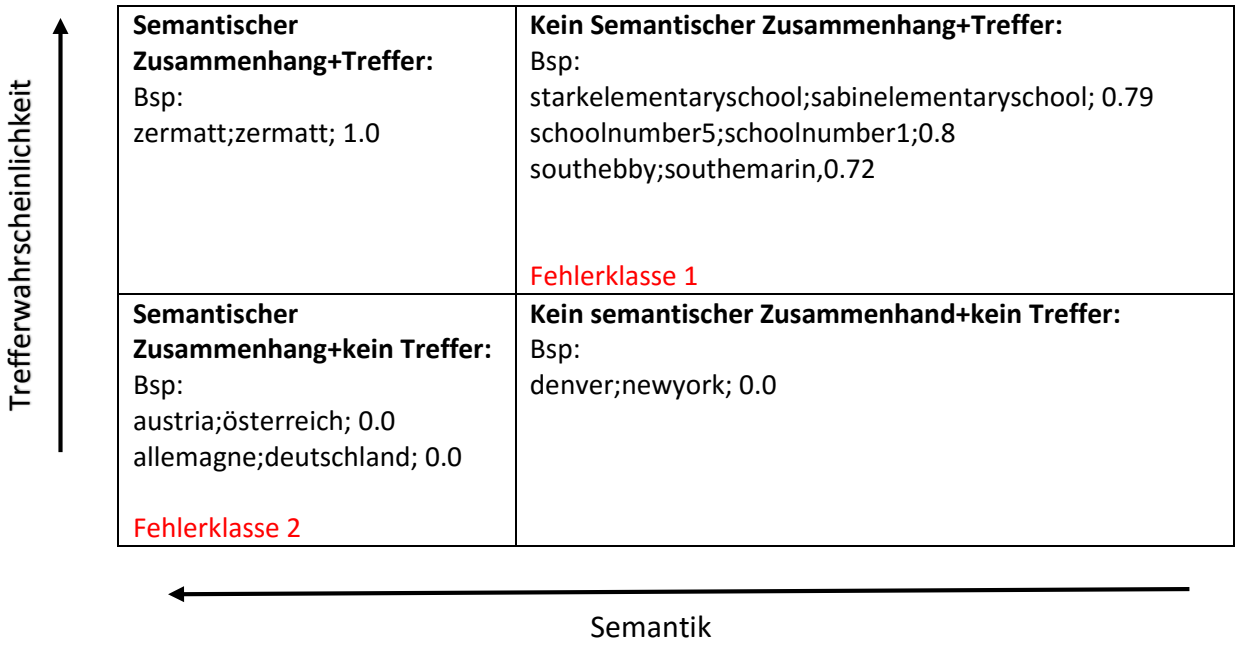
String Similarity Evaluation:

Legende:

Begriff	Erklärung
process	Prozess der in der Anwendung ausgeführt wird
bracketsmode	Filtermodus (true => Klammern und Inhalt nach Komma entfernen)
records	Anzahl Ergebnisdatensätze nach Ausführung
executiontime	Ausführungszeit (in Sekunden)
source	Datenquelle
target	Datenziel
tokensize	Größe der Token
simThreshold	Entscheidungswert für Gleichheit der Zeichenketten (Dice Metrik)

Fehlerbetrachtung:



Zur Fehlerklasse 1: Je nach Datensatz müssen spezifische Regeln gefunden und implementiert werden, um falsche Treffer zu verhindern.

Zur Fehlerklasse 2: Auf technischer Ebene schwierig zu beheben. Durch Einsatz von weiteren Quellen (Verzeichnisse, Wörterbücher, Übersetzungen,..) kann eine Identifizierung ermöglicht werden (jedoch komplex und aufwändig).

VM Specs: Ubuntu 16.04 LTS
4 Cores i5 4670K @3,4 GHz, 12GB RAM, 140GB SSD
Records Dataset perfect: 46039

process	bracketsmode	records	executiontime	target
createCompareCsv	true	28.422.030	35	perfect_Btrue.csv
createCompareCsv	false	28.422.030	36	perfect_Bfalse.csv

algorithm	simThreshold	records	tokensize	executiontime	source
stringCompare	0	28.422.030	-	23	perfect_Bfalse.csv
stringCompareNgran	0	28.422.030	3	85	perfect_Bfalse.csv
flinkSortMerge	0	28.422.030	3	71	perfect_Bfalse.csv
sortMerge	0	28.422.030	3	197	perfect_Bfalse.csv
simmetrics	0	28.422.030	-	131	perfect_Bfalse.csv
stringCompare	-	-	-	-	-
stringCompareNgran	0,7	6.571	3	76	perfect_Bfalse.csv
flinkSortMerge	0,7	6.571	3	78	perfect_Bfalse.csv
sortMerge	0,7	6.571	3	187	perfect_Bfalse.csv
simmetrics	0,7	6.571	-	125	perfect_Bfalse.csv
stringCompare	1	5.002	-	10	perfect_Bfalse.csv
stringCompareNgran	1	5.002	3	75	perfect_Bfalse.csv
flinkSortMerge	1	5.002	3	67	perfect_Bfalse.csv
sortMerge	1	5.002	3	188	perfect_Bfalse.csv
simmetrics	1	5.002		117	perfect_Bfalse.csv

algorithm	simThreshold	records	tokensize	executiontime	source
stringCompareNgran	0	28.422.030	3	83	perfect_Bfalse.csv
stringCompareNgran	0,1	1.955.417	3	76	perfect_Bfalse.csv
stringCompareNgran	0,3	148.741	3	78	perfect_Bfalse.csv
stringCompareNgran	0,5	28.315	3	76	perfect_Bfalse.csv
stringCompareNgran	0,7	6.571	3	75	perfect_Bfalse.csv
stringCompareNgran	0,9	5.002	3	75	perfect_Bfalse.csv
stringCompareNgran	1	5.002	3	75	perfect_Bfalse.csv

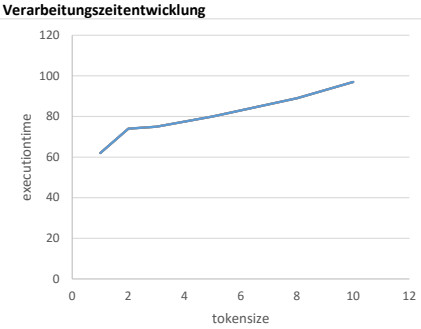
algorithm	simThreshold	records	tokensize	executiontime	source
stringCompareNgran	0,7	165.879	1	67	perfect_Bfalse.csv
stringCompareNgran	0,7	8.577	2	73	perfect_Bfalse.csv
stringCompareNgran	0,7	6.571	3	75	perfect_Bfalse.csv
stringCompareNgran	0,7	5.333	5	84	perfect_Bfalse.csv
stringCompareNgran	0,7	5.050	8	98	perfect_Bfalse.csv
stringCompareNgran	0,7	5.018	10	108	perfect_Bfalse.csv

algorithm	simThreshold	records	tokensize	executiontime	source
stringCompare	0	28.422.030	-	24	perfect_Btrue.csv
stringCompare	1	9.802	-	9	perfect_Btrue.csv

algorithm	simThreshold	records	tokensize	executiontime	source
stringCompareNgram	0,0	28.422.030	3	85	perfect_Btrue.csv
stringCompareNgram	0,1	2.402.933	3	76	perfect_Btrue.csv
stringCompareNgram	0,3	177.793	3	73	perfect_Btrue.csv
stringCompareNgram	0,5	30.032	3	73	perfect_Btrue.csv
stringCompareNgram	0,7	10.604	3	72	perfect_Btrue.csv
stringCompareNgram	0,9	9.806	3	72	perfect_Btrue.csv
stringCompareNgram	1,0	9.802	3	72	perfect_Btrue.csv

algorithm	simThreshold	records	tokensize	executiontime	source
stringCompareNgram	0,7	189.480	1	62	perfect_Btrue.csv
stringCompareNgram	0,7	12.246	2	74	perfect_Btrue.csv
stringCompareNgram	0,7	10.604	3	75	perfect_Btrue.csv
stringCompareNgram	0,7	9.960	5	80	perfect_Btrue.csv
stringCompareNgram	0,7	9.825	8	89	perfect_Btrue.csv
stringCompareNgram	0,7	9.819	10	97	perfect_Btrue.csv

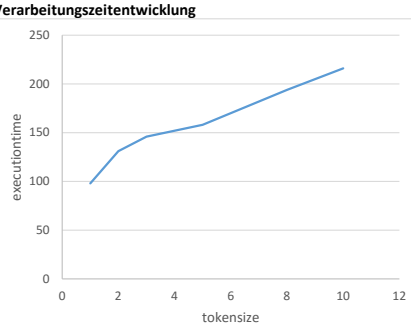
Verarbeitungszeitentwicklung



algorithm	simThreshold	records	tokensize	executiontime	source
sortMerge	0,0	28.422.030	3	154	perfect_Btrue.csv
sortMerge	0,1	2.402.933	3	158	perfect_Btrue.csv
sortMerge	0,3	177.793	3	146	perfect_Btrue.csv
sortMerge	0,5	30.032	3	146	perfect_Btrue.csv
sortMerge	0,7	10.604	3	146	perfect_Btrue.csv
sortMerge	0,9	9.806	3	147	perfect_Btrue.csv
sortMerge	1,0	9.802	3	146	perfect_Btrue.csv

algorithm	simThreshold	records	tokensize	executiontime	source
sortMerge	0,7	189.480	1	98	perfect_Btrue.csv
sortMerge	0,7	12.246	2	131	perfect_Btrue.csv
sortMerge	0,7	10.604	3	146	perfect_Btrue.csv
sortMerge	0,7	9.960	5	158	perfect_Btrue.csv
sortMerge	0,7	9.825	8	194	perfect_Btrue.csv
sortMerge	0,7	9.819	10	216	perfect_Btrue.csv

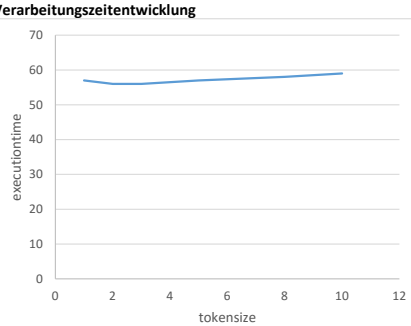
Verarbeitungszeitentwicklung



algorithm	simThreshold	records	tokensize	executiontime	source
flinkSortMerge	0,0	28.422.030	3	71	perfect_Btrue.csv
flinkSortMerge	0,1	2.402.933	3	59	perfect_Btrue.csv
flinkSortMerge	0,3	177.793	3	58	perfect_Btrue.csv
flinkSortMerge	0,5	30.032	3	58	perfect_Btrue.csv
flinkSortMerge	0,7	10.604	3	58	perfect_Btrue.csv
flinkSortMerge	0,9	9.806	3	57	perfect_Btrue.csv
flinkSortMerge	1,0	9.802	3	57	perfect_Btrue.csv

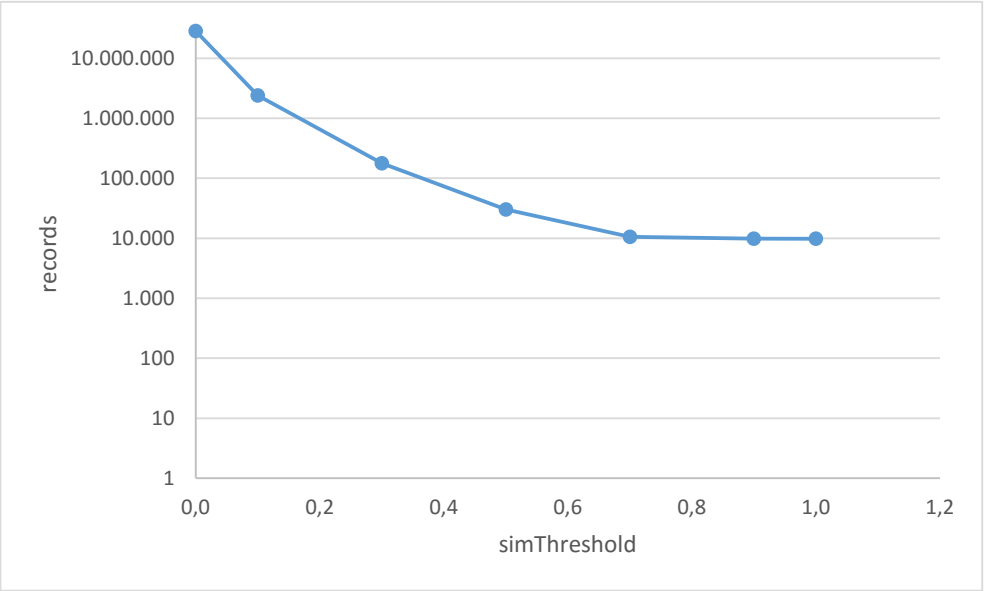
algorithm	simThreshold	records	tokensize	executiontime	source
flinkSortMerge	0,7	189.480	1	57	perfect_Btrue.csv
flinkSortMerge	0,7	12.246	2	56	perfect_Btrue.csv
flinkSortMerge	0,7	10.604	3	56	perfect_Btrue.csv
flinkSortMerge	0,7	9.960	5	57	perfect_Btrue.csv
flinkSortMerge	0,7	9.825	8	58	perfect_Btrue.csv
flinkSortMerge	0,7	9.819	10	59	perfect_Btrue.csv

Verarbeitungszeitentwicklung

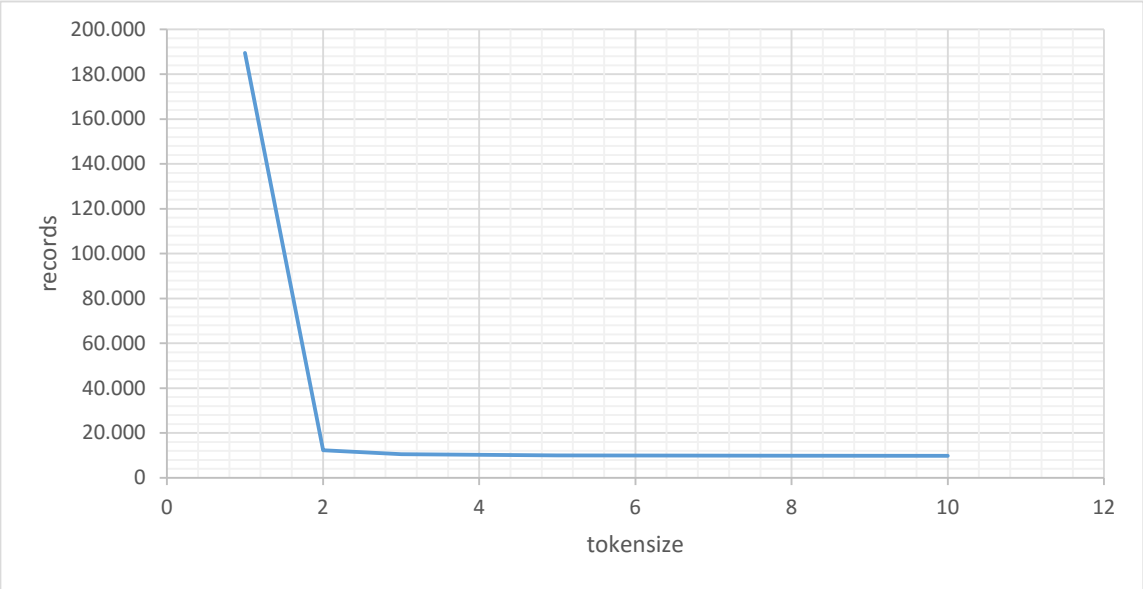


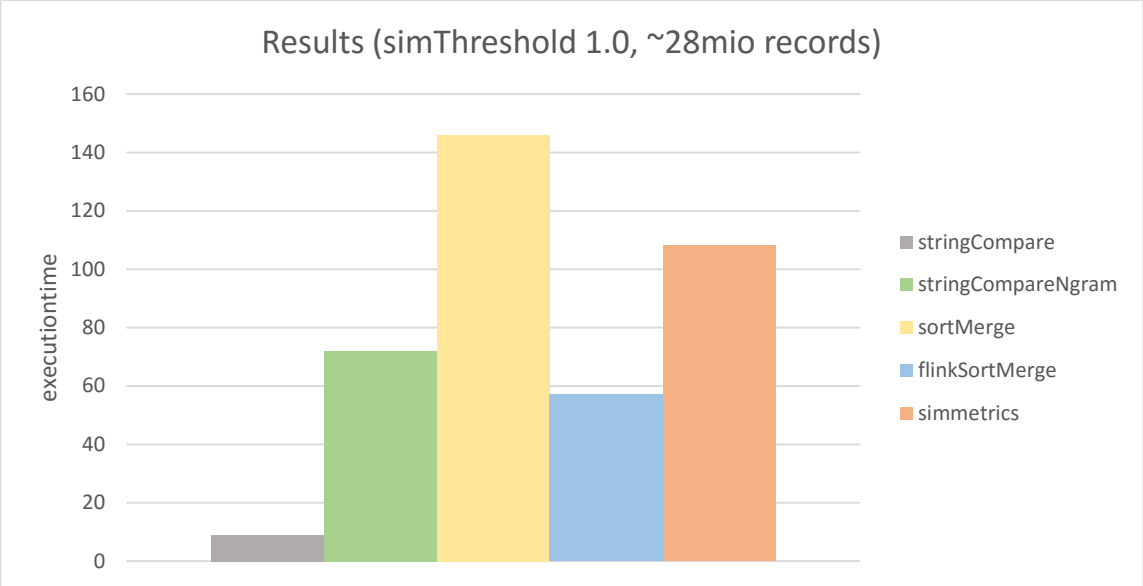
algorithm	simThreshold	records	tokensize	executiontime	source
simmetrics	0,0	28.422.030	-	109	perfect_Btrue.csv
simmetrics	0,1	2.402.933	-	108	perfect_Btrue.csv
simmetrics	0,3	177.793	-	108	perfect_Btrue.csv
simmetrics	0,5	30.032	-	106	perfect_Btrue.csv
simmetrics	0,7	10.604	-	110	perfect_Btrue.csv
simmetrics	0,9	9.806	-	108	perfect_Btrue.csv
simmetrics	1,0	9.802	-	108	perfect_Btrue.csv

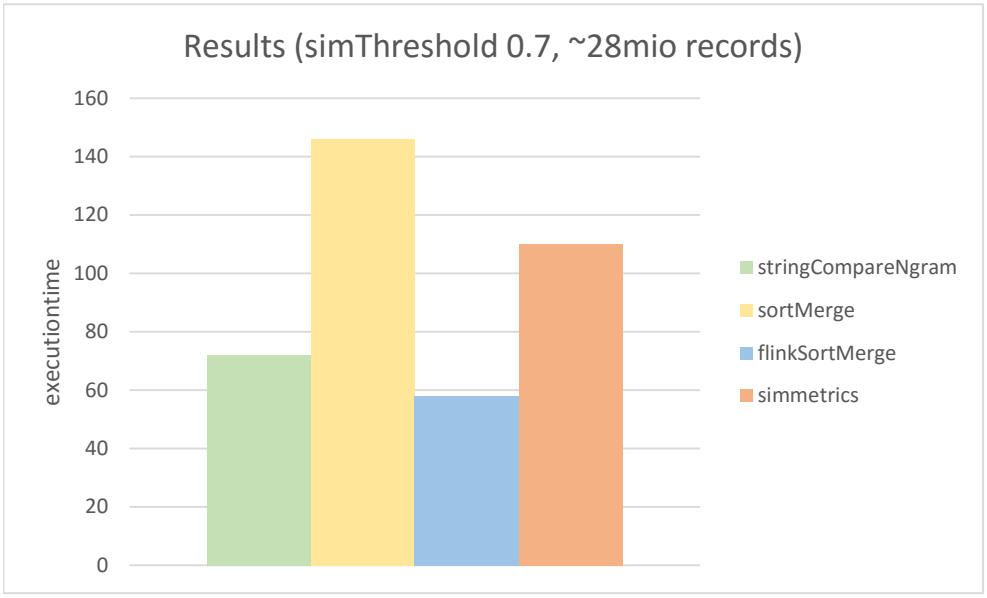
Trefferverteilung pro Threshold (gilt für alle Algorithmen mit Threshold)



Trefferverteilung pro Tokengröße (gilt für alle Algorithmen mit Tokennutzung)






Results (simThreshold 0.7, ~28mio records)
executiontime
160
140
120
100
80
60
40
20
0
stringCompareNgram
sortMerge
flinkSortMerge
simmetrics