Lösung - Übungsblatt 1

(String Matching)

Fabian Panse

panse@informatik.uni-hamburg.de

Universität Hamburg

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Aufgabe 1: Overlap Measure und Jaccard Koeffizient

Gegeben: x ='Henri Waternoose' und y ='Henry Waternose'

Tokenbildung durch 3-grams

$X = \{$'##H','#He','Hen','enr','nri','ri_','i_W','_Wa','Wat',
'ate','ter','ern','rno','noo','oos','ose','se#','e##'$\}$

$Y = \{$'##H','#He','Hen','enr','nry','ry_','y_W','_Wa','Wat',
'ate','ter','ern','rno','nos','ose','se#','e##'$\}$

$X = \{$'##H','#He','Hen','enr','nri','ri_','i_W','_Wa','Wat',
'ate','ter','ern','rno','noo','oos','ose','se#','e##'$\}$

$Y = \{$'##H','#He','Hen','enr','nry','ry_','y_W','_Wa','Wat',
'ate','ter','ern','rno','nos','ose','se#','e##'$\}$

$\Rightarrow O(X, Y) = 13$

$\Rightarrow Jacc(X, Y) = 13/22 = 0.591$

Aufgabe 2: Levenshtein Distanz/Ähnlichkeit

Gegeben: x ='Sean' und y ='Shawn'

	ε	s	h	a	w	n
ε	**0**	1	2	3	4	5
s	1	**0**	1	2	3	4
e	2	1	**1**	2	3	4
a	3	2	2	**1**	**2**	3
n	4	3	3	2	2	**2**

$\Rightarrow$ $LevDst(x, y) = 2$

$\Rightarrow$ $LevSim(x, y) = 1 - \frac{2}{max(4,5)} = 0.6$

Aufgabe 3: Affine Gap Distance

Gegeben: x ='Martin Thomas Doe' und y ='Martin T Do'

- Kosten für Öffnen einer Lücke: $w_g = 1$
- Kosten für Weiterführen einer Lücke: $w_s = 0.2$

- Die erste Lücke l_1 umfasst den substring 'homas'
- Gesamtkosten der Lücke: $w(l_1) = 1 + 4 \times 0.2 = 1.8$

- Die zweite Lücke l_2 umfasst den substring 'e'
- Gesamtkosten der Lücke: $w(l_2) = 1$

$\Rightarrow$ Gesamtkosten: 2.8

Aufgabe 4: Soundex Code

Gegeben: x ='depardieu', y ='debando' und z ='tepadeu'

	'depardieu'	**'debando'**	**'tepadeu'**
Step 1	'dprd'	'dbnd'	'tpd'
Step 2	'd163'	'd153'	't13'
Step 3	'd163'	'd153'	't13'
Step 4	**'d163'**	**'d153'**	**'t130'**

- Mit Ausnahme des ersten Buchstaben werden alle Vorkommnisse der Buchstaben 'a', 'e', 'i', 'o', 'u', 'y', 'h', und 'w' entfernt
- Mit Ausnahme des ersten Buchstaben werden alle verbliebende Buchstaben durch Ziffern ersetzt (b,p$\rightarrow$1, r$\rightarrow$6, d$\rightarrow$3, n$\rightarrow$5)
- Alle aufeinanderfolgenen Auftreten der gleichen Ziffer werden durch ein einzelnen Auftreten ersetzt
- Der Code wird auf die Länge vier beschränkt (Auffüllen mit '0')

Aufgabe 5: Extended Jaccard

Gegeben: x ='Tom John Kim' und y ='Tim Jon'
threshold $\theta = 0.5$

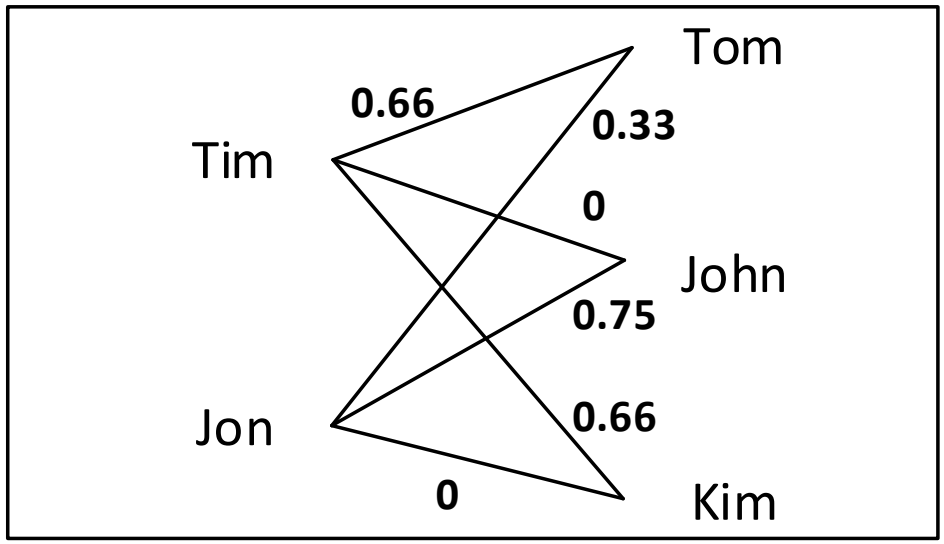
Levenshtein Ähnlichkeiten der Token

	'Tom'	'John'	'Kim'
'Tim'	2/3	0	2/3
'Jon'	1/3	3/4	0

- $shared(X, Y) = \{('Tom','Tim'),('John','Jon'),('Kim','Tim')\}$
- $unique(X) = \emptyset$
  $unique(Y) = \emptyset$
- $ExtJacc(X, Y) = \frac{3}{3+0+0} = \frac{3}{3} = 1$

Aufgabe 5: Generalized Jaccard

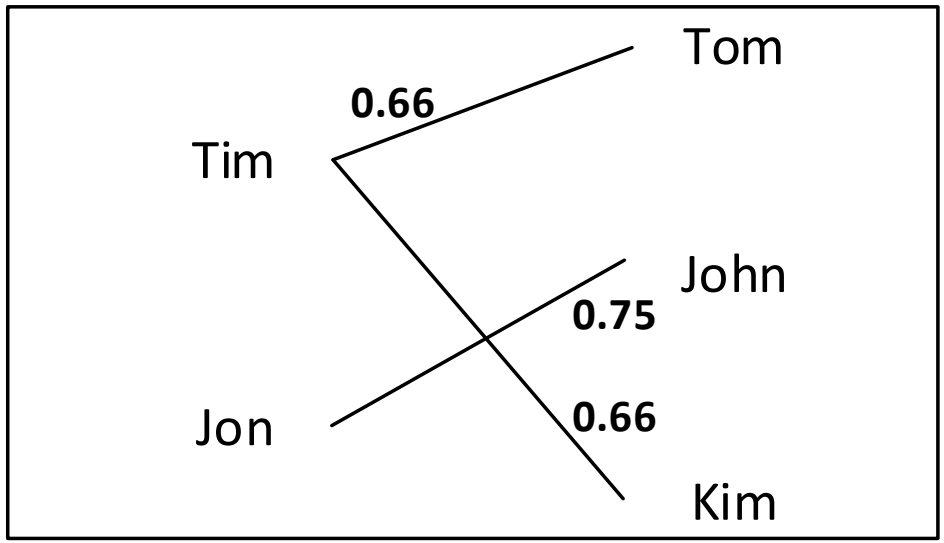
Gegeben: x ='Tom John Kim' und y ='Tim Jon'
threshold $\theta = 0.5$



$$GenJacc(X, Y) = \frac{0.66+0.75}{3+2-2} = \frac{1.41}{3} = 0.47$$

Aufgabe 5: Generalized Jaccard

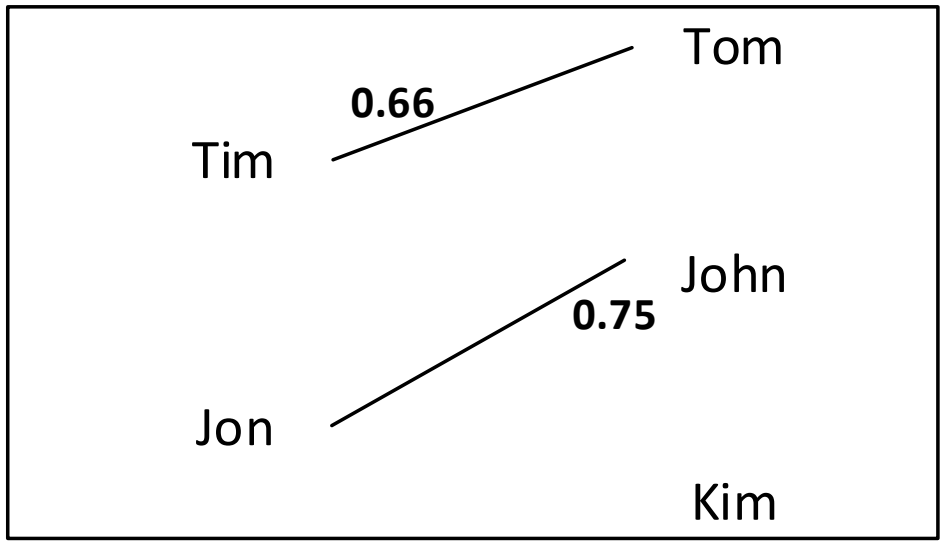
Gegeben: x ='Tom John Kim' und y ='Tim Jon'
threshold $\theta = 0.5$



$$GenJacc(X, Y) = \frac{0.66+0.75}{3+2-2} = \frac{1.41}{3} = 0.47$$

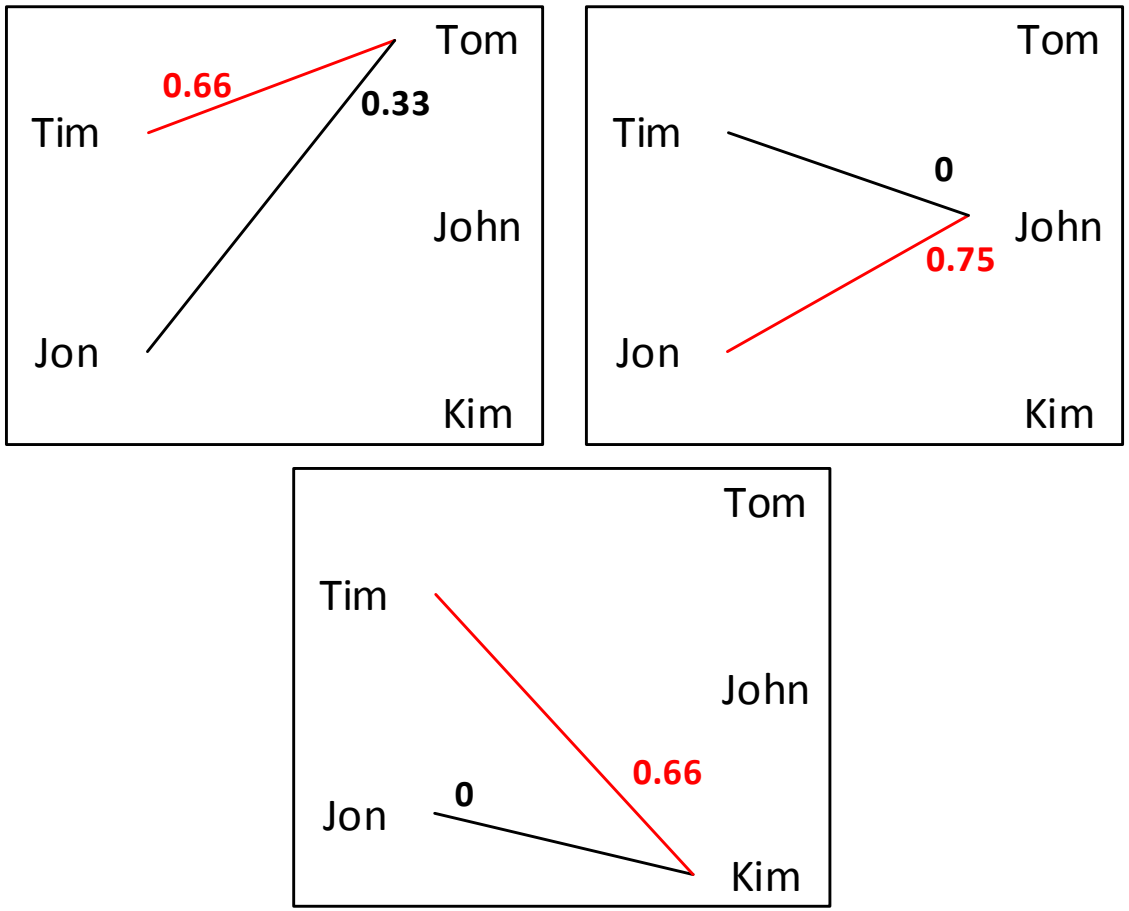
Aufgabe 5: Generalized Jaccard

Gegeben: x ='Tom John Kim' und y ='Tim Jon'
threshold $\theta = 0.5$



$$GenJacc(X, Y) = \frac{0.66+0.75}{3+2-2} = \frac{1.41}{3} = 0.47$$

Aufgabe 5: Monge-Elkan



$$MongeElkan(Y, X) = \frac{1}{3} \times (0.66 + 0.75 + 0.66) = 0.69$$

Aufgabe 6: TF/IDF

term frequency:

tf	x_1	x_2	x_3	x_4	x_5	x_6
'Insurance'	1	1	0	1	0	0
'Company'	1	0	1	0	1	0
'A&B'	0	1	0	0	1	0
'BC'	0	0	1	0	0	0
'AX'	0	0	1	1	0	0
'XY'	0	0	0	0	0	2
'Enterprises'	0	0	0	0	0	1

Aufgabe 6: TF/IDF

inverse document frequency:

idf	
'Insurance'	$6/3 = 2$
'Company'	$6/3 = 2$
'A&B'	$6/2 = 3$
'BC'	$6/1 = 6$
'AX'	$6/2 = 3$
'XY'	$6/1 = 6$
'Enterprises'	$6/1 = 6$

Aufgabe 6: TF/IDF

Kosinus Ähnlichkeit zwischen x_2 und x_4:

$$v_2 = \langle 2, 0, 3, 0, 0, 0, 0 \rangle$$
$$v_4 = \langle 2, 0, 0, 0, 3, 0, 0 \rangle$$

$$\Rightarrow \; CosSim(x_2, x_4) = (4+0+0+0+0+0+0)/(\sqrt{4+9} \times \sqrt{4+9})$$
$$= 4/(\sqrt{13} \times \sqrt{13}) = 4/13$$

Aufgabe 7: Skalierbarkeit

a) Aufbauen eines invertierten Indexes über die Token der einzelnen Stringwerte von B

Menge B

$Y_1 = \{\text{alpha, beta, delta, iota}\}$
$Y_2 = \{\text{gamma, delta}\}$
$Y_3 = \{\text{alpha, beta, zeta}\}$
$Y_4 = \{\text{alpha, gamma, iota, zeta}\}$
$Y_5 = \{\text{alpha, iota}\}$

Token in B	ID Listen
alpha	1,3,4,5
beta	1,3
gamma	2,4
delta	1,2
iota	1,4,5
zeta	3,4

- Kandidaten aufgrund Token 'alpha': {1,3,4,5}
- Kandidaten aufgrund Token 'delta': {1,2}
- Kandidaten aufgrund Token 'zeta': {3,4}
- Gesamte Menge an Kandidaten: {1,2,3,4,5} (also alle)

Aufgabe 7: Skalierbarkeit

b) Verwendung des Size Filterings

Menge B
$Y_1 = \{\text{alpha, beta, delta, iota}\}$
$Y_2 = \{\text{gamma, delta}\}$
$Y_3 = \{\text{alpha, beta, zeta}\}$
$Y_4 = \{\text{alpha, gamma, iota, zeta}\}$
$Y_5 = \{\text{alpha, iota}\}$

- Gegeben: $\theta = 0.8$ und $X = \{\text{alpha, delta, zeta}\}$
- Damit y_i ein Match von x sein kann, muss also gelten:

$$2.4 = 3 \times 0.8 \leq |Y_i| \leq 3/0.8 = 3.75$$

$\Rightarrow$ Das einzige $y_i \in B$ für welches dies gilt ist y_3

- Gesamte Menge an Kandidaten: $\{3\}$

Aufgabe 7: Skalierbarkeit

c) Verwendung des Prefix Filterings

Menge B

$Y_1 = \{\text{alpha, beta, delta, iota}\}$
$Y_2 = \{\text{gamma, delta}\}$
$Y_3 = \{\text{alpha, beta, zeta}\}$
$Y_4 = \{\text{alpha, gamma, iota, zeta}\}$
$Y_5 = \{\text{alpha, iota}\}$

Y_i	k	X_{k-1}
1	4	$\emptyset$
2	3	{alpha}
3	3	{alpha}
4	4	$\emptyset$
5	3	{alpha}

- Gegeben: $\theta = 0.8$, $|X| = 3$, Jaccard Koeffizient
- Damit y_i ein Match von x sein kann, muss also gelten:

$$|X \cap Y_i| \geq k = \lceil \frac{\theta}{1+\theta} \times (|X| + |Y_i|) \rceil = \lceil 0.444 \times (3 + |Y_i|) \rceil$$

- Y_i muss die ersten $|X| - (k-1)$ Token von X beinhalten

$\Rightarrow$ 1 und 4 scheiden aus, da sie 4 gemeinsame Token bräuchten

$\Rightarrow$ 2, 3 und 5 sind Kandidaten wenn sie 'alpha' beinhalten

- Gesamte Menge an Kandidaten: {3,5}