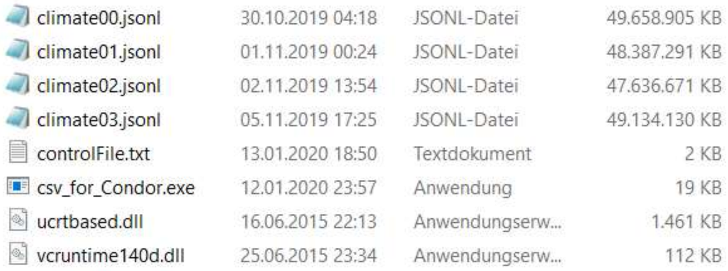
This readme file describes how to create csv-files as import files for Condor application.

requirements / tested with

- Laptop/PC with Windows 10 operating system
- *csv_for_Condor.exe* application
- *controlFile.txt* configuration file to give some information for csv_for_Condor.exe
- runtime DLL *vcruntime140d.dll* and *ucrtbased.dll* or MS Visual Studio installation

preparations

- copy the jsonl input files together with following files into one folder



climate00.jsonl	30.10.2019 04:18	JSONL-Datei	49.658.905 KB
climate01.jsonl	01.11.2019 00:24	JSONL-Datei	48.387.291 KB
climate02.jsonl	02.11.2019 13:54	JSONL-Datei	47.636.671 KB
climate03.jsonl	05.11.2019 17:25	JSONL-Datei	49.134.130 KB
controlFile.txt	13.01.2020 18:50	Textdokument	2 KB
csv_for_Condor.exe	12.01.2020 23:57	Anwendung	19 KB
ucrtbased.dll	16.06.2015 22:13	Anwendungserw...	1.461 KB
vcruntime140d.dll	25.06.2015 23:34	Anwendungserw...	112 KB

explanations for controlFile.txt

- don’t rename this file
- don’t insert or delete lines in this file
- don’t delete elements complete in column 1 and 2 (replacing is ok)
- don’t make any changes from line 31
- edit this file to your concerns with a text editor (e.g. *editor* or *wordpad*)

running the application

- first configure controlFile.txt and close this file
- turn off the screen saver
- double click *csv_for_Condor.exe* in windows explorer
- start process by pressing key *y*
- the console window should be focused during the application runs
- check in the task manager that the application don’t sleep, if it sleeps select the console window and press key ctrl+c
- wait for success message
- press return key for program ending

abort a running program

- press key ctrl+c
- follow the instructions

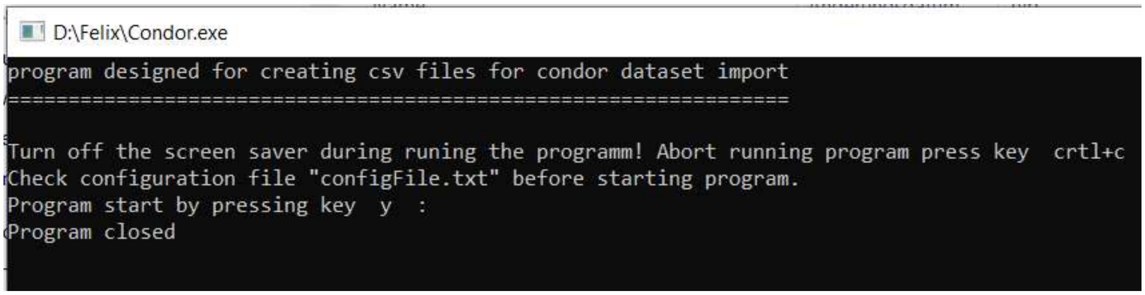
line_property	value	explanation
_0sourceFileName	climate00.jsonl	name of input file
_1sourceFileNo	0	postfix output folder
_2actorsFile<no>	actors_climate_tweets_	leader in actors output file name
_3linksFile<no>	links_climate_tweets_	leader in links output file name
_4endingDestFile	.csv	file extension output file name
_5folder_actors	d:\\Felix\\actors	leader in actors output folder name
_6folder_links	d:\\Felix\\links	leader in links output folder name
_7separator_csv_(;,|t)	;	separator in csv output file
_8noLinesDestFile	200000	number of lines in output file
_9enddate	2019-06-17	endtime in output files
10endtime	00:00:00	endtime in output files
11hlActorsCol_1	Uuid	column 1 heading name actors file
12hlActorsCol_2	Starttime	column 2 heading name actors file
13hlActorsCol_3	Endtime	column 3 heading name actors file
14hlLinksCol_1	Starttime	column 1 heading name links file
15hlLinksCol_2	fulltext	column 2 heading name links file
16hlLinksCol_3	SourceUuid	column 3 heading name links file
17hlLinksCol_4	TargetUuid	column 4 heading name links file
18hlLinksCol_5	Endtime	column 5 heading name links file
19separatorDate	-	separator char for date
20dummyActor	blank	name of dummy actor, e.g. one space
21addCharEndOfFullText	blank	add char at the end of fulltext
22replaceCharInFullx>y	;>,	replace char in fulltext, e.g. ; with ,
23replaceCharInFullx>y	">'	replace char in fulltext, e.g. “ with ‘
24replaceCharInFullx>y	not_in_use	replace char in fulltext, e.g. no replace
25cut_active_(y/n)	n	cut an xxl-file into a xl-file (jsonl) if y
26cutSourceFile	d:\\Felix\\climate00.jsonl	source xxl file
27cutDestFile	d:\\Felix\\xxx.jsonl	destination xl file
28cutNoLines	50000	number of lines in xl file
29reserve_9	not_in_use	reserve for later extentions
30reserve_10	not_in_use	reserve for later extentions
31noObjectsRead	7	don’t change
32object0	"created_at":	don’t change
33object1	"full_text":	don’t change
34object2	"user_mentions":	don’t change
35object3	"screen_name":	don’t change
36object4	"user":	don’t change
37object5	"screen_name":	don’t change
38object6	"created_at":	don’t change

remarks for controlFile.txt

- line 1 don't forget to change this number if choose another source file in line 0 e.g. Line 0 = climate00.jsonl -> Line 1 = 0
  Line 0 = climate01.jsonl -> Line 1 = 1
  The value in line 1 is part of the folder name for the output files, if forget already created output files can be overwritten
- line 7 possible tabulators in the csv file are **; , | t** t = tab
- line 20, 21 blank = <u>one</u> space
- line 20-24 not_in_use = no action in output files
- line 22-24 you can replace up to 3 chars in fulltext, e.g. **;>,** means every **;** in fulltext will be replaced with **,**
- line 20 dummy actor, necessary otherwise some import failure in Condor
- line 21 additional space at the end of fulltext, necessary otherwise some import failure in Condor, Condor extended hyperlinks at the end of fulltext over and over the separator if there is no additional space
- line 25 cut functionality, if **y** no csv files are created, instead of you can create a sub file (line 27) of the source file (line 26), this sub file contains the first n lines (line 28) of the source file

possible program sequences

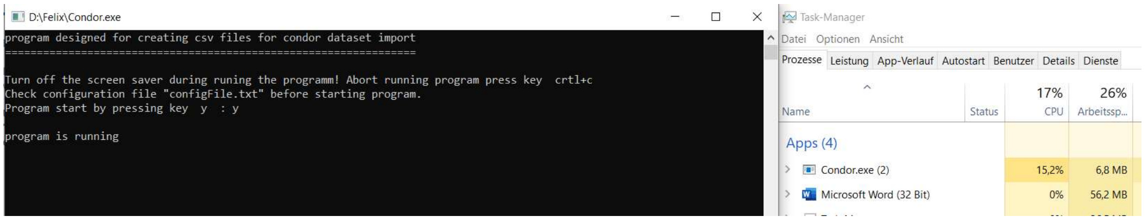
if you don't press key y the program will be closed



D:\Felix\Condor.exe

```
program designed for creating csv files for condor dataset import
=================================================================

Turn off the screen saver during runing the programm! Abort running program press key  crtl+c
Check configuration file "configFile.txt" before starting program.
Program start by pressing key  y  :
Program closed
```

start process press key y controlling in task manager, that the program don't sleep



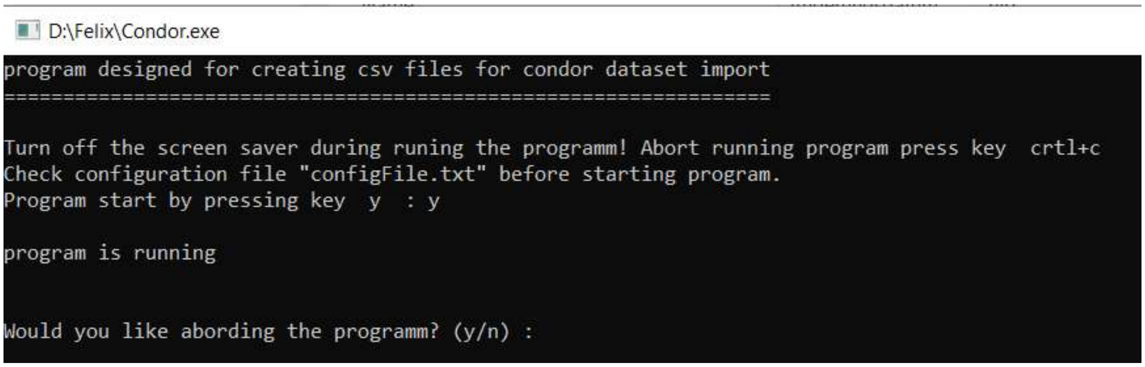
D:\Felix\Condor.exe

```
program designed for creating csv files for condor dataset import
=================================================================

Turn off the screen saver during runing the programm! Abort running program press key  crtl+c
Check configuration file "configFile.txt" before starting program.
Program start by pressing key  y  : y

program is running
```

Task-Manager
Datei Optionen Ansicht
Prozesse Leistung App-Verlauf Autostart Benutzer Details Dienste

Name	Status	17% CPU	26% Arbeitssp...
Apps (4)			
Condor.exe (2)		15,2%	6,8 MB
Microsoft Word (32 Bit)		0%	56,2 MB

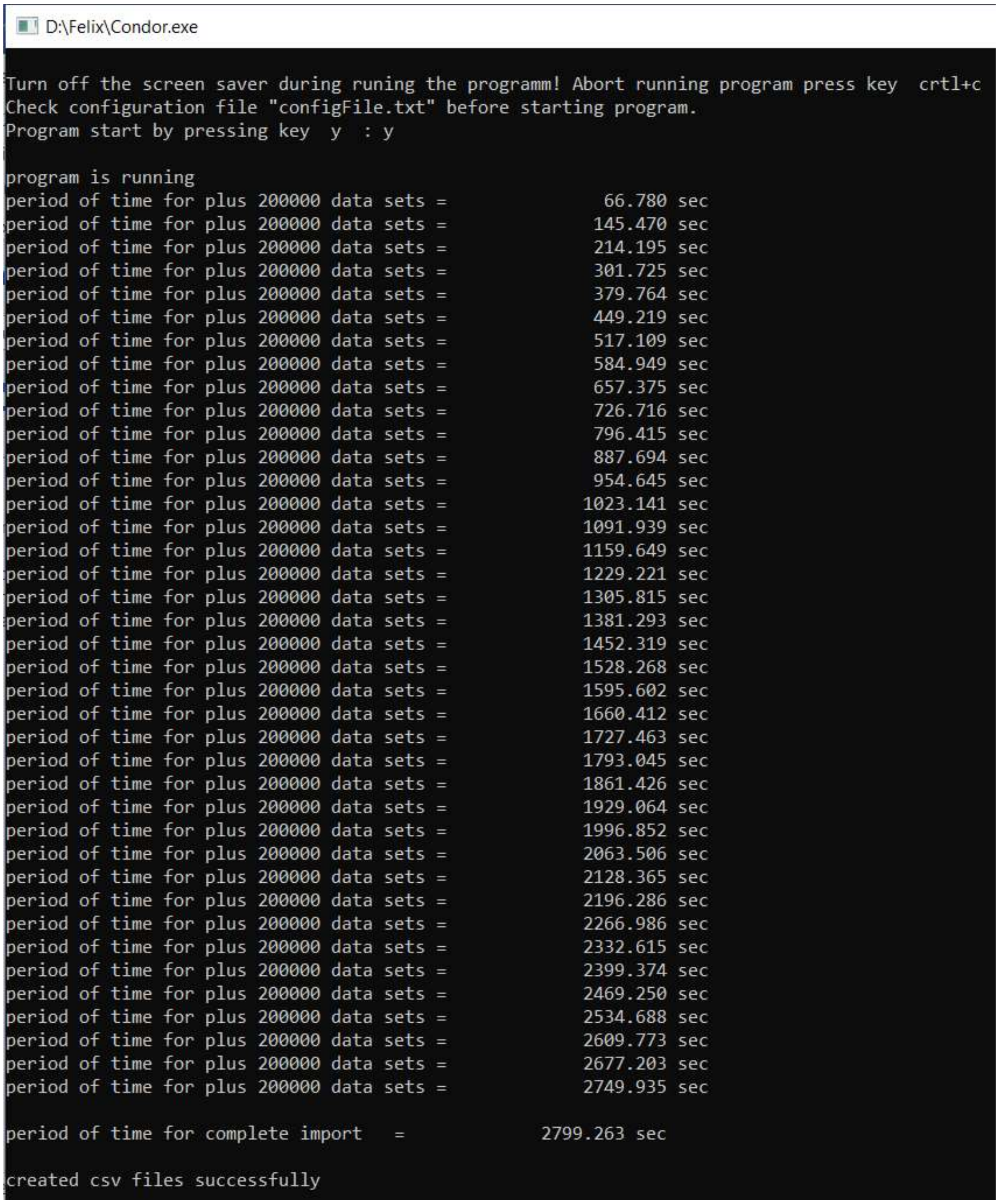
Abort program -> press key ctrl+c



D:\Felix\Condor.exe

```
program designed for creating csv files for condor dataset import
=================================================================

Turn off the screen saver during runing the programm! Abort running program press key  crtl+c
Check configuration file "configFile.txt" before starting program.
Program start by pressing key  y  : y

program is running


Would you like abording the programm? (y/n) :
```

program successfully finished



```
Turn off the screen saver during runing the programm! Abort running program press key  crtl+c
Check configuration file "configFile.txt" before starting program.
Program start by pressing key  y  : y

program is running
period of time for plus 200000 data sets =             66.780 sec
period of time for plus 200000 data sets =            145.470 sec
period of time for plus 200000 data sets =            214.195 sec
period of time for plus 200000 data sets =            301.725 sec
period of time for plus 200000 data sets =            379.764 sec
period of time for plus 200000 data sets =            449.219 sec
period of time for plus 200000 data sets =            517.109 sec
period of time for plus 200000 data sets =            584.949 sec
period of time for plus 200000 data sets =            657.375 sec
period of time for plus 200000 data sets =            726.716 sec
period of time for plus 200000 data sets =            796.415 sec
period of time for plus 200000 data sets =            887.694 sec
period of time for plus 200000 data sets =            954.645 sec
period of time for plus 200000 data sets =           1023.141 sec
period of time for plus 200000 data sets =           1091.939 sec
period of time for plus 200000 data sets =           1159.649 sec
period of time for plus 200000 data sets =           1229.221 sec
period of time for plus 200000 data sets =           1305.815 sec
period of time for plus 200000 data sets =           1381.293 sec
period of time for plus 200000 data sets =           1452.319 sec
period of time for plus 200000 data sets =           1528.268 sec
period of time for plus 200000 data sets =           1595.602 sec
period of time for plus 200000 data sets =           1660.412 sec
period of time for plus 200000 data sets =           1727.463 sec
period of time for plus 200000 data sets =           1793.045 sec
period of time for plus 200000 data sets =           1861.426 sec
period of time for plus 200000 data sets =           1929.064 sec
period of time for plus 200000 data sets =           1996.852 sec
period of time for plus 200000 data sets =           2063.506 sec
period of time for plus 200000 data sets =           2128.365 sec
period of time for plus 200000 data sets =           2196.286 sec
period of time for plus 200000 data sets =           2266.986 sec
period of time for plus 200000 data sets =           2332.615 sec
period of time for plus 200000 data sets =           2399.374 sec
period of time for plus 200000 data sets =           2469.250 sec
period of time for plus 200000 data sets =           2534.688 sec
period of time for plus 200000 data sets =           2609.773 sec
period of time for plus 200000 data sets =           2677.203 sec
period of time for plus 200000 data sets =           2749.935 sec

period of time for complete import   =           2799.263 sec

created csv files successfully
```

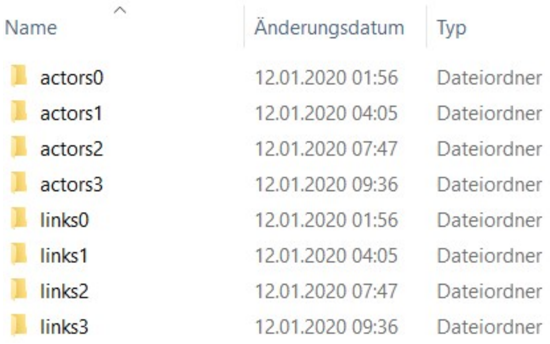
- program duration 2800 seconds for a 50GB input file
- leads to 2GB output files (actors & links)
- performance data for used laptop (16GB Memory, SSD, Windows 10)

x64-basierter PC

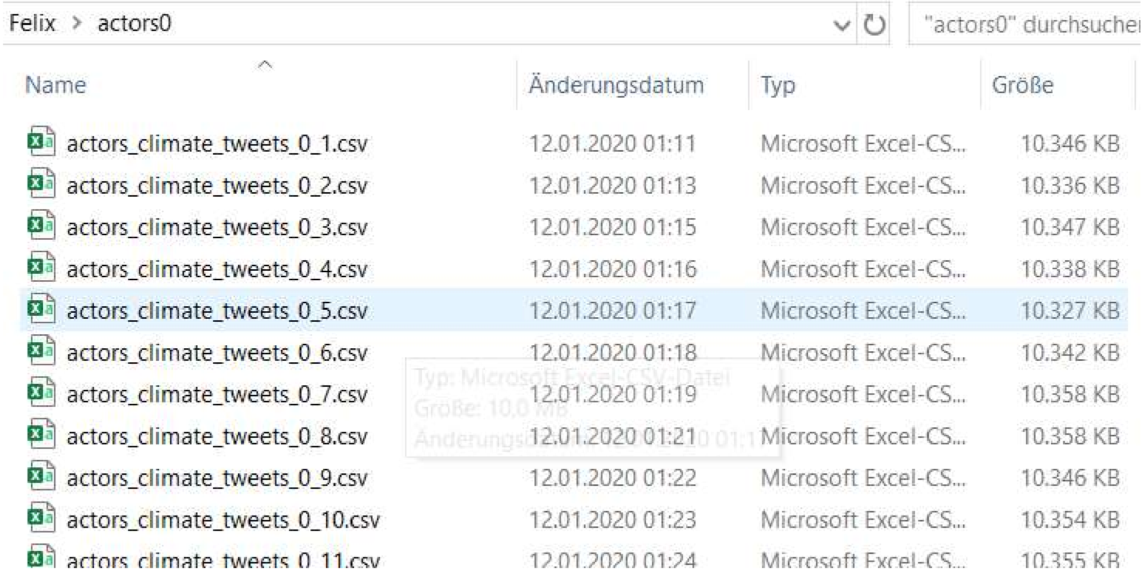
LENOVO_MT_20QQ_BU_Think_FM_ThinkPad P53

Intel(R) Core(TM) i7-9850H CPU @ 2.60GHz, 2592 MHz, 6 Kern(e),

data storage for 4 input files

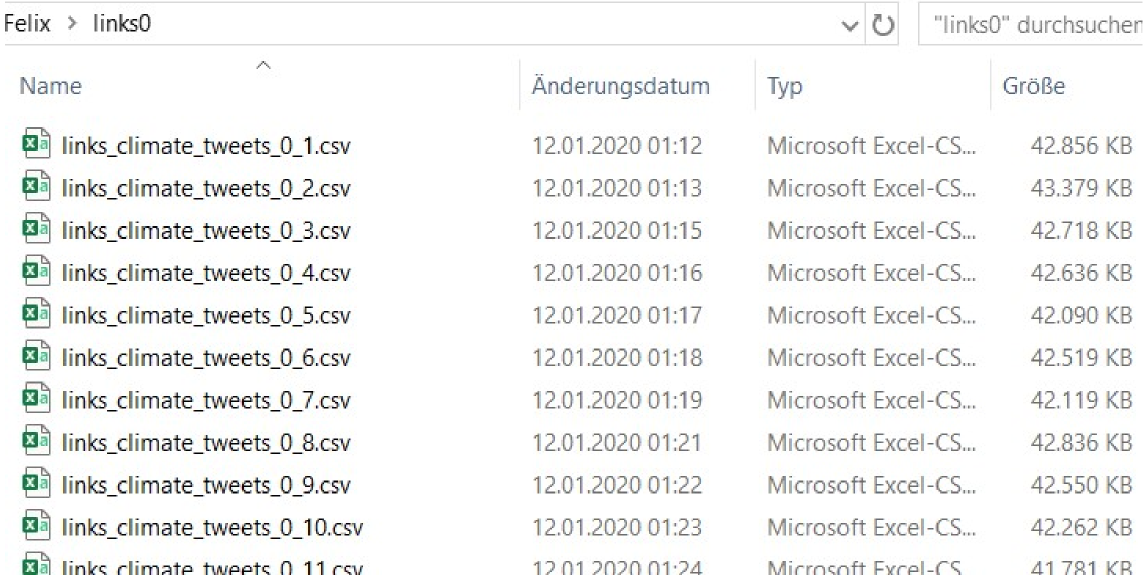


Name	Änderungsdatum	Typ
actors0	12.01.2020 01:56	Dateiordner
actors1	12.01.2020 04:05	Dateiordner
actors2	12.01.2020 07:47	Dateiordner
actors3	12.01.2020 09:36	Dateiordner
links0	12.01.2020 01:56	Dateiordner
links1	12.01.2020 04:05	Dateiordner
links2	12.01.2020 07:47	Dateiordner
links3	12.01.2020 09:36	Dateiordner



Name	Änderungsdatum	Typ	Größe
actors_climate_tweets_0_1.csv	12.01.2020 01:11	Microsoft Excel-CS...	10.346 KB
actors_climate_tweets_0_2.csv	12.01.2020 01:13	Microsoft Excel-CS...	10.336 KB
actors_climate_tweets_0_3.csv	12.01.2020 01:15	Microsoft Excel-CS...	10.347 KB
actors_climate_tweets_0_4.csv	12.01.2020 01:16	Microsoft Excel-CS...	10.338 KB
actors_climate_tweets_0_5.csv	12.01.2020 01:17	Microsoft Excel-CS...	10.327 KB
actors_climate_tweets_0_6.csv	12.01.2020 01:18	Microsoft Excel-CS...	10.342 KB
actors_climate_tweets_0_7.csv	12.01.2020 01:19	Microsoft Excel-CS...	10.358 KB
actors_climate_tweets_0_8.csv	12.01.2020 01:21	Microsoft Excel-CS...	10.358 KB
actors_climate_tweets_0_9.csv	12.01.2020 01:22	Microsoft Excel-CS...	10.346 KB
actors_climate_tweets_0_10.csv	12.01.2020 01:23	Microsoft Excel-CS...	10.354 KB
actors_climate_tweets_0_11.csv	12.01.2020 01:24	Microsoft Excel-CS...	10.355 KB

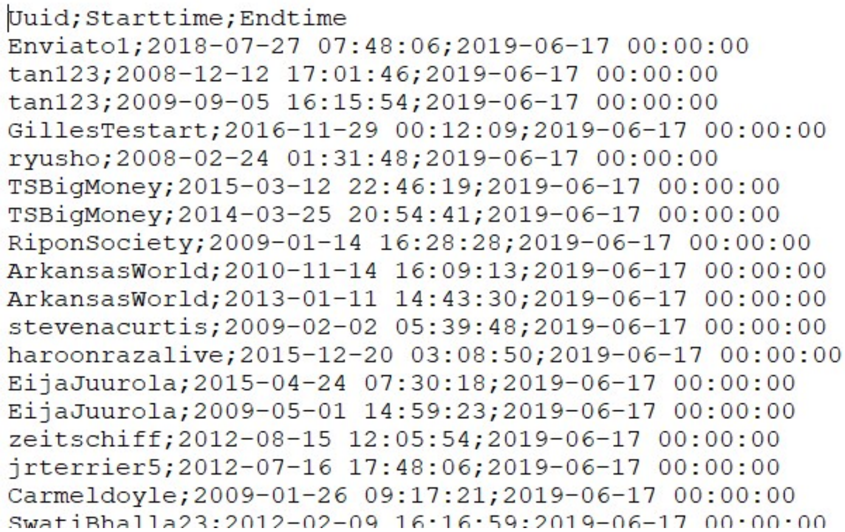
Typ: Microsoft Excel-CSV-Datei
Größe: 10,0 MB
Änderungsdatum: 12.01.2020 01:1



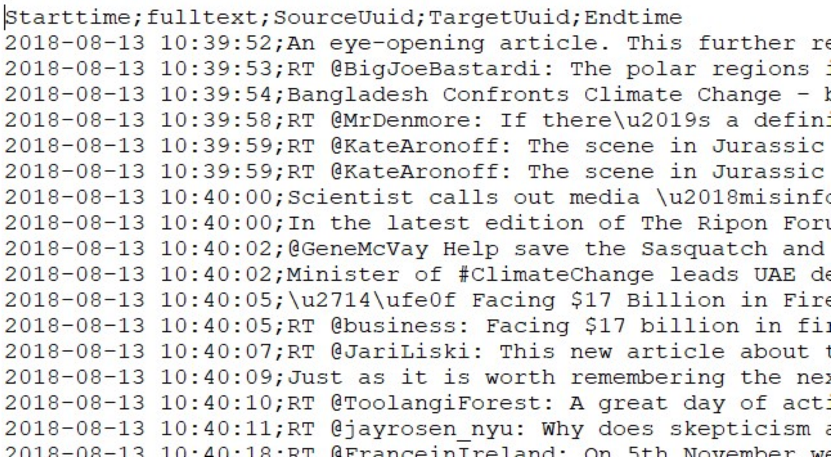
Name	Änderungsdatum	Typ	Größe
links_climate_tweets_0_1.csv	12.01.2020 01:12	Microsoft Excel-CS...	42.856 KB
links_climate_tweets_0_2.csv	12.01.2020 01:13	Microsoft Excel-CS...	43.379 KB
links_climate_tweets_0_3.csv	12.01.2020 01:15	Microsoft Excel-CS...	42.718 KB
links_climate_tweets_0_4.csv	12.01.2020 01:16	Microsoft Excel-CS...	42.636 KB
links_climate_tweets_0_5.csv	12.01.2020 01:17	Microsoft Excel-CS...	42.090 KB
links_climate_tweets_0_6.csv	12.01.2020 01:18	Microsoft Excel-CS...	42.519 KB
links_climate_tweets_0_7.csv	12.01.2020 01:19	Microsoft Excel-CS...	42.119 KB
links_climate_tweets_0_8.csv	12.01.2020 01:21	Microsoft Excel-CS...	42.836 KB
links_climate_tweets_0_9.csv	12.01.2020 01:22	Microsoft Excel-CS...	42.550 KB
links_climate_tweets_0_10.csv	12.01.2020 01:23	Microsoft Excel-CS...	42.262 KB
links_climate_tweets_0_11.csv	12.01.2020 01:24	Microsoft Excel-CS...	41.781 KB

- e.g. **actors_climate_tweets_0_5.csv** corresponds with **links_climate_tweets_0_5.csv**

example for output files



```
Uuid;Starttime;Endtime
Enviato1;2018-07-27 07:48:06;2019-06-17 00:00:00
tan123;2008-12-12 17:01:46;2019-06-17 00:00:00
tan123;2009-09-05 16:15:54;2019-06-17 00:00:00
GillesTestart;2016-11-29 00:12:09;2019-06-17 00:00:00
ryusho;2008-02-24 01:31:48;2019-06-17 00:00:00
TSBigMoney;2015-03-12 22:46:19;2019-06-17 00:00:00
TSBigMoney;2014-03-25 20:54:41;2019-06-17 00:00:00
RiponSociety;2009-01-14 16:28:28;2019-06-17 00:00:00
ArkansasWorld;2010-11-14 16:09:13;2019-06-17 00:00:00
ArkansasWorld;2013-01-11 14:43:30;2019-06-17 00:00:00
stevenacurtis;2009-02-02 05:39:48;2019-06-17 00:00:00
haroonrazalive;2015-12-20 03:08:50;2019-06-17 00:00:00
EijaJuurola;2015-04-24 07:30:18;2019-06-17 00:00:00
EijaJuurola;2009-05-01 14:59:23;2019-06-17 00:00:00
zeitschiff;2012-08-15 12:05:54;2019-06-17 00:00:00
jrterrier5;2012-07-16 17:48:06;2019-06-17 00:00:00
Carmeldoyle;2009-01-26 09:17:21;2019-06-17 00:00:00
SwatiBhalla23;2012-02-09 16:16:59;2019-06-17 00:00:00
```



```
Starttime;fulltext;SourceUuid;TargetUuid;Endtime
2018-08-13 10:39:52;An eye-opening article. This further re
2018-08-13 10:39:53;RT @BigJoeBastardi: The polar regions i
2018-08-13 10:39:54;Bangladesh Confronts Climate Change - k
2018-08-13 10:39:58;RT @MrDenmore: If there\u2019s a defini
2018-08-13 10:39:59;RT @KateAronoff: The scene in Jurassic
2018-08-13 10:39:59;RT @KateAronoff: The scene in Jurassic
2018-08-13 10:40:00;Scientist calls out media \u2018misinfo
2018-08-13 10:40:00;In the latest edition of The Ripon Foru
2018-08-13 10:40:02;@GeneMcVay Help save the Sasquatch and
2018-08-13 10:40:02;Minister of #ClimateChange leads UAE de
2018-08-13 10:40:05;\u2714\ufe0f Facing $17 Billion in Fire
2018-08-13 10:40:05;RT @business: Facing $17 billion in fir
2018-08-13 10:40:07;RT @JariLiski: This new article about t
2018-08-13 10:40:09;Just as it is worth remembering the nex
2018-08-13 10:40:10;RT @ToolangiForest: A great day of acti
2018-08-13 10:40:11;RT @jayrosen_nyu: Why does skepticism a
2018-08-13 10:40:18;RT @FranceinIreland: On 5th November we
```