# REGEX + scrapinng

«Высшая школа экономики»

24.9

# Outline

re templates

scraping basics

Регулярные выражения - язык описания шаблонов (patterns) для извлечения информации из текста.

Pure rule-based approach.

- Поиск точного совпадения.
- Поиск шаблонного совпадения.
- Возможность введения переменных.
- Жадность.
- Метасимволы.

Точное совпадение:

Строка на вход «aaa bbb ccc»

шаблон: r"a"

на выходе: [а, а, а]

```
□ - СИМВОЛ «ИЛИ»
```

Строка на вход «aaa bbb ccc»

шаблон: r"[ab]"

на выходе: [a, a, a, b, b, b]

```
() - группировка
```

Строка на вход «aaa bbb ccc»

шаблон: r"(aaa) bbb"

на выходе: [ааа]

```
Если группа не нужна, то в группе ставим :?
```

```
Строка на вход «aaa bbb ccc»
```

```
шаблон: r"(aaa) (?:bbb)"
```

```
на выходе: [ааа]
```

#### Метасимволы

- \d все цифры [0-9]
- w все буквы [а-Zа-Я]
- \s пробельные символы
- \b слова

#### Метасимволы

- \D Не цифры [0-9]
- W Не буквы [а̂-Zа-Я]
- \S Не пробельные символы
- \В Не слова

#### Квантификаторы

- {n} n повторений
- {n,m} минимум n, максимум m повторений
- {n,} минимум n повторений
- {,m} максимум m повторений

Пример кватификатора

Строка на вход «На дворе - трава, на траве - дрова.»

на выходе: ['дворе', 'трава', 'траве', 'дрова']

## Синонимы кватификаторов

Синоним	Расшифровка	Квантификатор
+	1 и более раз	{1,}
*	0 и более раз	{0,}
?	0 или 1 раз	{0,1}

Жадность - по умолчанию регулярные выражения захватывают максимум символов, которые помещаются под шаблон.

Строка на вход «( dfghvb ) sdvsd ( sdcvkjnh ) sdvsd ( dkjhvgr ) sdvfv.»

шаблон:  $r'' \setminus (.+ \setminus)''$ 

на выходе: ['dfghvb ) sdvsd ( sdcvkjnh ) sdvsd ( dkjhvgr']

```
Чтобы убрать жаность необходимо добавить ?
```

Строка на вход «( dfghvb ) sdvsd ( sdcvkjnh ) sdvsd ( dkjhvgr ) sdvfv.»

шаблон: 
$$r" \setminus (.+? \setminus)"$$

на выходе: ['dfghvb', 'sdcvkjnh', 'dkjhvgr']

По умолчанию перенос строки  $(\n)$  явялется концом поиска регулярного выражения.

(?s) в начале строки шаблона включает перенос строк.

Скрапинг - процесс сбора информации с веб страниц.

Парсинг - процесс обработки текста, часто подразумевается разбор текста на составные части.

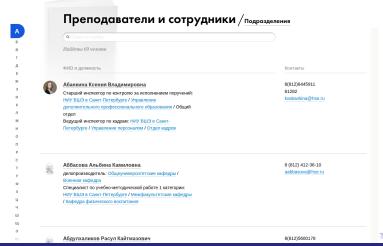
- requests + bs4 низкоуровневый поиск.
- selenium имитация клиента.
- scrapy готовое решение для скрапинга, основано на ассинхронных запросах.

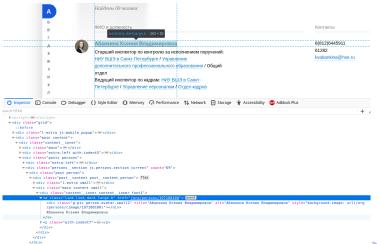
Из чего состоит веб страница

- html
- CSS
- js

Возможные проблемы:

- Скорость...
- Тайминги, фризы.
- АЈАХ страницы.





▶ <div class="post person"> = </div>