Поиск в тексте

- Допустим, нам надо найти в тексте слово **Händel**. С умляутами в текстах бывают проблемы, и оно может быть написано так:

- **Händel**
- **Handel**
- **Haendel**

- Как справиться с этой проблемой?

Поиск в тексте

Варианты решения:

- Искать 3 раза.
- Искать, предварительно заменив все *ä* и *ae* на *a*.

- Всё это слишком сложно для такой простой задачи.

Регулярные выражения

- Регулярное выражение — это выражение на специальном языке, позволяющее искать нужные фрагменты текста.

H(ä|ae?)ndel

- | (vertical bar) означает «или» (дизъюнкция, disjunction)
- ? (question mark) означает «предыдущего символа может и не быть»
- скобки (parentheses) используются для группировки

Регулярные выражения

- Регулярные выражения можно использовать во многих языках программирования, в текстовых редакторах (Notepad++), в специальных программах для поиска файлов и т. п.

- Языки регулярных выражений немного различаются, но имеют общую основу

- В Питоне функции для работы с ними находятся в модуле `re`

Регулярные выражения

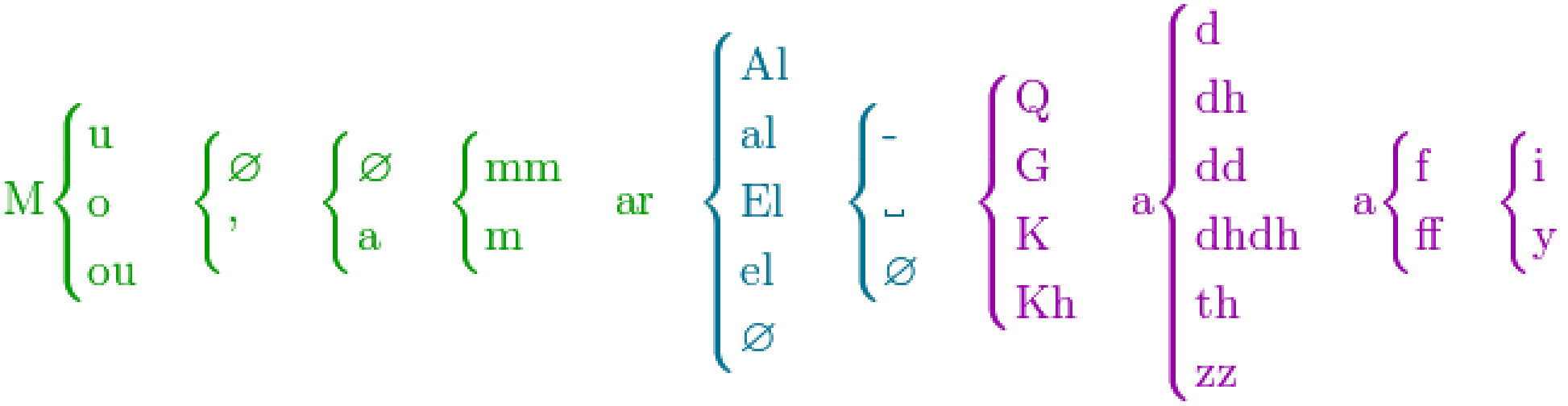
- Могут использоваться для поиска или замены фрагментов текста

- Часть строки, подходящая под регулярное выражение, называется совпадением (match)

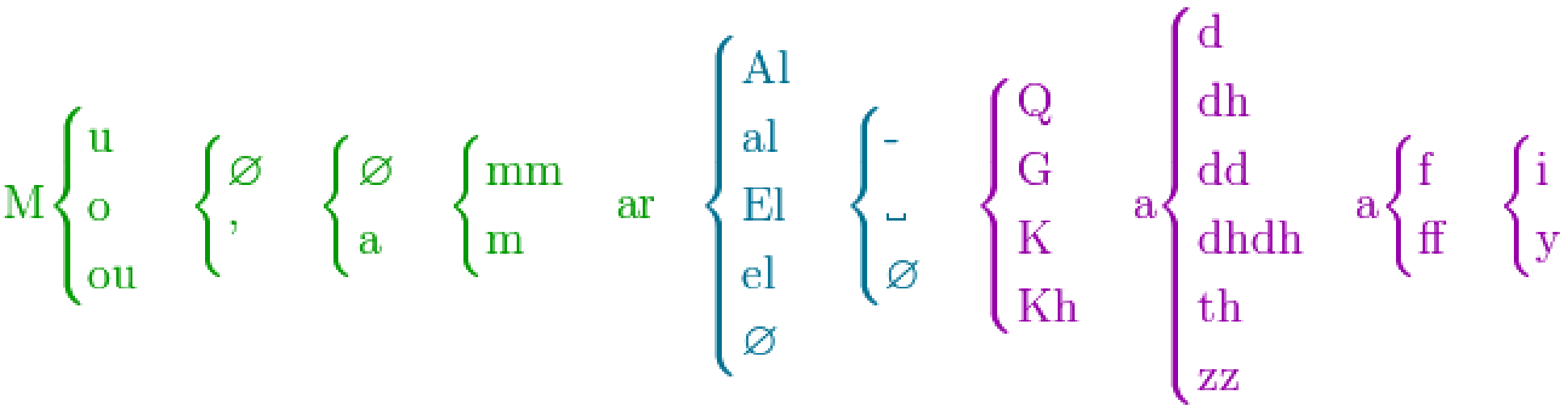
- Возможности регулярных выражений широки, но не безграничны

Пример посложнее

over 9000 способов написать «Муаммар Каддафи» по-английски



«Муаммар Каддафи»



M(u|ou?)'?a?mm?ar

((A|a|E|e)l)?-?

(Q|G|Kh?)a(d(d|h(dh)?)|th|zz)aff?(i|y)

Язык регулярных выражений

- Часть р. в., ограниченная скобками, называется группой (group)
- (...|...|...) — перечисление вариантов в группе
- ? — возможное отсутствие предыдущего символа или группы
- . — один любой символ

Язык регулярных выражений

- * — предыдущий символ или группа, повторённые любое количество раз (включая 0)
- + — предыдущий символ или группа, повторённые любое положительное количество раз
- соответственно, .* — любое количество любых символов

Пример

ко

- Кошка, или домашняя кошка (лат. Felis silvestris catus) — домашнее животное, одно из наиболее популярных (наряду с собакой) «животных-компаньонов»

Пример

и(б|в)?ол?

- Кошка, или домашняя кошка (лат. Felis silvestris catus) — домашнее животное, одно из наиболее популярных (наряду с собакой) «животных-компаньонов»

Пример

до.*е+

- Кошка, или домашняя кошка (лат. Felis silvestris catus) — домашнее животное, одно из наиболее популярных (наряду с собакой) «животных-компаньонов»

Язык регулярных выражений

- [...] — один из перечисленных символов (например: [абв])
- [...-...] — один символ из диапазона (например: [а-я])
- можно комбинировать: [а-яА-Яbq]
- экранирование метасимволов: \[, \], \(, \), \., * и т. п.

Пример

- Выражение, находящее все формы слова *have*:

```
ha(s|d|v(e|ing))
```

- Выражение, дающее совпадение на словах, имеющих хотя бы два слога:

```
.*[аеёиоуиыэюя].*[аеёиоуиыэюя].*
```

Пример

- Выражение, находящее электронные адреса:

```
[a-z0-9_\.-]+@[a-z0-9\.-]+\.[a-z][a-z][a-z]?[a-z]?
```

- Выражение, находящее даты в формате число.месяц.год:

```
[0-3]?[0-9]+\.[01]?[0-9]\.[1-9][0-9][0-9][0-9]
```

  - На самом деле, тут есть неточности: допускается дата 0, 32 и т. п.

Рег. выражения и Питон

- Модуль **`re`**
- Для проверки, совпадает ли строка **`s`** с регулярным выражением **`regex`**: функция **`re.search`**

```
m = re.search(regex, s)
if m != None:
    ...
```