Редакционное расстояние
Расстояние Левенштейна

Задача о выравнивании

SUNNY SNOWY

1. SUNNY
   SNOWY 3

   ДОЖДЬ
   ДРОЖЬ 3

2. SUN-NY
   S-NOWY 3

   Д-ОЖДЬ
   ДРОЖ-Ь 2

Задача о редакционном расстоянии

Три типа удалений:

1. замена $\frac{a}{в}$
2. удаление $\frac{a}{-}$
3. вставка $\frac{-}{в}$

Вопрос: сколько нужно редакций для преобразования одного слова в другое?

S_1 = POLYNOMIAL

S_2 = EXPONENTIAL

$E(S_1, S_2) = ?$

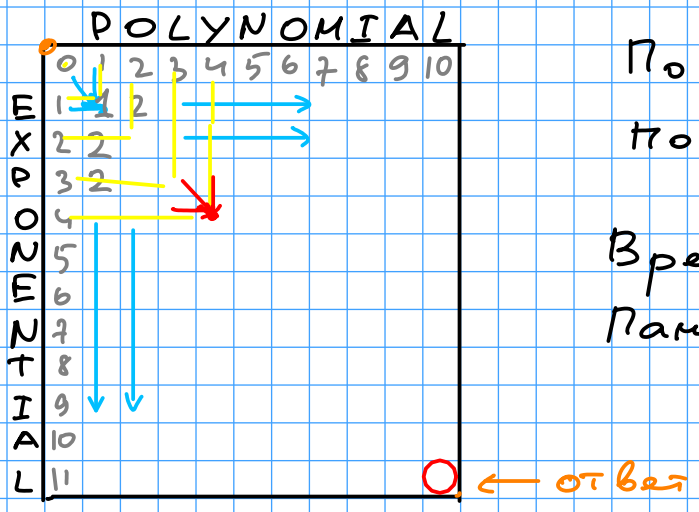
Динамическое программирование:

1. $E[i, j]$ – выравнивание $S_1[1, i]$ и $S_2[1, j]$
2. $E[0, j] = j$
   $E[i, 0] = i$

$$E[i, j] = \min \begin{cases} E[i-1, j-1] + [S_1[i] \neq S_2[j]] & \text{замена} \\ E[i, j-1] + 1 & \text{вставку} \\ E[i-1, j] + 1 & \text{удаление} \end{cases}$$

стоимость операций

3. Порядок вычисления



По строкам / по столбцам

Время работы: $O(|S_1| \cdot |S_2|)$

Память: $O(|S_1| \cdot |S_2|)$

Можно восстановить путь, запомнив min ребра.

Пусть $|S_1| = |S_2| = 10^9$

Время: $|S_1| \cdot |S_2| = 10^{18}$

Умеем: $3 \cdot 10^5$ операций в секунду

Дней: $3600 \cdot 24 = 10^5$

Потребуется $\sim 3 \cdot 10^3$ дней.

Память: проблема.

Можно хранить только две строки $\Rightarrow$

Память $O(|S_1| + |S_2|)$

Как восстановить редакц. последовательность?

Алгоритм Хиршберга

1. $E(S_1, S_2) = E(rev(S_1), rev(S_2))$
2. $\forall\, S_{11} \circ S_{12} = S_1 \Rightarrow \exists\, S_{21} \circ S_{22}$:
   $E(S_{11}, S_{21}) + E(S_{12}, S_{22}) = E(S_1, S_2)$

S_1 = POLYN|OMIAL (S_{11} | S_{12})

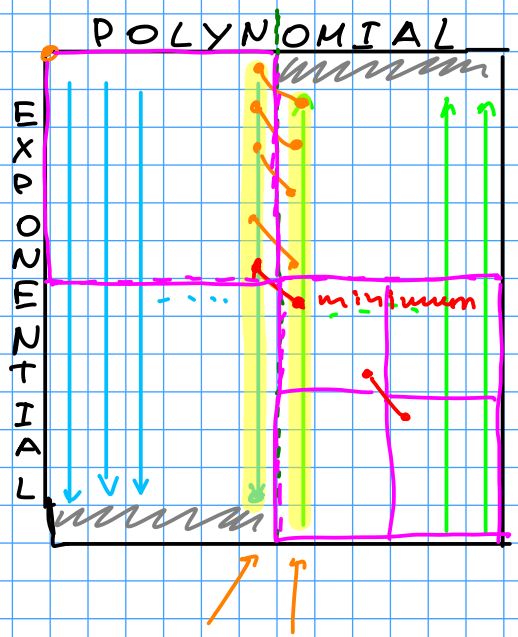
S_2 = EXPO|NENTIAL (S_{21} | S_{22})

Д	-	О	*	Д	б
Д	р	О	*	-	б

2

S_{11} = ДО* S_{12} = Дб

S_{21} = ДРО* S_{22} = б



$E(S_{11}, S_2[0, j])$

$E(rev(S_{12}), rev(S_2[j, n]))$

Первая итерация:

1. Разбиваем $S_1 = S_{11} \circ S_{12}$
   $|S_{11}| = |S_{12}| = |S_1| / 2$
2. Вычисляем:
   - $E(S_{11}, S_2)$
   - $E(rev(S_{12}), rev(S_2))$

Повторяем рекурсивно для левого верхнего и правого нижнего прямоугольников.

Время: $|S_1| \cdot |S_2| + \frac{|S_1| \cdot |S_2|}{2} + \frac{|S_1| \cdot |S_2|}{4} \dots =$

$= 2|S_1| \cdot |S_2| = O(|S_1| \cdot |S_2|)$

Память: $O(|S_1| + |S_2|)$

NB: Meet-in-the-middle