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)

$d_H(GCAT,CGAT) = 2$

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

$d_E(CGACG, GTCGA) = 3$

Подсчет **расстояния Левенштейна**

i \ j		“	T	E	S	T
	“					
	S					
	E					
	T					

Подсчет **расстояния Левенштейна**

		0			
	“	T	E	S	T
0 “	0				
S					
E					
T					

Подсчет **расстояния Левенштейна**

	“	T	E	S	T
“	0				
S	1				
E					
T					

Подсчет **расстояния Левенштейна**

	“	T	E	S	T
“	0				
S	1				
E	2				
T					

Подсчет **расстояния Левенштейна**

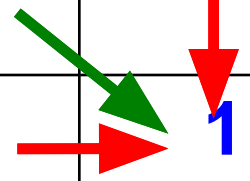
	“	T	E	S	T
“	0				
S	1				
E	2				
T	3				

Подсчет **расстояния Левенштейна**

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

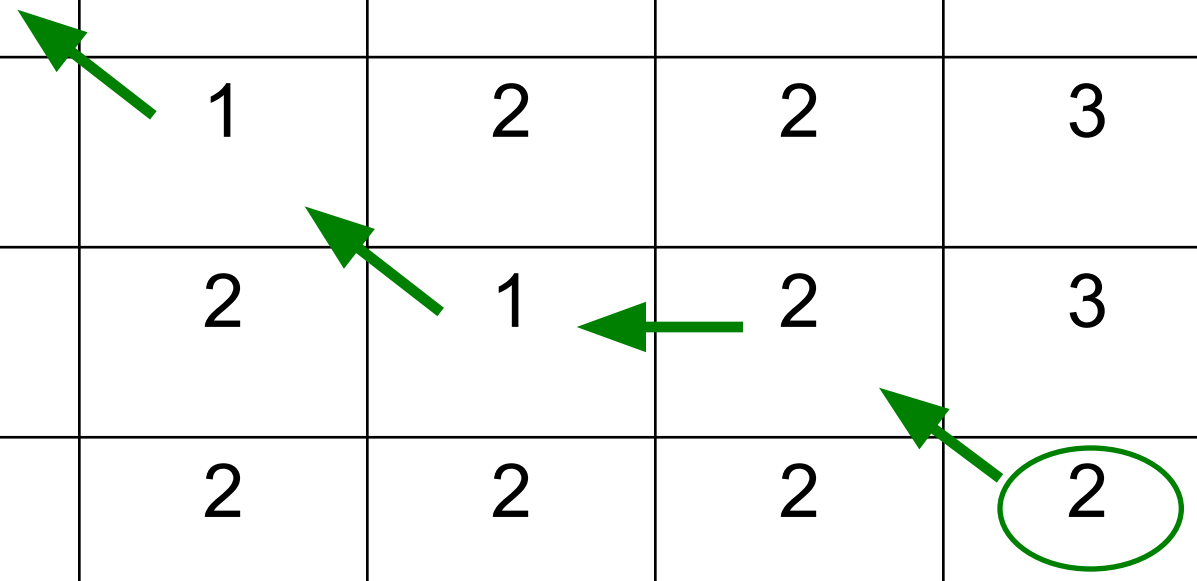
Подсчет **расстояния Левенштейна**

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



Подсчет **расстояния Левенштейна**

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



- **Расстояние Дамерау-Левенштейна**

  (перестановка соседних символов)

  d_{DL}(GCAT,CGAT) = 1

- **Метрика Нидлмана-Вунша**

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

  delete (c) = 1 insert (c) = 2 substitute (c) = 3

- **Метрика Смита-Вотермана**

  (штраф за операцию зависит от символа)

  delete (‘A’) = 2 delete (‘B’) = 0.1

Штраф за пропуски

- **Константный штраф**

  d_C(“gov”, “government”) = 3

- **Линейный штраф**

  d_L(“gov”, “government”) = 3 * 7 = 21

- **Афинный штраф**

  d_A(“gov”, “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) \ M(t) = { ‘a’, ‘d’ } M(t) \ M(s) = { ‘e’ }

$bag_{dist}(s,t) = max(|\{ ‘a’, ‘d’ \}|, |\{ ‘e’ \}|) = 2$

Jaro metric
(Winkler, 1999)

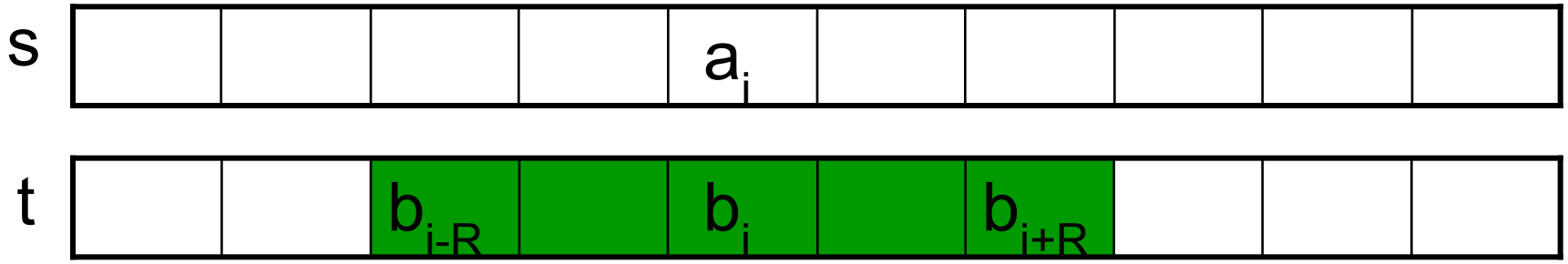
$$J(s,t) = \frac{1}{3}*(|s'|/|s| + |t'|/|t| + (|s'| - [T_{s',t'}/2])/|s'|)$$

- $s = a_1 a_2 \ldots a_k \quad t = b_1 \ldots b_L$
- 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 = “CRETA” t = “TRACES”

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

3. s’ = “REA” t’ = “RAE”

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

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

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”

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

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
$G_2(s)$ = { “#M”,“MA”, “AR”, “RT”, “TH”, “HA”, “A#”}

q = 3
$G_3(s)$ = { “##M”,“#MA”,“MAR”, “ART”, “RTH”,
“THA”,“HA#”,“A##”}

q-grams metric

s = “MARTHA” t = “MARCH”

$G_2(s)$ = { “#M”,“MA”, “AR”, “RT”, “TH”, “HA”, “A#”}

$G_2(t)$ = { “#M”,“MA”, “AR”, “RC”, “CH”, “H#”}

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 = “MARTHA” q = 2 skip{0,1}

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

Общий префикс(Common Prefix)

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

s = “MARTHA” t = “MARCH”

Lcp(s,t) = 3 α = 1

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

Наибольшая общая подстрока

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

s = “abcdeftg” t = “bcdaefg” k = 2

- s = “abcdeftg” t = “bcdaefg”
  LCS(s,t) = 3 + LCS(“aeftg”, “aefg”)
- $s_{-Lcs(s,t)}$ = “aeftg” $t_{-Lcs(s,t)}$ = “aefg”
  LCS(s,t)= 3 + 3 + LCS(“tg”, “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)

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

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

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

Наборы тестирующих данных

1. Польские имена (1457)

2. Полные польские имена (1219)

Результаты исследования

Конец доклада

Вопросы?