Framework for Syntactic String Similarity Measures

Najlah Gali, Radu Mariescu-Istodor, Damien Hostettler, Pasi Fränti

24.4.2019

N. Gali, R. Mariescu-Istodor, D. Hostettler and P. Fränti,
"Framework for syntactic string similarity measures",
Expert Systems with Applications, 2019.

Introduction

Application examples

Titles of web pages:

V-café
Viet-Café

Place names:

Ting Tsi River
Tingtze River

Keywords and keyphrases:

Theater
theatre

Ontology alignments:

associate professor
senior lecturer

Named entities:

U.S State Department
US Department of State

Short segments of text:

Apple computer
Apple pie

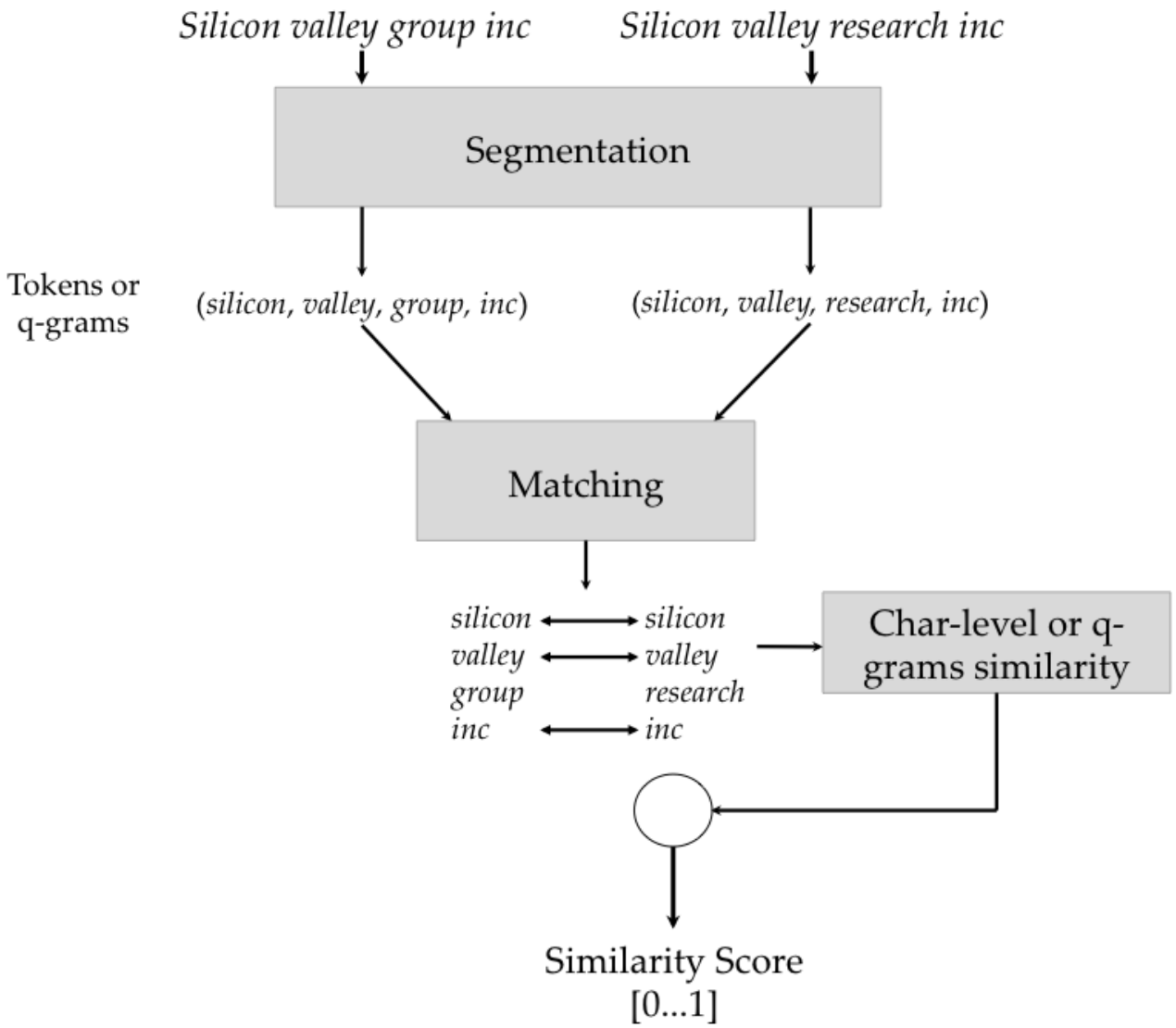
Personal names:

Gail Vest
Gayle Vesty

Sentences:

I haven't watched television for ages
It's been a long time since I watched television

Similarity framework



Existing packages

Year	Package	Language	Type	Measures	Source
2003	SecondString[3]	Java	Character Token Soft	38	Cohen et al. (2003)
2005	SimMetric [4]	Java	Character Q-gram Token	23	---
2013	DKPro[5]	Java	Character Q-gram Token, Soft	20	Bär et al. (2013)
2014	Stringdist[6]	C	Character Q-gram	10	Van der Loo (2014)
2016	Harry[7]	C	Character Token	21	Rieck & Wressnegger (2016)
2017	StringSim[8]	Java	Character Q-gram Token, Soft	143	Gali et al (2019)

[3] https://sourceforge.net/projects/secondstring
[4] https://sourceforge.net/projects/simmetrics
[5] https://dkpro.github.io/dkpro-similarity
[6] http://www.markvanderloo.eu/yaRb/category/string-metrics
[7] http://www.mlsec.org/harry
[8] http://cs.uef.fi/sipu/soft/stringsim

StringSim package

Existing measure

√ New combination

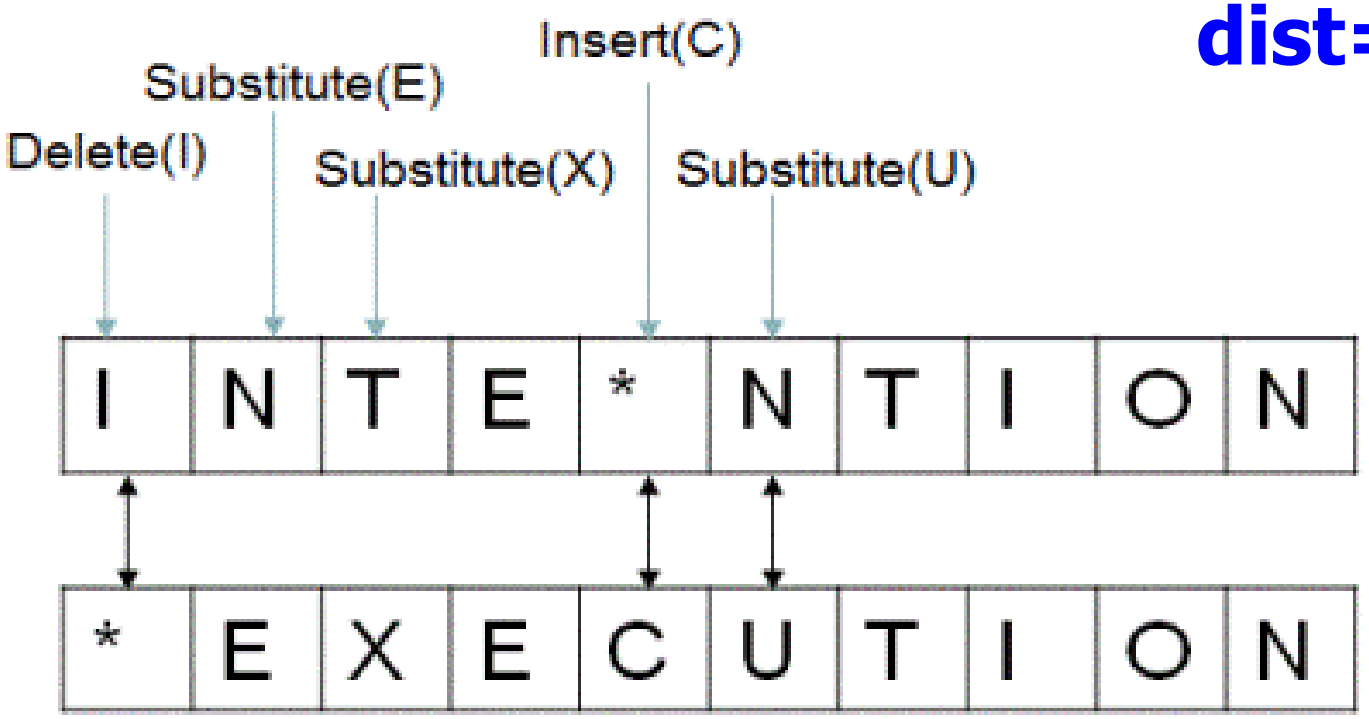
			Token-level									
			Set-matching						Bag-of-tokens			Seq.
		Ch/Q	Bra-Ban	Simpson	Jacc	Dice	Rouge	Mon-Elk	Cos	Eucl	Manh	Edit
	Exact match											
Character-level	Hamming		√	√	√	√	√	√	√	√	√	√
	Levenshtein						√			√	√	
	Dam-Levenshtein		√	√	√	√	√	√	√	√	√	√
	Needle-Wunsch		√	√	√	√	√	√	√	√	√	√
	SW		√	√	√	√	√	√	√	√	√	√
	SWG		√	√	√	√	√		√	√	√	√
	Jaro						√			√	√	
	Jaro-Winkler		√	√			√	√	√	√	√	√
	LCS		√	√	√	√	√	√	√	√	√	√
Grams	2-Grams						√			√	√	
	3-Grams		√	√	√	√	√	√	√	√	√	√
Semantic	Word2Vec		√	√			√	√				√

Character-level measures

- Exact match
- Transformation
- Longest common substring (LCS)

Edit distance

Levenshtein 1966



dist=5

Solved by dynamic programming algorithm

Character-level measures

Similarity measure	Equation	Edit operation costs			
		Insert	Delete	Substitute	Swap
Levenshtein (1966)	$1 - \frac{edit(s_1, s_2)}{max(\|s_1\|, \|s_2\|)}$	1	1	1	-
Damerau-Levenshtein (Damerau 1964)	$1 - \frac{edit(s_1, s_2)}{max(\|s_1\|, \|s_2\|)}$	1	1	1	1
Needleman and Wunsch (1970)	$1 - \frac{edit(s_1, s_2)}{2 \times max(\|s_1\|, \|s_2\|)}$	variable	variable	1	-
Smith and Waterman (1981)	$\frac{edit(s_1, s_2)}{min(\|s_1\|, \|s_2\|)}$	variable	variable	-2	-
Smith-Waterman-Gotoh (Gotoh 1982)	$\frac{edit(s_1, s_2)}{min(\|s_1\|, \|s_2\|)}$	variable	variable	-3 +3	-
Hamming (1950)	$1 - \frac{edit(s_1, s_2)}{max(\|s_1\|, \|s_2\|)}$	-	-	1	-
Jaro (1989)	$\frac{1}{3} \times \left(\frac{m}{\|s_1\|} + \frac{m}{\|s_2\|} + \frac{m - x}{m} \right)$	-	-	-	-
Jaro-Winkler (Winkler 1990)	$J(s_1, s_2) + (l \times p(1 - J(s_1, s_2))$	-	-	-	-
Longest common substring (Friedman and Sideli 1992)	$\frac{\|sub(s_1, s_2)\|}{max(\|s_1\|, \|s_2\|)}$	-	-	-	-

String segmentation

- Tokenization
- Q-grams

Segmentation examples

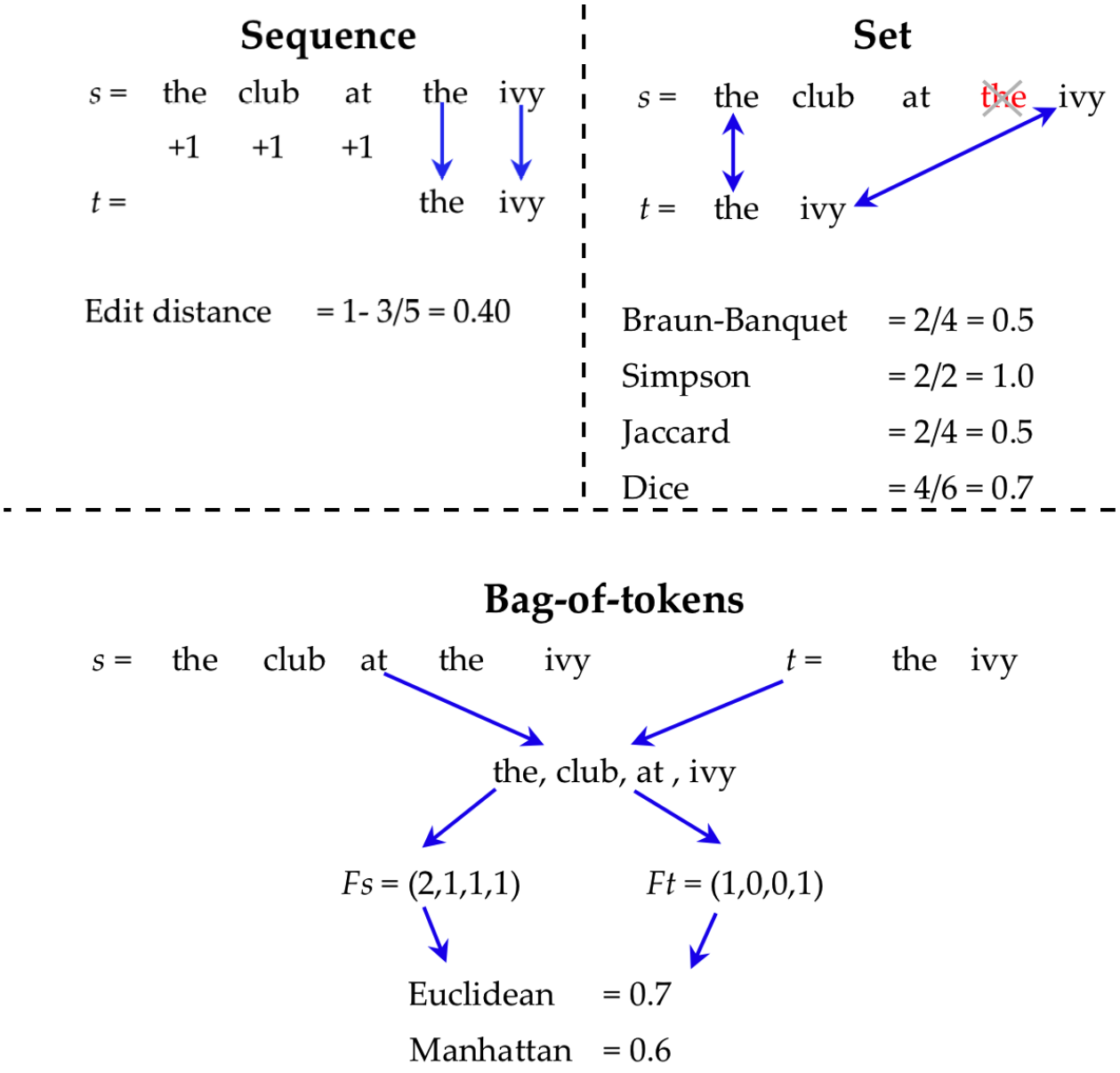
The club at the Ivy

Segmentation method	**Output**
None (char sequence)	the club at the ivy
q-grams (q = 3)	the, he_, e_c, _cl, clu, lub, ub_ , b_a, _at, at_, t_t, _th, the, he_ , e_i, _iv, ivy
q-grams with padding	##t, #th, the, he_ , e_c, _cl, clu, lub, ub_ , b_a, _at, at_ , t_t, _th, the, he_ , e_i, _iv, ivy, vy%, y%%
1-skip-grams	t*e, h*c, e*l, c*u, l*b, u*a, b*t, a*t, t*h, t*e, h*i, e*v, i*y
Tokenization	the, club, at, the, ivy

Matching techniques

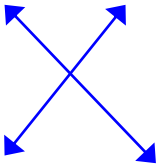
- Sequence matching
- Set matching
- Bag-of-tokens

String matching at token level



Problem of crisp sets

gray color



color gray

Similarity = 1.0

gray color

?

colour grey

Similarity = 0.0

Soft set-matching

Smith-Waterman-Gotoh

	the	grey	colour
gray	0.20	0.90	0.30
color	0.20	0.30	0.80
Max.	**0.20**	**0.90**	**0.80**

Similarity = $\frac{1}{3} \cdot (0.2 + 0.9 + 0.8) = 0.63$

Soft cardinalities of sets

{gray, grey} …. {gray, color}

$$|T|_{soft} = \sum_{i=1}^{n} \left[\frac{1}{\sum_{j=1}^{n} d(T^i, T^j)} \right]$$

	gray	grey	**Sum**	**1/sum**
gray	1.00	0.90	1.90	0.53
grey	0.90	1.00	1.90	0.53
			$\|T\|_{soft}$	1.06

	gray	color	**Sum**	**1/sum**
gray	1.00	0.30	1.30	0.77
color	0.30	1.00	1.30	0.77
			$\|T\|_{soft}$	1.54

Soft cardinalities of sets

{gray, grey} …. {gray, color}

$$|T_1 \cap T_2|_{soft} = |T_1|_{soft} + |T_2|_{soft} - |T_1 \cup T_2|_{soft}$$

$$Jaccard(T_1, T_2) = \frac{|T_1 \cap T_2|_{soft}}{|T_1 \cup T_2|_{soft}}$$

Another example

	gray	color	the	grey	colour	**Sum**	**1/sum**
gray	1.00	0.30	0.20	0.90	0.30	2.70	0.37
color	0.30	1.00	0.20	0.30	0.80	2.60	0.38
the	0.20	0.20	1.00	0.33	0.20	1.93	0.52
grey	0.90	0.30	0.33	1.00	0.30	2.83	0.35
colour	0.30	0.80	0.20	0.30	1.00	2.60	0.38
						$\lvert T_1 \cup T_2 \rvert_{soft}$	2.01

Summary of measures

Sequence and set-matching

Sequence		Soft variant
Chaudhuri et al. (2003)	$\mathrm{sim}_{ij} = \begin{cases} \mathrm{sim}_{i-1,j-1} & if\ T_1^i = T_2^j \\ \min \begin{cases} \mathrm{sim}_{i-1,j} + 1 \\ \mathrm{sim}_{i,j-1} + 1 \\ \mathrm{sim}_{i-1,j-1} + 1 \end{cases} & otherwise \end{cases}$	$\mathrm{sim}_{ij} = \begin{cases} \mathrm{sim}_{i-1,j-1} & if\ T_1^i = T_2^j \\ min \begin{cases} \mathrm{sim}_{i-1,j} + (1 - d(T_1^i, T_2^j)) \\ \mathrm{sim}_{i,j-1} + (1 - d(T_1^i, T_2^j)) \\ \mathrm{sim}_{i-1,j-1} + (1 - d(T_1^i, T_2^j)) \end{cases} & otherwise \end{cases}$
Set		**Soft variant**
Braun- Banquet (Choi et al., 2010)	$\frac{\lvert T_1 \cap T_2 \rvert}{\max(\lvert T_1 \rvert, \lvert T_2 \rvert)}$	$\frac{\lvert T_1 \cap T_2 \rvert_{soft}}{\max(\lvert T_1 \rvert_{soft}, \lvert T_2 \rvert_{soft})}$
Simpson (Choi et al., 2010)	$\frac{\lvert T_1 \cap T_2 \rvert}{\min(\lvert T_1 \rvert, \lvert T_2 \rvert)}$	$\frac{\lvert T_1 \cap T_2 \rvert_{soft}}{\min(\lvert T_1 \rvert_{soft}, \lvert T_2 \rvert_{soft})}$
Jaccard (Rezaei and Fränti, 2016)	$\frac{\lvert T_1 \cap T_2 \rvert}{\lvert T_1 \cup T_2 \rvert}$	$\frac{\lvert T_1 \cap T_2 \rvert_{soft}}{\lvert T_1 \cup T_2 \rvert_{soft}}$
Dice (Brew and McKelvie, 1996)	$\frac{2 \times \lvert T_1 \cap T_2 \rvert}{\lvert T_1 \rvert + \lvert T_2 \rvert}$	$\frac{2 \times \lvert T_1 \cap T_2 \rvert_{soft}}{\lvert T_1 \rvert_{soft} + \lvert T_2 \rvert_{soft}}$
Rouge-N (Lin, 2004)	$\left(\left(\frac{1}{p}\right) + \left(\frac{1}{r}\right)\right)^{-1}$ $p = \frac{\lvert [T_1] \cap [T_2] \rvert}{\lvert [T_1] \rvert}, r = \frac{\lvert [T_1] \cap [T_2] \rvert}{\lvert [T_2] \rvert}$	$\left(\left(\frac{1}{p}\right) + \left(\frac{1}{r}\right)\right)^{-1}$ $p = \frac{\lvert [T_1] \cap [T_2] \rvert_{soft}}{\lvert [T_1] \rvert_{soft}}, r = \frac{\lvert [T_1] \cap [T_2] \rvert_{soft}}{\lvert [T_2] \rvert_{soft}}$
Monge-Elkan (1996)		$\frac{1}{\lvert [T_1] \rvert} \sum_{i=0}^{\lvert [T_1] \rvert} \max_{1 \le j \le \lvert [T_2] \rvert} d(T_1^i, T_2^j)$

Summary of measures

Bag-of-tokens

	Bag-of-tokens	Soft variant
Cosine (Cohen et al., 2003b)	$\frac{\sum_{i=1}^{n} \boldsymbol{v}_1^i \boldsymbol{v}_2^i}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{v}_1^i)^2} \sqrt{\sum_{i=1}^{n} (\boldsymbol{v}_2^i)^2}}$	$\frac{\sum_{i,j=1}^{n} d(T_1^i, T_2^j) \boldsymbol{v}_1^i \boldsymbol{v}_2^j}{\sqrt{\sum_{i,j=1}^{n} d(T_1^i, T_1^j) \boldsymbol{v}_1^i \boldsymbol{v}_1^j} \sqrt{\sum_{i,j=1}^{n} d(T_2^i, T_2^j) \boldsymbol{v}_2^i \boldsymbol{v}_2^j}}$
Euclidean (Malakasiotis and Androutsopoulos, 2007)	$1 - \frac{\sqrt{\sum_{i=1}^{n} (\boldsymbol{v}_1^i - \boldsymbol{v}_2^i)^2}}{\sqrt{\lvert \boldsymbol{v}_1^i \rvert^2 + \lvert \boldsymbol{v}_2^i \rvert^2}}$	$1 - \frac{\sqrt{\sum_{i,j=1}^{n} d(T_1^i, T_2^j) (\boldsymbol{v}_1^i - \boldsymbol{v}_2^j)^2}}{\sqrt{\left(\sum_{i,j=1}^{n} d(T_1^i, T_1^j) \boldsymbol{v}_1^i \boldsymbol{v}_1^j\right)^2 + \left(\sum_{i,j=1}^{n} d(T_2^i, T_2^j) \boldsymbol{v}_2^i \boldsymbol{v}_2^j\right)^2}}$
Manhattan (Malakasiotis and Androutsopoulos, 2007)	$1 - \frac{\sum_{i=1}^{n} \lvert \boldsymbol{v}_1^i - \boldsymbol{v}_2^i \rvert}{\lvert \boldsymbol{v}_1^i \rvert + \lvert \boldsymbol{v}_2^i \rvert}$	$1 - \frac{\sum_{i,j=1}^{n} d(T_1^i, T_2^j) \lvert \boldsymbol{v}_1^i - \boldsymbol{v}_2^j \rvert}{\sum_{i,j=1}^{n} d(T_1^i, T_1^j) \boldsymbol{v}_1^i \boldsymbol{v}_1^j + \sum_{i,j=1}^{n} d(T_2^i, T_2^j) \boldsymbol{v}_2^i \boldsymbol{v}_2^j}$

Results

Datasets

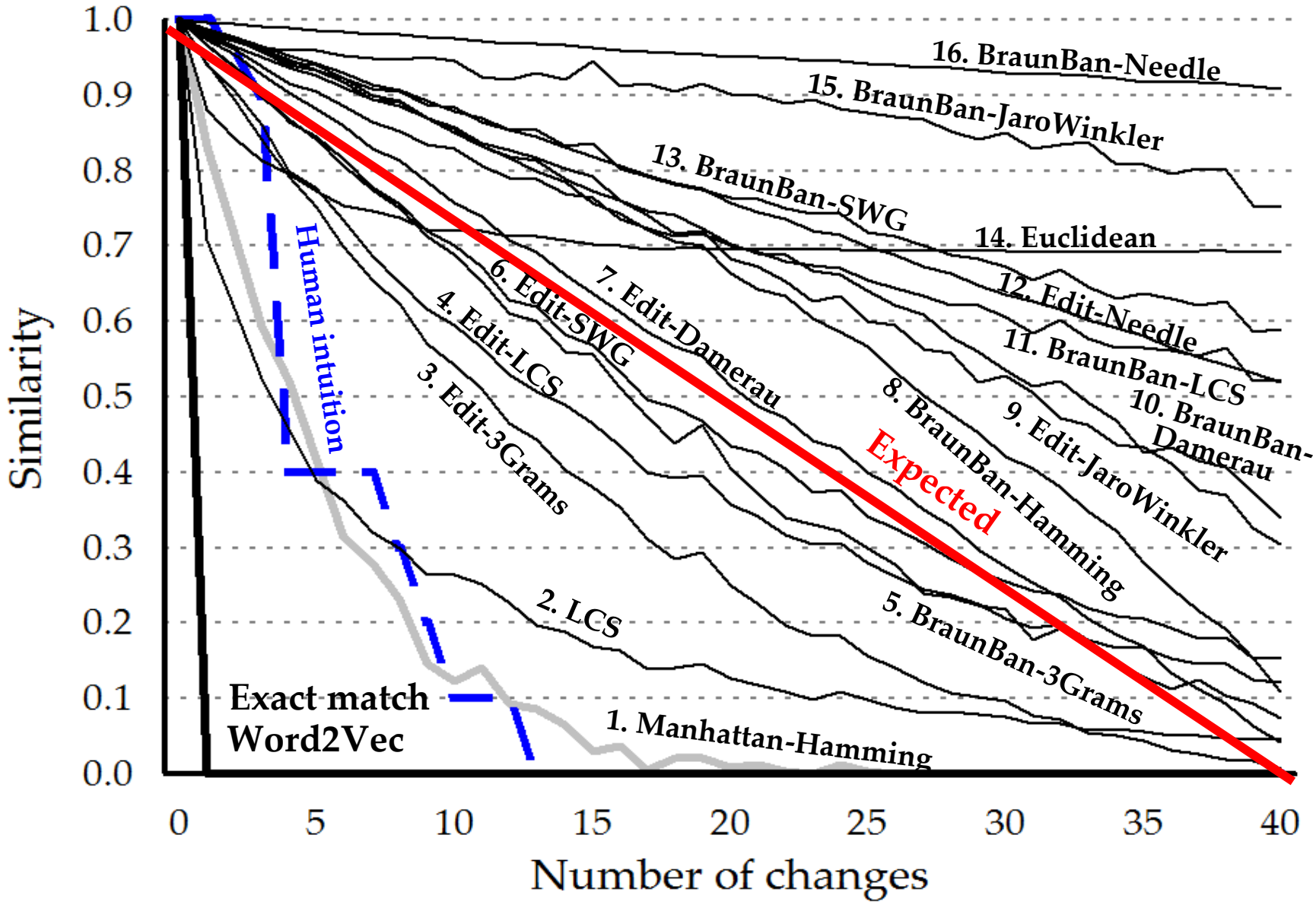
Source	Data set	Size	Language	String length					
				Token			Character		
				Min	Av.	Max	Min	Av.	Max
Gali et al. (2017)	Titler	4,968	English	1	3	8	4	14	39
Gali et al. (2019)	Mopsi photos	1,000	English Finnish	1	3	26	6	17	65
Cohen et al. (2003)	Bird Nybird	982	English	1	3	69	4	21	321
	Bird Scott1	38		2	3	8	7	20	58
	Bird Scott2	719		3	4	9	15	35	83
	Business	2,139		1	3	8	4	19	51
	Game	855		1	5	55	4	27	255
	Park	654		2	3	12	6	16	58
	Restaurant	863		7	11	21	40	59	102

Text manipulation

Character changes

	Ch/ Q	Edit	MongeElkan	Brau-Ban	Simpson	Jaccard	Dice & Rouge	Cosine	Manhattan	Euclidean
Exact match						2				
Hamming										
Levenshtein & Damerau-Levenshtein				6	12	6	9	6		
Needleman Wunch				14						
Smith Waterman & SWG	2		6	10	15	9	12	15		
Jaro	7			11	13	12			4	10
Jaro Winkler	8									
LCS	1			9	12	6	9	6		
2Grams					5	2	3			
3Grams				3						
Word2Vec										

Effect of char changes

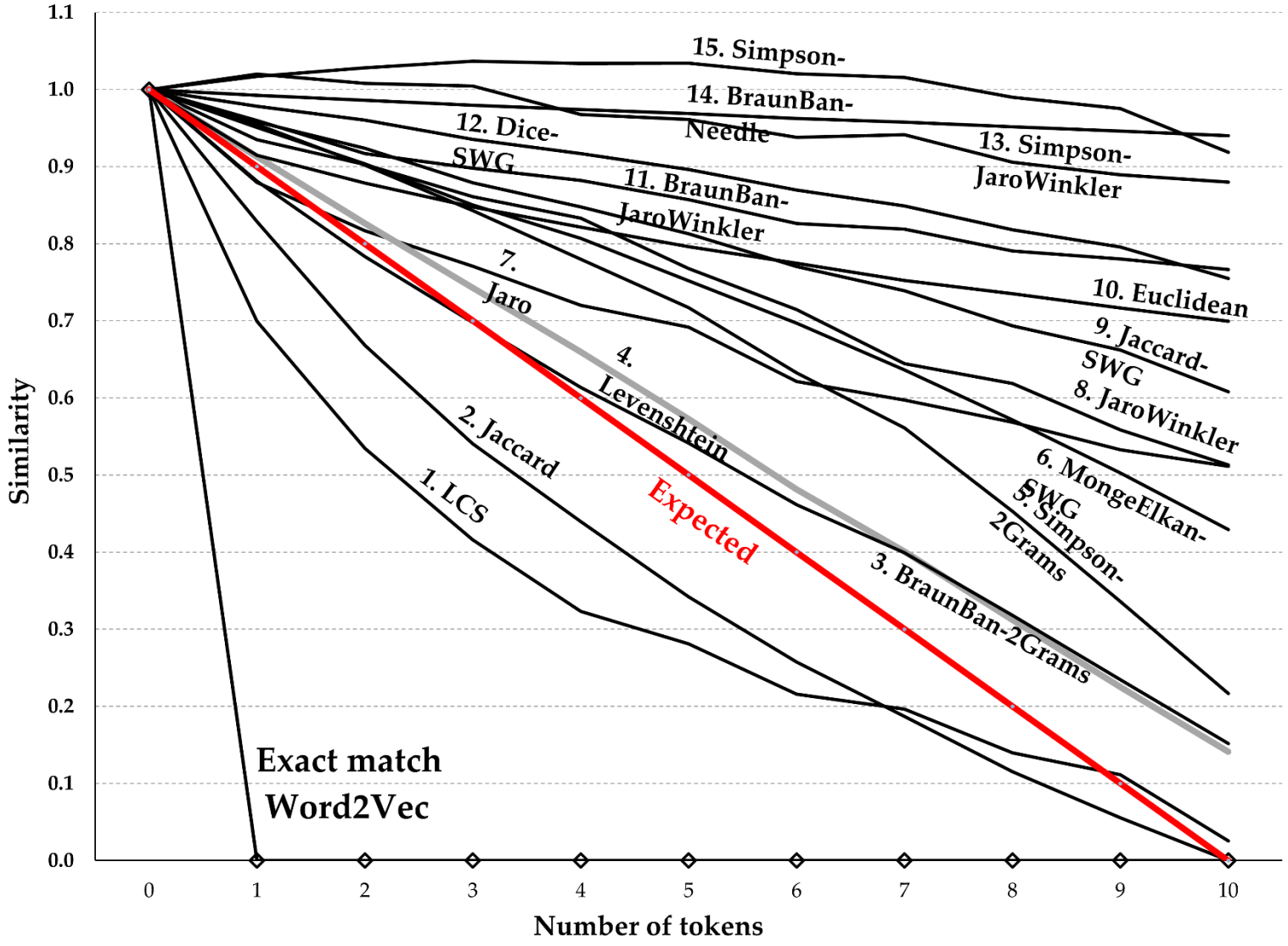


Text manipulation

Token changes

	Ch/ Q	Edit	MongeElkan	Brau-Ban	Simpson	Jaccard	Dice & Rouge	Cosine	Manhattan	Euclidean
Exact match						2				
Hamming										
Levenshtein & Damerau-Levenshtein				6	12	6	9	6		
Needleman Wunch				14						
Smith Waterman & SWG	2		6	10	15	9	12	15		
Jaro	7			11	13	12			4	10
Jaro Winkler	8									
LCS	1			9	12	6	9	6		
2Grams					5	2	3			
3Grams				3						

Effect of token changes



Correlation to human intuition

			Token-level									
			Set-matching						Bag-of-tokens			Seq.
		Ch/Q	Bra-Ban	Simpson	Jacc	Dice	Rouge	Mon-Elk	Cos	Eucl	Manh	Edit
	Exact match	40	46	14	46	45	45	46	48	44	46	48
Character-level	Hamming	41	47	14	48	47	47	48	49	44	46	50
	Levenshtein	52	48	7	49	48	48	50	49	44	46	52
	Dam-Levenshtein	52	48	6	49	48	48	50	49	44	46	52
	Needle-Wunsch	49	43	4	45	34	34	48	42	43	47	51
	SW	16	46	-1	46	44	44	49	45	43	46	51
	SWG	16	44	-4	44	40	40	47	43	42	47	51
	Jaro	51	43	-1	42	39	39	47	43	44	47	49
	Jaro-Winkler	46	43	-1	42	39	39	46	43	44	47	49
	LCS	47	47	6	48	47	47	50	48	44	46	52
Grams	2-Grams	51	49	13	50	**50**	**50**	50	52	44	46	52
	3-Grams	52	50	14	50	**50**	**50**	50	51	44	46	52
Semantic	Word2Vec	4	34	-5	34	34	34	35	34	36	26	36

Qualitative examples

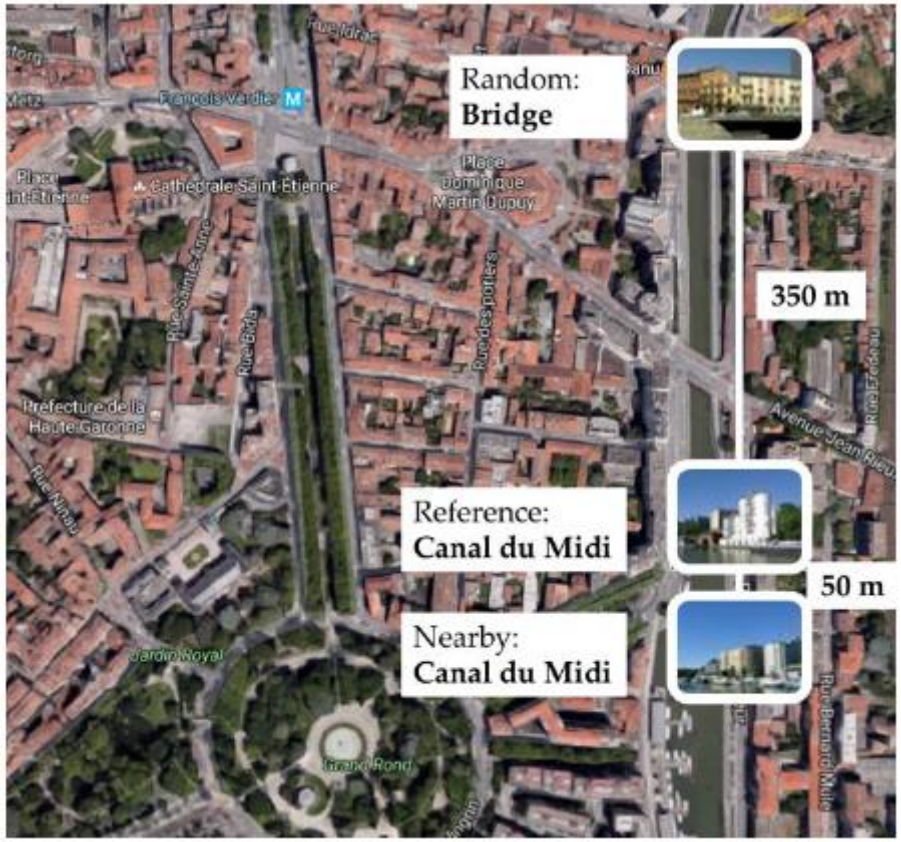
Excellent match

- Freda's – Fredas
- Drom UK – Dröm UK
- Hot Spring – HotSpring
- Park Hotel and Spa – Park Hotel & Spa
- Holiday Inn Bristol Filton – Holiday Inn Filton-Bristol

Poor match

- Out of the Blue – Out the Blue
- Arcata Pizzeria – At Arcata Pizzeria
- 3 Degrees – Degrees

Correlation to distance



Correlation to distance

			Token-level									
			Set-matching						Bag-of-tokens			Seq.
		Ch/Q	Bra-Ban	Simpson	Jacc	Dice	Rouge	Mon-Elk	Cos	Eucl	Manh	Edit
	Exact match	62	65	65	65	65	65	65	65	67	65	65
Character-level	Hamming	63	67	66	66	66	67	66	66	67	65	65
	Levenshtein	70	67	67	67	67	67	67	69	65	65	62
	Dam-Levenshtein	70	67	67	67	67	67	66	68	65	64	65
	Needle-Wunsch	69	70	65	70	70	70	70	70	67	65	62
	SW	62	68	68	69	69	69	69	68	66	65	70
	SWG	62	72	69	72	72	72	67	70	65	65	67
	Jaro	64	67	59	67	67	67	67	66	65	65	62
	Jaro-Winkler	64	67	60	67	67	67	67	67	65	65	61
	LCS	67	69	68	69	69	69	70	71	66	66	65
Grams	2-Grams	70	71	69	70	**70**	**70**	70	70	69	65	70
	3-Grams	67	72	70	72	**72**	**72**	71	71	69	65	68
Semantic	Word2Vec	62	73	73	73	73	73	74	67	62	73	73

Clustering experiment

- 180 photos
- 15 clusters

Keyword: *talo*

Keywords: *kahvi, cafe, kafe*

Keyword: *hotel*

Clustering results

			Token-level									
			Set-matching						Bag-of-tokens			Seq.
		Ch/Q	Bra-Ban	Simpson	Jacc	Dice	Rouge	Mon-Elk	Cos	Eucl	Manh	Edit
	Exact match	47	63	74	67	66	66	66	67	58	66	63
Character-level	Hamming	42	60	69	59	69	69	71	69	58	69	61
	Levenshtein	64	63	72	66	62	62	68	69	61	68	63
	Dam-Levenshtein	64	58	71	66	70	70	68	72	61	69	63
	Needle-Wunsch	53	61	76	59	61	61	67	66	60	70	64
	SW	78	62	70	66	70	70	74	75	59	66	70
	SWG	72	60	62	63	64	64	73	67	61	65	65
	Jaro	59	49	50	49	53	53	63	51	60	69	54
	Jaro Winkler	57	48	55	56	54	54	67	52	61	69	57
	LCS	67	66	74	67	78	78	74	74	58	67	66
Grams	2-Grams	71	69	81	68	**74**	**74**	73	73	56	65	67
	3-Grams	72	69	75	72	**77**	**77**	69	73	60	65	69
Semantic	Word2Vec	46	60	74	60	61	61	67	58	65	71	57

Name matching

Similarity (%)	String 1	String 2	Key 1	Key 2
100	Hyperstudio	Hyperstudio	hyperstudio	hyperstudio
90.7	Mario Teaches Typing	Mario Teaches Typing 2	mariotype	foobar
74.9	Green Eggs and Ham	Green Eggs and Ham by Dr. Seuss	greeneggs	greeneggs
69.2	Fisher Price's Pirate Ship	Pirate Ship	pirateship	pirateship
69.1	Let's Color	Let's Learn Shapes & Colors	none	foobar
58.7	Catz	Catz, Your Computer Petz	catz	catz

Name matching

			Token-level									
			Set-matching						Bag-of-tokens			Seq.
		Ch/Q	Bra-Ban	Simpson	Jacc	Dice	Rouge	Mon-Elk	Cos	Eucl	Manh	Edit
	Exact match	13	80	78	80	80	80	81	80	66	79	74
Character-level	Hamming	16	75	75	78	78	78	79	78	65	79	72
	Levenshtein	68	72	59	79	79	79	86	83	61	79	77
	Dam-Levenshtein	68	72	59	80	80	80	86	82	61	79	77
	Needle-Wunsch	58	69	56	80	80	79	85	84	57	80	70
	SW	73	63	21	62	62	61	84	62	59	80	72
	SWG	74	59	21	60	60	60	84	62	57	80	71
	Jaro	60	30	6	17	17	17	86	20	60	80	64
	Jaro Winkler	59	30	6	17	17	17	85	19	61	80	64
	LCS	65	75	69	81	81	81	86	83	62	80	77
Grams	2-Grams	76	80	78	**86**	**86**	**86**	87	83	64	78	79
	3-Grams	77	82	83	**87**	**87**	**87**	87	84	65	78	80
Semantic	Word2Vec	3	74	69	81	81	81	77	64	86	84	74

Summary of the results

Char level	Both combined	Token level	Word2Vec
Text manipulation:			
Most methods (+) LCS oversensitive (-) Q-grams oversensitive (-)	Most methods (+)	Oversensitive (-)	Oversensitive (-)
Human intuition:			
Most methods (+) Q-grams (+) Smith-Waterman/Gotoh (-)	Most token level + Q-grams (+) Edit distance + Any char level (+) Simpson + Any char level (-)	Edit distance (+) Simpson (-)	
Correlation to distance:			
Most methods (+/-) Damerau/Levenshtein (+) 2-grams (+)	Most token level + Q-grams (+) Euclidean/ Manhattan/Edit (+/-)	Most methods (+/-)	Mostly best (+)
Clustering:			
Smith-Waterman/Gotoh (+) Q-grams (+) Hamming (-)	Most token level + Q-grams (+) Bran-Ban worse (-)	Simpson (+)	
Names matching:			
Hamming (-)	Most token level + Q-grams (+) Monge-Elkan + Most char level (+) Most token level + Jaro /Winkler (-)		

Conclusions

Token level measures

- *Well-maintained databases*: ok as such
- *Free text*: soft variants improves!

Semantic similarity

- Suffers from single char changes

Recommendation:

- Dice or Rouge (token level) + Q-grams
- No single measure work for all applications

http://cs.uef.fi/sipu/soft/stringsim

The end