878),(4698),(2337),(4694),(121),(233
8),(1407),(121),(2102),(1408),(4690),(3890),(4689),(4689),(4689),(4689),(781),(4685),(1414),(4685),(2097),(1417),(3109),
(3108),(2346),(3107),(127),(1418),(2093),(3104),(2093),(3908),(3100),(3098),(1422),(1422),(1422),(1426),(760),(4673),(14
26),(1427),(3091),(1427),(1431),(136),(3928),(757),(756),(3085),(2087),(752),(3081),(1438),(751),(747),(3078),(3077),(74
4),(3077),(15),(743),(3072),(1450),(739),(3068),(2361),(738),(736),(3960),(2077),(3961),(141),(3962),(4658),(1453),(2366),
(1454),(3056),(3053),(2370),(145),(4459),(3052),(4655),(3051),(1457),(1462),(1462),(2375),(2068),(3035),(3991),(3034),
(2066),(715),(1467),(2063),(711),(3028),(4001),(708),(3021),(1483),(2381),(4011),(1486),(4012),(3014),(3014),(4013),(169
),(169),(701),(4015),(4640),(4017),(2059),(695),(3008),(171),(3007),(2389),(4023),(2389),(2058),(3006),(2058),(691),(299
7),(2057),(4033),(4033),(2054),(4034),(180),(181),(2399),(4038),(4628),(181),(1498),(4039),(4040),(2991),(2049),(2991),(
2403),(4041),(4047),(681),(1499),(2046),(2404),(185),(4048),(4048),(2989),(186),(2045),(2224),(188),(1501),(2044),(677),
(189),(4052),(4059),(671),(1512),(4062),(4065),(2410),(4067),(192),(4615),(666),(2412),(2413),(2964),(1519),(2959),(4087),
(659),(4089),(4090),(2957),(2413),(655),(4610),(652),(4093),(4094),(2955),(2953),(1528),(4096),(1531),(2032),(2949),(2
1,1 Top
```

Рисунок 4 Результат выполнения задания 1

```
3,6,0.5
~
~
~
~
```

Рисунок 5 Результат выполнения задания 2 (статистические результаты)

```
login2,Jane Smith,1
login3,Bob Lee,1
login4,Alice Kim,1
~
~
~
~
```

Рисунок 6 Результат выполнения задания 2 (кто сделал меньше 3 твитов)

## Листинги программ:

*task1.pig:*

```
STOPWORDS = LOAD 'input/stopwords.csv' USING PigStorage(',') AS
(word:chararray);

BOOK = LOAD 'input/book.txt' USING PigStorage('\n') AS (line:chararray);

BOOK_WITH_NUM = RANK BOOK;
```

```

BOOK_LOWER = FOREACH BOOK_WITH_NUM GENERATE
    rank BOOK AS line_num,
    LOWER(line) AS line;

BOOK_WORDS = FOREACH BOOK_LOWER GENERATE
    line_num,
    FLATTEN(TOKENIZE(line)) AS word;

CLEAN_WORDS = FOREACH BOOK_WORDS GENERATE
    line_num,
    REPLACE(LOWER(word), '[^a-z]', '') AS word;

FILTERED_WORDS = FILTER CLEAN_WORDS BY word IS NOT NULL AND word != '';

STOPWORDS_LOWER = FOREACH STOPWORDS GENERATE LOWER(word) AS word;

JOINED = JOIN FILTERED_WORDS BY word, STOPWORDS_LOWER BY word USING
'replicated';

GROUPED = GROUP JOINED BY FILTERED_WORDS::word;

RESULT = FOREACH GROUPED GENERATE
    group AS word,
    JOINED.FILTERED_WORDS::line_num AS lines_bag;

STORE RESULT INTO 'output/task1_2025-10-20' USING PigStorage('\t');

```

### *task2.pig:*

```

TWEETS = LOAD 'input/tweets.csv' USING PigStorage(',') AS (tweet_id:int,
tweet:chararray, login:chararray);
USERS = LOAD 'input/users.csv' USING PigStorage(',') AS (login:chararray,
user_name:chararray, state:chararray);

TWEET_COUNT = GROUP TWEETS BY login;
USER_TWEETS = FOREACH TWEET_COUNT GENERATE
    group AS login,
    COUNT(TWEETS) AS tweet_count;

FEW_TWEETS = FILTER USER_TWEETS BY tweet_count < 3;

FEW_USERS = JOIN FEW_TWEETS BY login, USERS BY login;

FEW_USERS_FINAL = FOREACH FEW_USERS GENERATE
    FEW_TWEETS::login AS login,
    USERS::user_name AS user_name,
    FEW_TWEETS::tweet_count AS tweet_count;

FEW_TWEETS_SUM = FOREACH (GROUP FEW_USERS_FINAL ALL) GENERATE
SUM(FEW_USERS_FINAL.tweet_count) AS few_total;

TOTAL_TWEETS = FOREACH (GROUP TWEETS ALL) GENERATE COUNT(TWEETS) AS total;

JOINED = CROSS FEW_TWEETS_SUM, TOTAL_TWEETS;
RESULT_STATS = FOREACH JOINED GENERATE
    few_total,
    total,
    (double)few_total / (double)total AS fraction;

STORE RESULT_STATS INTO 'output/task2_stats_2025-10-20' USING
PigStorage(',');

```

```
STORE FEW_USERS_FINAL INTO 'output/task2_users_2025-10-20' USING  
PigStorage(',');
```

**Вывод:** в ходе выполнения лабораторной работы были получены практические навыки по написанию программ на языке Pig Latin.