

# Методы оценки близости строк

Татьяна Кривошеева

# 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                    kill → bill
- Вставка                         kill → skill
- Удаление                        fear → ear

## 1. Расстояние Хэмминга (подстановка)

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

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

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

# Подсчет расстояния Левенштейна

j

	“	T	E	S	T
“					
i	S				
E					
T					

# Подсчет расстояния Левенштейна

0

	“	Т	Е	С	Т
0	“	0			
S					
E					
Т					

# Подсчет расстояния Левенштейна

	“	Т	Е	С	Т
“	0				
S	1				
E					
T					

# Подсчет расстояния Левенштейна

	“	Т	Е	С	Т
“	0				
S	1				
E	2				
T					

# Подсчет расстояния Левенштейна

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

# Подсчет расстояния Левенштейна

	“	Т	Е	S	Т
“	0	1	2	3	4
S	1				
E	2				
T	3				

# Подсчет расстояния Левенштейна

	“	Т	Е	S	Т
“	0	1	2	3	4
S	1	1			
E	2				
Т	3				

The diagram illustrates the Levenshtein distance matrix for the strings "СЕМЯ" and "СЫРЬЯ". The matrix is a 4x6 grid where rows represent the first string ("СЕМЯ") and columns represent the second string ("СЫРЬЯ"). The diagonal elements are black, while off-diagonal elements are red. A green arrow points from the top-left cell to the cell at index (1,1), and a red arrow points from the cell at index (1,1) to the cell at index (2,1).

	“	Т	Е	S	Т
“	0	1	2	3	4
S	1	1			
E	2				
Т	3				

# Подсчет расстояния Левенштейна

	“	Т	Е	S	Т
“	0	1	2	3	4
S	1	1	2	2	3
E	2	2	1	2	3
Т	3	2	2	2	2

The diagram shows a 5x6 grid representing the Levenshtein distance between the strings "TEXT" (rows) and "TEST" (columns). The grid contains the following values:

	“	Т	Е	S	Т
“	0	1	2	3	4
S	1	1	2	2	3
E	2	2	1	2	3
Т	3	2	2	2	2

Green arrows highlight specific transitions: one from the first 'T' in "TEST" to the first 'T' in "TEXT", another from the second 'E' in "TEST" to the second 'E' in "TEXT", and a third from the 'S' in "TEST" to the second 'T' in "TEXT". A green oval surrounds the value '2' in the bottom-right cell, which is circled.

- **Расстояние Дамерау-Левенштейна**  
(перестановка соседних символов)

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

- **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**

$$d_C("gov", "gover\textcolor{red}{n}ment") = 3$$

- **Линейный штраф**

$$d_L("gov", "gover\textcolor{red}{n}ment") = 3 * 7 = 21$$

- **Афинный штраф**

$$d_A("gov", "gover\textcolor{blue}{n}ment") = 3 + 6 * 2 = 15$$

# Bag distance

(Bartolini, 2002)

# Bag distance metric

$s = \text{"bread"}$

$t = \text{"beer"}$

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

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

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

# Jaro metric (Winkler, 1999)

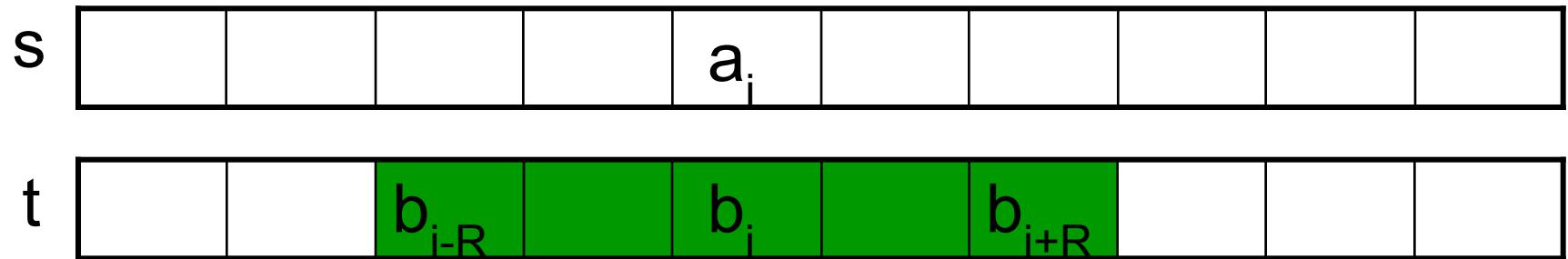
$$J(s,t) = \frac{1}{3} * (|s'|/|s| + |t'|/|t| + (|s'| - [T_{s',t'} / 2]) / |s'|)$$

- $s = a_1a_2\dots a_k$      $t = b_1\dots b_L$
- $s'$  и  $t'$  строки общих символов  $s$  и  $t$
- $T_{s',t'}$  количество транспозиций

# Jaro metric (Winkler, 1999)

Общие символы

$$a_i = b_j$$



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

# 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

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

$s = \text{"DIXON"}$        $t = \text{"DICKSONX"}$

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

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

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

# 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

$s = \text{"MARTHA"}$        $t = \text{"MARCH"}$

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

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

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

# Skip-gram metric

(Keskustalo, 2003)

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

$s = \text{"MARTHA"} \quad q = 2 \quad \text{skip}\{0,1\}$

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

# Общий префикс(Common Prefix)

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

$s = "MARTHA"$        $t = "MARCH"$

$Lcp(s,t) = 3$        $\alpha = 1$

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

# Наибольшая общая подстрока

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

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

- $s = \text{"ab}\color{red}{cdeftg"} \quad t = \text{"bc}\color{red}{daefg"}$   
 $\text{LCS}(s,t) = 3 + \text{LCS}(\text{"aeftg"}, \text{"aefg"})$
- $s_{-\text{Lcs}(s,t)} = \text{"ae}\color{blue}{ftg"} \quad t_{-\text{Lcs}(s,t)} = \text{"a}\color{blue}{efg"}$   
 $\text{LCS}(s,t) = 3 + 3 + \text{LCS}(\text{"tg"}, \text{"g"}) = 6$

# Weighted LCS

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

# 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?