# Trefferwahrscheinlichkeit .

## **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:

# . c.mer betraentan

Semantischer	Kein Semantischer Zusammenhang+Treffer:
Zusammenhang+Treffer:	Bsp:
Bsp:	starkelementaryschool; sabinelementaryschool; 0.79
zermatt;zermatt; 1.0	schoolnumber5;schoolnumber1;0.8
	southebby;southemarin,0.72
	Fehlerklasse 1
Semantischer	Kein semantischer Zusammenhand+kein Treffer:
Zusammenhang+kein Treff	er: Bsp:
Bsp:	denver;newyork; 0.0
austria;österreich; 0.0	
allemagne; deutschland; 0.0	
Fehlerklasse 2	

### Semantik

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

simThreshold	records	tokensize	executiontime	source	
0	28.422.030	-		perfect Btrue.csv	
1		-		perfect_Btrue.csv	
_	0.002		_	pa	
simThreshold	records	tokensize	executiontime	source	Verarbeitungszeitentwicklung
0,0		3		perfect Btrue.csv	
					120
				_	100
				_	
					į į
				_	Light 60
1,0	5.002	3	12	perrect_btrue.csv	executiontime 60 09 04 04 04 04 04 04 04 04 04 04 04 04 04
simThreshold	records	tokensize	evecutiontime	source	<b>™</b> 40
					20
					0
				_	
0,7	9.819	10	97	perfect_Btrue.csv	
cimThroshold	rocords	tokonsize -	ovocutiontime.	cource	Vorarhoitungszaitantwicklurs
					Verarbeitungszeitentwicklung
				_	
				_	
					200
				_	<u>§</u> 150
					on the state of th
1,0	9.802	3	146	perfect_Btrue.csv	
					₩ 100 /
0,7					
0,7	12.246		131	perfect_Btrue.csv	
0,7	10.604		146	perfect_Btrue.csv	0
0,7	9.960	5	158	perfect_Btrue.csv	0 2 4 6 8 10 1
0,7	9.825	8	194	perfect_Btrue.csv	tokensize
0,7	9.819	10	216	perfect_Btrue.csv	
simThreshold	records	tokensize	executiontime	source	Verarbeitungszeitentwicklung
0,0	28.422.030				
0,1	2.402.933		59	perfect_Btrue.csv	
0,3	177.793	3	58	perfect_Btrue.csv	60
0,5	30.032	3	58	perfect_Btrue.csv	50
0,7	10.604	3	58	perfect_Btrue.csv	
0,9	9.806	3	57	perfect_Btrue.csv	140 High 40
1,0	9.802	3	57	perfect_Btrue.csv	nie die die die die die die die die die d
					executiontime 40
simThreshold	records	tokensize	executiontime	source	20
0,7	189.480	1	57	perfect_Btrue.csv	
	12.246	2		perfect Btrue.csv	10
0,7				perfect_Btrue.csv	
	10.604	3			
0,7	10.604	3 5			
			57	perfect_Btrue.csv	0 2 4 6 8 10 1
0,7 0,7 0,7	10.604 9.960	5	57 58	perfect_Btrue.csv perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7	10.604 9.960 9.825	5 8	57 58	perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7	10.604 9.960 9.825 9.819	5 8 10	57 58 59	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7 0,7	10.604 9.960 9.825 9.819	5 8 10 tokensize	57 58 59 executiontime	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv source	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7 0,7 simThreshold 0,0	10.604 9.960 9.825 9.819 records 28.422.030	5 8 10 tokensize	57 58 59 executiontime	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv source perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7 0,7 simThreshold 0,0 0,1	10.604 9.960 9.825 9.819 records 28.422.030 2.402.933	5 8 10 tokensize	57 58 59 executiontime 109 108	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv source perfect_Btrue.csv perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7 0,7 simThreshold 0,0 0,1 0,3	10.604 9.960 9.825 9.819 records 28.422.030 2.402.933 177.793	5 8 10 tokensize - -	57 58 59 executiontime 109 108 108	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv source perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7 0,7 simThreshold 0,0 0,1 0,3 0,5	10.604 9.960 9.825 9.819 records 28.422.030 2.402.933 177.793 30.032	5 8 10 tokensize - - -	57 58 59 executiontime 109 108 108	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv source perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
0,7 0,7 0,7 0,7 0,7 simThreshold 0,0 0,1 0,3	10.604 9.960 9.825 9.819 records 28.422.030 2.402.933 177.793 30.032 10.604	5 8 10 tokensize - - -	57 58 59 executiontime 109 108 108 106	perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv source perfect_Btrue.csv perfect_Btrue.csv perfect_Btrue.csv	0 2 4 6 8 10 1 tokensize
	0,7 0,7 0,7 0,7 0,7 0,7 0,7 0,7 0,7 simThreshold 0,0 0,1 0,3 0,5 0,7 0,9 1,0 simThreshold 0,7 0,7 0,7 0,7 0,7 0,7 0,7 0,7 0,7 0,7	0,3 177.793 0,5 30.032 0,7 10.604 0,9 9.806 1,0 9.802    SimThreshold   records	0,3	0,3         177.793         3         73           0,5         30.032         3         73           0,7         10.604         3         72           0,9         9.806         3         72           1,0         9.802         3         72           1,0         9.802         3         72           simThreshold         records         tokensize         executiontime           0,7         12.246         2         74           0,7         10.604         3         75           0,7         9.960         5         80           0,7         9.825         8         89           0,7         9.819         10         97           simThreshold records         tokensize         executiontime           0,0         28.422.030         3         154           0,1         2.402.933         3         146           0,5         30.032         3         146           0,7         10.604         3         146           0,9         9.806         3         147           1,0         9.802         3         146           simThr	0,3







