Спелл-чекинг и редакционное расстояние

Занятие 4

Спелл-чекинг

Симпатичный	вряд ли	сегодня
Симпотичный	вряд-ли	седня

Спелл-чекинг

Симпатичный	вряд ли	сегодня
Симпотичный	вряд-ли	се* *дня

Нам, людям, просто исправлять эти ошибки. Компьютеру сложно. Какой алгоритм сделать, чтобы научить его?

Расстояние Левенштейна

Самое популярное редакционное расстояние

Метрика, позволяющая определить «**схожесть**» двух строк — минимальное количество операций

- **вставки** одного символа
- **удаления** одного символа
- **замены** одного символа на другой,

необходимых для превращения одной строки в другую.

Расстояние Левенштейна

Самое популярное редакционное расстояние

Метрика, позволяющая определить «**схожесть**» двух строк — минимальное количество операций

- **вставки** одного символа
- **удаления** одного символа
- **замены** одного символа на другой,

необходимых для превращения одной строки в другую.

Чтобы узнать редакционное расстояние между двумя строками, нужно **посчитать минимальное количество операций**, которые нужно сделать, чтобы превратить первую строку во вторую.

Операции

Расстояние Левенштейна

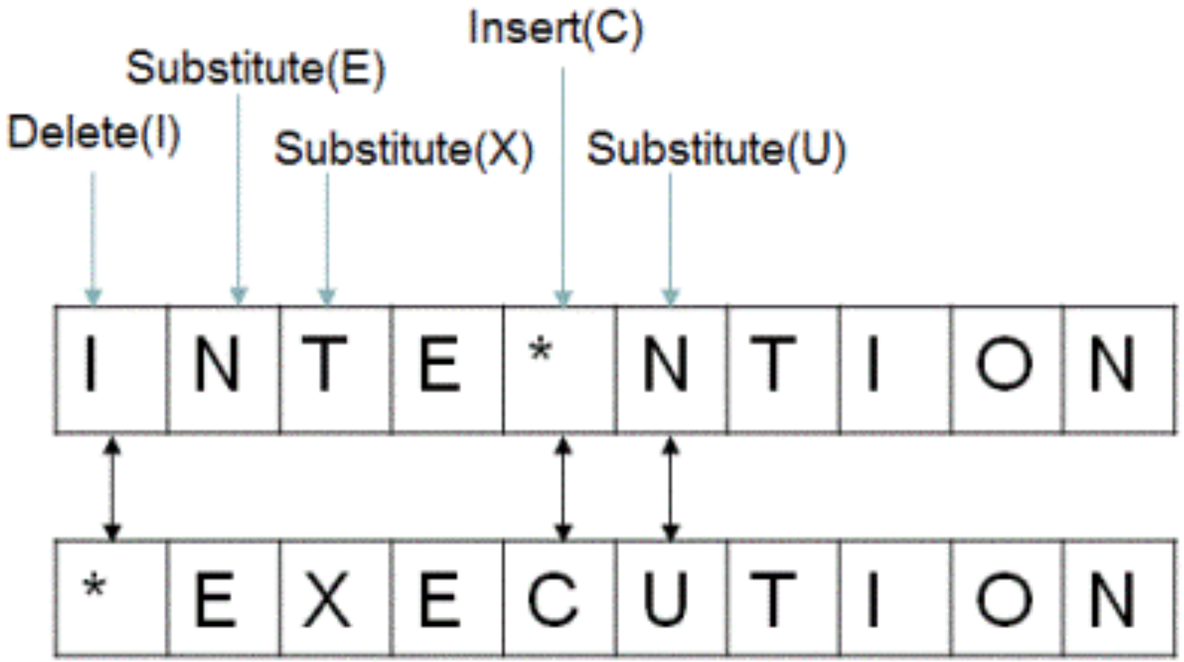
Вставка	сверсник -> сверстник
Замена	черезчур -> чересчур
Удаление	мне нравиться -> мне нравится

Расстояние Левенштейна

Самое популярное редакционное расстояние

- Каждый раз в операции может участвовать один символ в строке (**посимвольная** операция)
- Каждая **вставка**, **удаление** и **замена** стоят 1 балл
- (Иногда вставка и удаление = 1 балл, замена = 2 балла)
- В конце баллы суммируются - вуаля, количественная мера готова!

Пример подсчета



1 удаление + 3 замены + 1 вставка = 5 баллов

или 1 удаление + 3 замены (* 2) + 1 вставка = 8 баллов

Выравнивание

alignment

- Посимвольное выравнивание в задачах спелл-чекинга

В параллельных корпусах:

- выравнивание по словам
- выравнивание по предложениям

Нужно для **машинного перевода**

Выравнивание

alignment

I guess the Feds didn't do such a good job on the protection part.	Полагаю, федералы не очень справились с охраной.

Проблемы пословного выравнивания

No! No, you leave your cash safe and sound where it is.	Пусть ваши деньги останутся в целости и сохранности.

Проблемы выравнивания по предложениям

Выравнивание

	#	H	O	C
#				
O				
C				

Выравнивание

было / стало	#	н	о	с
#	0	1	2	3
о	1	2	1	2
с	2	3	2	1

Выравнивание

было / стало	#	н	о	с
#	0	1	2	3
о	1	2	1	2
с	2	3	2	1

Выравнивание

было / стало	#	н	о	с
#	0	1	2	3
о	1	2	1	2
с	2	3	2	1

Выравнивание

было / стало	#	н	о	с
#	0	1	2	3
о	1	2	1	2
с	2	3	2	1

Редакционное расстояние

Реальный пример

Пользователь: Новосиб

Словарь:

- Новосиль
- Новосибирск

Задача: Посчитайте кол-во баллов для обоих вариантов. Что выберет алгоритм?

Редакционное расстояние

Реальный пример

Пользователь: Новосиб

Словарь:

- Новосиль — — — — 1 замена, 1 вставка. Р. Левенштейна = 3
- Новосибирск — — 4 вставки. Р. Левенштейна = 4

К сожалению, победит Новосиль

Редакционное расстояние

Алгоритм

Нам дан текст. Как исправить ошибки?

1. Проверить все слова по словарю (*предварительно разбив на токены и удалив пунктуацию*). Не найденные в словаре слова записать в список ошибочных.

2. Найти кандидатов на роль ошибочных слов. Посчитаем, сколько их может быть…

Количество возможных кандидатов

Множество всех вариантов посимвольных удалений, вставок и замен.

Удаление

Сколько вариантов удаления возможно в слове «интригант»?

Удаление

Сколько вариантов удаления возможно в слове «интригант»?

9

Удаление

Удаление = n вариантов, где n - длина слова

```
 нтригант
и тригант
ин ригант
инт игант
интр гант
интри ант
интриг нт
интрига т
интриган
```

```
1 len('интригант')
```

```
9
```

Замена

Сколько замен возможно в слове, в котором 3 буквы?

(Например, слово «чир»)

Замена

Сколько замен возможно в слове, в котором 3 буквы?

(Например, слово «чир»)

96

Замена

- Каждую букву можем заменить на любую другую букву алфавита
- 1 буква заменяется на другие 32
- Всего 32 * n операций

```
word = 'чир'
```

```
с ир
м ир
и ир
т ир
ь ир
б ир
ю ир
ё ир
```

Вставка

word = ‘сонце’, len = 5

вставить буквы можно на _ позиций

Вставка

word = ‘сонце’, len = 5

вставить буквы можно на 6 позиций (n + 1):

1. _сонце
2. с_онце
3. со_нце
4. сон_це
5. сонц_е
6. сонце_

Вставка

Сколько вставок возможно в слове «сонце»?

Вставка

Сколько вставок возможно в слове «сонце»?

198

Вставка

word = ‘сонце’, len = 5

- Вставить буквы можно на 6 позиций (n + 1)
- Вставить можно любую букву алфавита (33)
- Всего возможных операций (n + 1) * 33

Количество возможных кандидатов

Множество всех вариантов посимвольных удалений, вставок и замен. len(word) = n

Удаление: n вариантов

Замена: n * 32 (в алфавите 33 буквы, одна из них - это наш символ)

Вставка: (n + 1) * 33

Количество возможных кандидатов

Множество всех вариантов посимвольных удалений, вставок и замен. len(word) = n

Удаление: n вариантов

Замена: n * 32 (в алфавите 33 буквы, одна из них - это наш символ)

Вставка: (n + 1) * 33

Итого: n + 32n + 33(n + 1) = 33n + 33n + 33 = 66n + 33

Количество возможных кандидатов

Итого: $n + 32n + 33(n + 1) = 33n + 33n + 33 = 66n + 33$

Посчитайте, сколько возможно кандидатов для слова из 3 букв? Из 5?

Количество возможных кандидатов

Итого: n + 32n + 33(n + 1) = 33n + 33n + 33 = **66n + 33**

Посчитайте, сколько возможно кандидатов для слова из 3 букв?

231

Из 5?

363

Быстро будет работать такой алгоритм, если проверять на орфографию большой текст?

Редакционное расстояние

Алгоритм

Нам дан текст

1. Проверить все слова по словарю (предварительно разбив на токены и удалив пунктуацию). Не найденные в словаре слова записать в список ошибочных.

2. Найти кандидатов на роль ошибочных.

3. Оставляем только словарных кандидатов

4. Как выбрать лучшего? Самый частотный в нашем корпусе? Учитывать контекст (N-граммы)?

Виды редакционных расстояний

Редакционное расстояние - общий термин

Включает несколько метрик

1. Расстояние Левенштейна (замена, удаление, вставка)
2. Расстояние Дамерау-Левенштейна (замена, удаление, вставка, **перестановка соседних**) - покрывает 80% чел. ошибок
3. Расстояние Хэмминга (замена)
4. ...