



Министерство науки и высшего образования Российской Федерации
Калужский филиал
федерального государственного бюджетного
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»

Лабораторная работа №4

«Язык Pig Latin»

ДИСЦИПЛИНА: «Технологии обработки больших данных»

Выполнил: студент гр. ИУК4-72Б _____ (_____)
(подпись) (Ф.И.О.)

Проверил: _____ (_____)
(подпись) (Ф.И.О.)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:

- Оценка:

Калуга, 2023

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

Постановка задачи

Вариант 4

Задание 1

Построить индекс файла. Для каждого слова в файле результат должен содержать номера всех строка, 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. Выполнить задание по варианту, используя Pig Latin. Файл tweets.csv имеет формат: tweet_id, tweet, login Файл users.csv имеет формат: login, user_name, state Файлы: tweets.csv, users.csv

Вывести имена пользователей, опубликовавших хотя бы 2 твита. Отсортировать результат по активности пользователя (пользователи с наибольшим числом твитов должны быть вверху списка).

Ход выполнения работы

Задание 1

Листинг программы

```
raw_data = LOAD '/user/hduser/lab4/task1/test.txt' USING PigStorage('\n')
AS (line:chararray);
with_indices = RANK raw_data;
words_with_indices = FOREACH with_indices GENERATE $0,
FLATTEN(TOKENIZE(line, ' '));
grouped_words = GROUP words_with_indices BY $1;
result = FOREACH grouped_words GENERATE $0 as word, BagToTuple($1.$0) as
indices;
```

```
RMF /user/hduser/lab4/task2-output;
STORE result INTO '/user/hduser/lab4/task1-output.txt' using
PigStorage(',');
DUMP result;
```

Результаты выполнения скрипта

```
This ebook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this ebook or online
at www.gutenberg.org. If you are not located in the United States,
you will have to check the laws of the country where you are located
before using this eBook.

Title: Less than kin

Author: Alice Duer Miller

Release date: September 18, 2023 [eBook #71674]

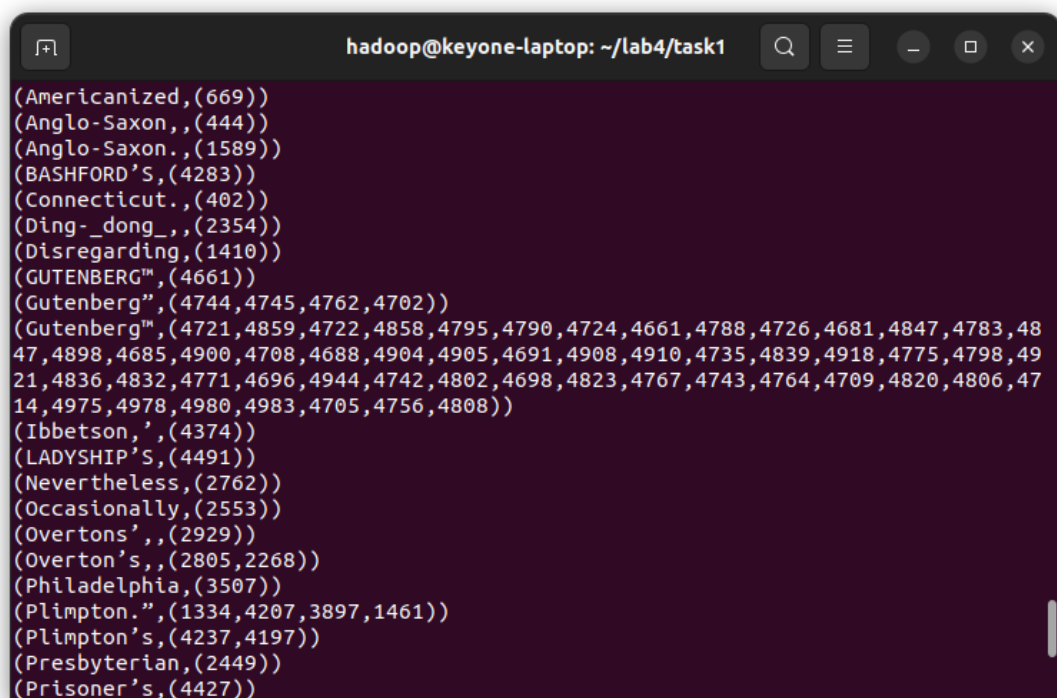
Language: English

Original publication: New York: Henry Holt and Company, 1909

Credits: Steve Mattern and the Online Distributed Proofreading Team at https://www.pgdp.net (This boo

*** START OF THE PROJECT GUTENBERG EBOOK LESS THAN KIN ***
```

Рисунок 1. Входной файл



```
hadoop@keyone-laptop: ~/lab4/task1
(Americanized,(669))
(Anglo-Saxon,,(444))
(Anglo-Saxon.,(1589))
(BASHFORD'S,(4283))
(Connecticut.,(402))
(Ding_dong_.,(2354))
(Disregarding,(1410))
(GUTENBERG™,(4661))
(Gutenberg", (4744,4745,4762,4702))
(Gutenberg™,(4721,4859,4722,4858,4795,4790,4724,4661,4788,4726,4681,4847,4783,48
47,4898,4685,4900,4708,4688,4904,4905,4691,4908,4910,4735,4839,4918,4775,4798,49
21,4836,4832,4771,4696,4944,4742,4802,4698,4823,4767,4743,4764,4709,4820,4806,47
14,4975,4978,4980,4983,4705,4756,4808))
(Ibbetson,',(4374))
(LADYSHIP'S,(4491))
(Nevertheless,(2762))
(Occasionally,(2553))
(Overtons',,(2929))
(Overton's,,(2805,2268))
(Philadelphia,(3507))
(Plimpton.",(1334,4207,3897,1461))
(Plimpton's,(4237,4197))
(Presbyterian,(2449))
(Prisoner's,(4427))
```

Рисунок 2. Демонстрация результата выполнения скрипта

Задание 2

Листинг программы

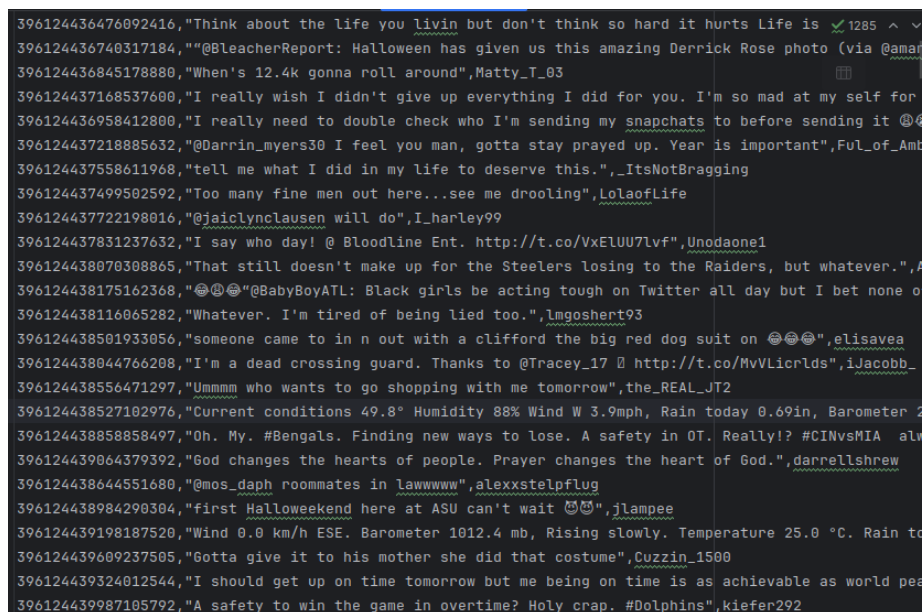
```
REGISTER hdfs:///tmp/piggybank.jar;
DEFINE CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

tweets = LOAD '/user/hduser/lab4/task2/tweets.csv' using CSVLoader() AS
(tweet_id: long, tweet: chararray, login: chararray);
users = LOAD '/user/hduser/lab4/task2/users.csv' using CSVLoader() AS
(login: chararray, user_name: chararray, state: chararray);

grouped_tweets = GROUP tweets BY login;
login_with_tweet_count = FOREACH grouped_tweets GENERATE $0 as login,
COUNT($1) as tweet_count;
filtered_logins = FILTER login_with_tweet_count BY tweet_count >= 2;
joined_logins = JOIN filtered_logins BY login, users BY login;
user_tweet_count = FOREACH joined_logins GENERATE users::user_name,
filtered_logins::tweet_count;
result = ORDER user_tweet_count BY $1 DESC;

RMF /user/hduser/lab4/task2-output;
STORE result INTO '/user/hduser/lab4/task2-output' using PigStorage(',');
DUMP result;
```

Результаты выполнения скрипта



```
396124436476092416,"Think about the life you livin but don't think so hard it hurts Life is ✓ 1285 ^ v
396124436740317184,"@BleacherReport: Halloween has given us this amazing Derrick Rose photo (via @amar
396124436845178880,"When's 12.4k gonna roll around",Matty_I_03
396124437168537600,"I really wish I didn't give up everything I did for you. I'm so mad at my self for
396124436958412800,"I really need to double check who I'm sending my snapchat's to before sending it @C
396124437218885632,"@Darrin_myers30 I feel you man, gotta stay prayed up. Year is important",Ful_of_Amb
396124437558611968,"tell me what I did in my life to deserve this.",_ItsNotBragging
396124437499502592,"Too many fine men out here...see me drooling",LolaofLife
396124437722198016,"@jaiclynclausen will do",I_harley99
396124437831237632,"I say who day! @ Bloodline Ent. http://t.co/VxELUW7lvf",Unodaone1
396124438070308865,"That still doesn't make up for the Steelers losing to the Raiders, but whatever.",A
396124438175162368,"@BabyBoyATL: Black girls be acting tough on Twitter all day but I bet none o
396124438116065282,"Whatever. I'm tired of being lied too.",lmgoshert93
396124438501933056,"someone came to in n out with a clifford the big red dog suit on @elisevea
396124438044766208,"I'm a dead crossing guard. Thanks to @Tracey_17 http://t.co/MvVLicrlds",iJacobb_
396124438556471297,"Ummm who wants to go shopping with me tomorrow",the_REAL_JT2
396124438527102976,"Current conditions 49.8° Humidity 88% Wind W 3.9mph, Rain today 0.69in, Barometer 2
396124438858858497,"Oh. My. #Bengals. Finding new ways to lose. A safety in OT. Really!? #CINvsMIA al
396124439064379392,"God changes the hearts of people. Prayer changes the heart of God.",darrellshrew
396124438644551680,"@mos_daph roommates in lawwww",alexxtelpflug
396124438984290304,"first Halloweekend here at ASU can't wait 🎃",jlampee
396124439198187520,"Wind 0.0 km/h ESE. Barometer 1012.4 mb, Rising slowly. Temperature 25.0 °C. Rain to
396124439609237505,"Gotta give it to his mother she did that costume",Cuzzin_1500
396124439324012544,"I should get up on time tomorrow but me being on time is as achievable as world pea
396124439987105792,"A safety to win the game in overtime? Holy crap. #Dolphins",Kiefer292
```

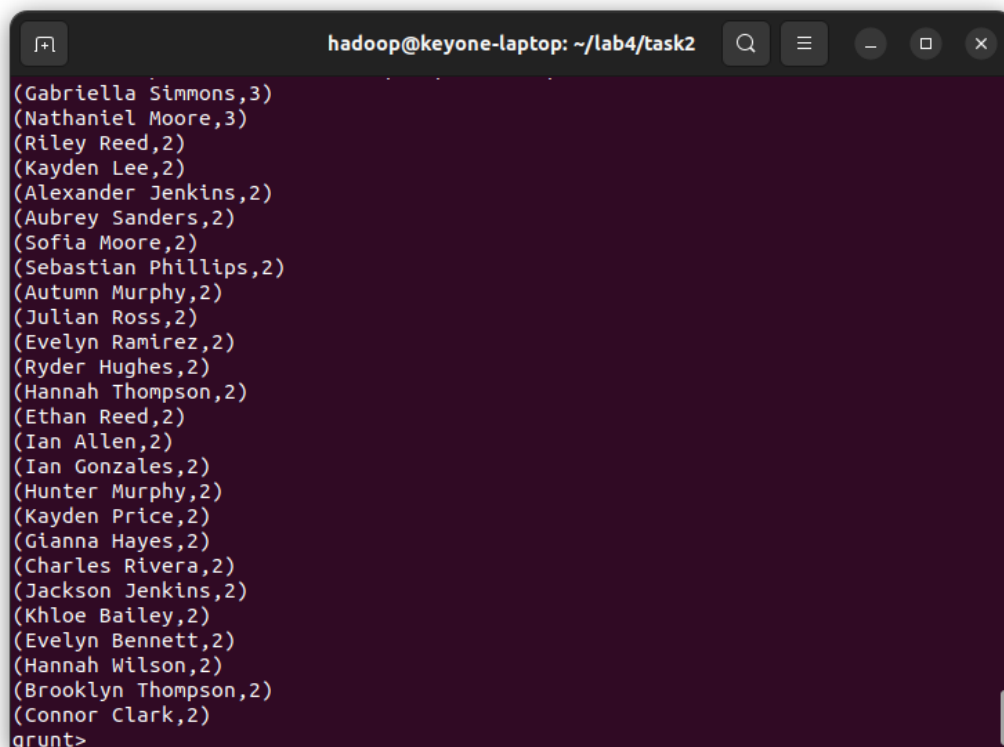
Рисунок 3. Входной файл tweets.csv

```

Obey_Jony09,Riley Bell,MA
Colten_stamkos,Lily King,RI
Matty_T_03,Jackson Jenkins,GU
savava143,Lauren Henderson,MA
julianpham,Noah Walker,PR
Ful_of_Ambition,Levi Watson,OR
_ItsNotBragging,John Hill,MT
LolaofLife,Lily Simmons,UT
I_harley99,Aubrey Butler,GA
Unodaone1,Kennedy Gray,LA
AVar14,Layla Richardson,MO
AmberChadwick_,Riley Diaz,IN
lmgoshert93,Madeline Campbell,VT
elisavea,Wyatt Kelly,PW
iJacobb_,Piper Baker,MS
the_REAL_JT2,Noah Garcia,CT
OakvilleWx,Gianna Williams,LA
pgaynor14,Genesis Walker,OR
darrellshrew,Carlos Powell,DE
alexstelpflug,Jocelyn Cooper,ND
jlampee,Mason Walker,WI
DW9577SW,Peyton Martinez,PW
Cuzzin_1500,Isaac Allen,AR
Brookiebabyy86,Ethan Butler,AZ
kiefer292,Brody Hernandez,HI

```

Рисунок 4. Входной файл users.csv



```

hadoop@keyone-laptop: ~/lab4/task2
(Gabriella Simmons,3)
(Nathaniel Moore,3)
(Riley Reed,2)
(Kayden Lee,2)
(Alexander Jenkins,2)
(Aubrey Sanders,2)
(Sofia Moore,2)
(Sebastian Phillips,2)
(Autumn Murphy,2)
(Julian Ross,2)
(Evelyn Ramirez,2)
(Ryder Hughes,2)
(Hannah Thompson,2)
(Ethan Reed,2)
(Ian Allen,2)
(Ian Gonzales,2)
(Hunter Murphy,2)
(Kayden Price,2)
(Gianna Hayes,2)
(Charles Rivera,2)
(Jackson Jenkins,2)
(Khloe Bailey,2)
(Evelyn Bennett,2)
(Hannah Wilson,2)
(Brooklyn Thompson,2)
(Connor Clark,2)
grunt>

```

Рисунок 5. Результат выполнения скрипта

Вывод: в ходе выполнения лабораторной работы были сформированы практические навыки реализации `rig`-скриптов для обработки больших данных.