

# Методы оценки близости строк

Татьяна Кривошечеева

# Строковые метрики

- Расстояние Хэмминга
- Расстояние Левенштейна
- Расстояние Дамерай-Левенштейна,  
Метрика Нидлмана-Вунша,  
Метрика Смита-Вотермана
- Bag distance
- Метрики Jaro, Jaro-Winkler
- q-grams, skip-grams
- Общий префикс
- Наибольшая общая подстрока
- Метрика Monge-Elkan

# Операции преобразования строк

- Подстановка  $\text{kill} \rightarrow \text{bill}$
- Вставка  $\text{kill} \rightarrow \text{skill}$
- Удаление  $\text{fear} \rightarrow \text{ear}$

# 1. Расстояние Хэмминга (подстановка)

$$d_H(\text{GCAT, CGAT}) = 2$$

# 2. Расстояние Левенштейна (удаление, вставка, подстановка)

$$d_E(\text{CGACG, GTCGA}) = 3$$

# Подсчет расстояния Левенштейна

j


	“	T	E	S	T
“					
S					
E					
T					

# Подсчет расстояния Левенштейна


0

	“	Т	Е	С	Т
0	“	0			
С					
Е					
Т					

# Подсчет расстояния Левенштейна

	“	Т	Е	S	Т
“	0 				
S	1				
Е					
Т					

# Подсчет расстояния Левенштейна

	“	Т	Е	С	Т
“	0				
С	1 				
Е	2				
Т					



# Подсчет расстояния Левенштейна

	“	Т	Е	S	Т
“	0				
S	1				
Е	2				
Т	3				

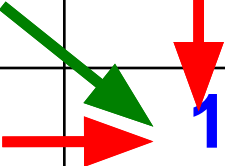
A green arrow points down from the cell containing '2' to the cell containing '3'. The number '3' is colored blue.

# Подсчет расстояния Левенштейна

	“	Т	Е	С	Т
“	0	1	2	3	4
С	1				
Е	2				
Т	3				

# Подсчет расстояния Левенштейна

	“	Т	Е	С	Т
“	0	1	2	3	4
С	1	1			
Е	2				
Т	3				



The diagram illustrates the calculation of the Levenshtein distance between the words "С" and "Т". The table shows the distance values for each character pair. A green arrow points from the cell (row "С", column "Т") to the cell (row "С", column "Т"), indicating the current step in the calculation. A red arrow points from the cell (row "С", column "Т") to the cell (row "С", column "Т"), indicating the current step in the calculation. A blue arrow points from the cell (row "С", column "Т") to the cell (row "С", column "Т"), indicating the current step in the calculation.

# Подсчет расстояния Левенштейна

	“	Т	Е	С	Т
“	0	1	2	3	4
С	1	1	2	2	3
Е	2	2	1	2	3
Т	3	2	2	2	2

The diagram illustrates the calculation of the Levenshtein distance between the strings "С" and "Т". The table shows the distance values for all subproblems. Green arrows indicate the path of the optimal alignment, starting from the bottom-right cell (2) and moving back to the top-left cell (0). The bottom-right cell (2) is circled in green.

- **Расстояние Дамерау-Левенштейна**

(перестановка соседних символов)

$$d_{DL}(\text{GCAT,CGAT}) = 1$$

- **Метрика Нидлмана-Вунша**

(за операции вставки, удаления, подстановки можно получить разный штраф)

$$\text{delete (c)} = 1 \quad \text{insert (c)} = 2 \quad \text{substitute (c)} = 3$$

- **Метрика Смита-Вотермана**

(штраф за операцию зависит от символа)

$$\text{delete ('A')} = 2$$

$$\text{delete ('B')} = 0.1$$

# Штраф за пропуски

- **Константный штраф**

$$d_C(\text{"gov"}, \text{"government"}) = 3$$

- **Линейный штраф**

$$d_L(\text{"gov"}, \text{"government"}) = 3 * 7 = 21$$

- **Афинный штраф**

$$d_A(\text{"gov"}, \text{"government"}) = 3 + 6 * 2 = 15$$

# Bag distance

(Bartolini, 2002)

# Bag distance metric

s = “bread”

t = “beer”

$M(s) = \{ 'b', 'r', 'e', 'a', 'd' \}$      $M(t) = \{ 'b', 'e', 'e', 'r' \}$

$M(s) \setminus M(t) = \{ 'a', 'd' \}$      $M(t) \setminus M(s) = \{ 'e' \}$

$\text{bag}_{\text{dist}}(s,t) = \max (|\{ 'a', 'd' \}|, |\{ 'e' \}|) = 2$



# Jaro metric

(Winkler, 1999)

$$J(s,t) = \frac{1}{3} * (|s'|/|s| + |t'|/|t| + (|s'| - [T_{s',t'} / 2]) / |s'|)$$

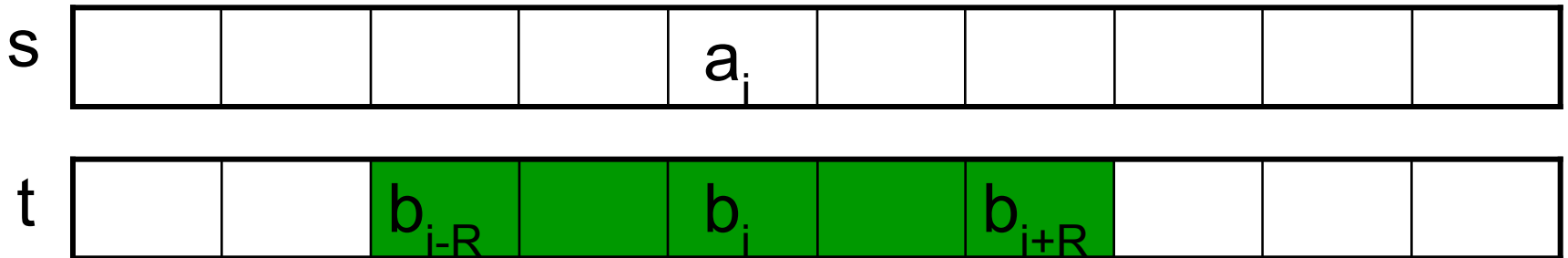
- $s = a_1 a_2 \dots a_k$      $t = b_1 \dots b_L$
- $s'$  и  $t'$  строки общих символов  $s$  и  $t$
- $T_{s',t'}$  количество транспозиций

# Jaro metric

(Winkler, 1999)

Общие символы

$$a_i = b_j$$



$$R = \lceil \max(|s|, |t|) / 2 \rceil - 1$$

# Jaro metric

1.  $s = \text{"CRETA"}$                        $t = \text{"TRACES"}$

2.  $R = [\max(|s|, |t|)/2] - 1 = [\max(5, 6)/2] - 1 = 2$

3.  $s' = \text{"REA"}$                                $t' = \text{"RAE"}$

4.  $T_{s',t'} = 2$

$$\begin{aligned} J(s,t) &= \frac{1}{3} * (|s'|/|s| + |t'|/|t| + (|s'| - [T_{s',t'}/2])/|s'|) \\ &= \frac{1}{3} * (3/5 + 3/6 + (3 - [2/2])/3) = 0.59 \end{aligned}$$

# Jaro-Winkler metric

$$JW(s,t) = J(s,t) + \alpha * \text{boost}(s,t) * (1 - J(s,t))$$
$$\text{boost}(s,t) = \min(|Lcp(s,t)|, p)$$

s = “DIXON”

t = “DICKSONX”

$$J(s,t) = 0.767 \quad \alpha = 0.1 \quad p = 2$$

Lcp(s,t) = “DI”

$$\text{boost}(s,t) = \min(2, 2) = 2$$

$$JW(s,t) = 0.767 + 0.1 * 2 * (1 - 0.767) = 0.813$$

# q-grams metric

(Gravano, 2001)

q-gram – подстрока заданной строки длины q

s = "MARTHA"

q = 2

$G_2(s) = \{ \text{"#M"}, \text{"MA"}, \text{"AR"}, \text{"RT"}, \text{"TH"}, \text{"HA"}, \text{"A#"} \}$

q = 3

$G_3(s) = \{ \text{"##M"}, \text{"#MA"}, \text{"MAR"}, \text{"ART"}, \text{"RTH"}, \text{"THA"}, \text{"HA#"}, \text{"A##"} \}$

# q-grams metric

s = "MARTHA"                      t = "MARCH"

$G_2(s) = \{ \text{"#M"}, \text{"MA"}, \text{"AR"}, \text{"RT"}, \text{"TH"}, \text{"HA"}, \text{"A#"} \}$

$G_2(t) = \{ \text{"#M"}, \text{"MA"}, \text{"AR"}, \text{"RC"}, \text{"CH"}, \text{"H#"} \}$

$q\text{-gram}(s,t) = 3 / \max(7, 6) = 0.43$

# Skip-gram metric

(Keskustalo, 2003)

Skip-gram – “q-грамма”, которая может состоять из несоседних символов

$s = \text{“MARTHA”}$   $q = 2$   $\text{skip}\{0,1\}$

$G_{\text{skip}\{0,1\}}(s) = \{ \text{“\#M”}, \text{“\#A”}, \text{“MA”}, \text{“MR”}, \text{“AR”}, \text{“AT”},$   
 $\text{“RH”}, \text{“TA”}, \text{“RT”}, \text{“TH”}, \text{“HA”}, \text{“A\#”},$   
 $\text{“H\#”} \}$

# Общий префикс(Common Prefix)

$$CP_{\alpha}(s,t) = (|Lcp(s,t)| + \alpha)^2 / (|s| * |t|)$$

s = "MARTHA"      t = "MARCH"

$$Lcp(s,t) = 3 \quad \alpha = 1$$

$$CP_1(s,t) = (3 + 1)^2 / (6 * 5) = 0.53$$



# Наибольшая общая подстрока

$$\text{LCS}(s,t) = \begin{cases} 0, & |\text{Lcs}(s,t)| < k \\ |\text{Lcs}(s,t)| + \text{LCS}(s_{-\text{Lcs}(s,t)}, t_{-\text{Lcs}(s,t)}) \end{cases}$$

$s = \text{"abcdeftg"}$      $t = \text{"bcdaeftg"}$      $k = 2$

•  $s = \text{"abcdeftg"}$      $t = \text{"bcdaeftg"}$

$$\text{LCS}(s,t) = 3 + \text{LCS}(\text{"aeftg"}, \text{"aeftg"})$$

•  $s_{-\text{Lcs}(s,t)} = \text{"aeftg"}$      $t_{-\text{Lcs}(s,t)} = \text{"aeftg"}$

$$\text{LCS}(s,t) = 3 + 3 + \text{LCS}(\text{"tg"}, \text{"g"}) = 6$$

# Weighted LCS

$$w_{\text{Lcs}(s,t)} = \frac{|\text{Lcs}(s,t)| + \alpha - \max(\alpha, p)}{|\text{Lcs}(s,t)| + \alpha}$$

# Monge-Elkan

(Monge and Elkan, 1996)

$$s = \{s_1 s_2 \dots s_K\} \quad t = \{t_1 t_2 \dots t_L\}$$

$$\text{Monge-Elkan}(s,t) = 1/K * \sum_{i=1}^K \max_{j=1..L} \text{sim}(s_i, t_j)$$

$\text{sim}(s_i, t_j)$  – любая метрика для сравнения  
двух строк

# Наборы тестирующих данных

1. Польские имена (1457)

2. Полные польские имена (1219)

# Результаты исследования

Конец доклада

Вопросы?