Простые приемы работы с текстовыми данными:
применение метода регулярных выражений

Карасёва Екатерина

Коваленко Кира

Что это такое?

- ***Регулярное выражение*** *— это обозначение критериев, которым должен соответствовать искомый текст. С помощью регулярных выражений можно найти текстовый фрагмент - один или более, - который соответствует заданным условиям, а также произвести замену.*

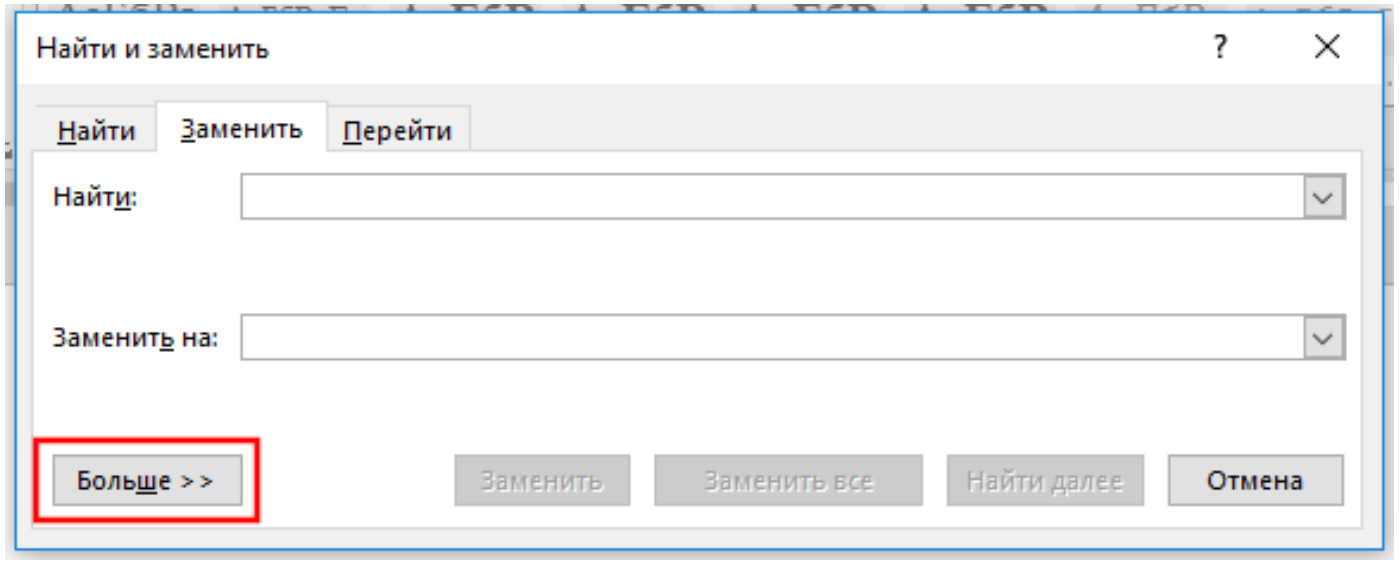
Где это используется?

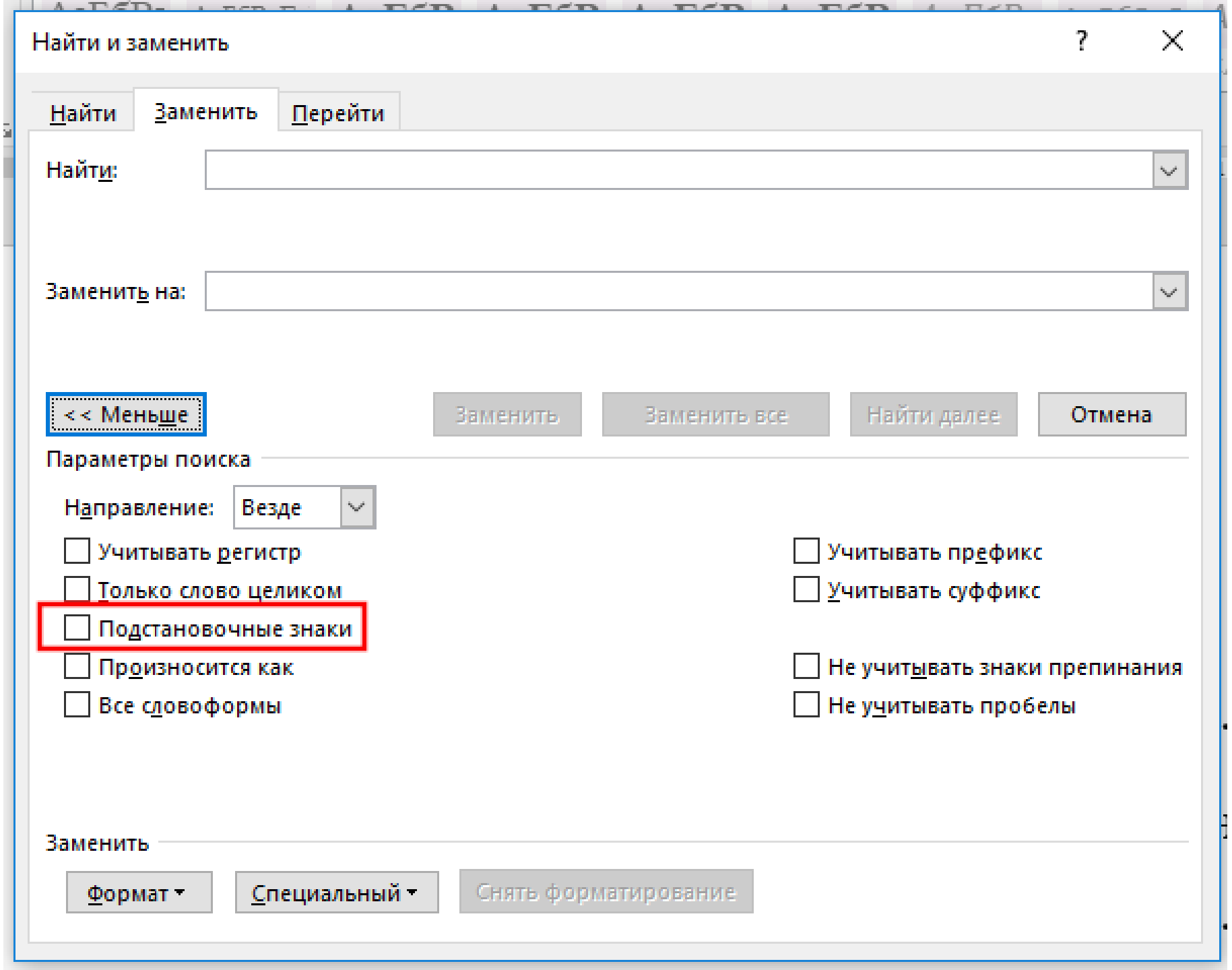
- Текстовые редакторы: **MS Word**
- Редакторы таблиц: **Google Sheets**
- Базы данных: **MySQL, PostgreSQL, Oracle**
- Языки программирования: **Python, R, Java, JavaScript, Perl, PHP, XQuery, XPath…**
- Специализированные инструменты: **grep, PowerGREP…**

Сегодня:

- Текстовый редактор: **MS Word**
  - ✓ поиск и замена
  - ✓ форматирование
- Редактор таблиц: Таблицы **Google**
  - ✓ поиск и замена
  - ✓ реструктурирование данных

Microsoft Word




Найти и заменить
Найти
Заменить
Перейти
Найти:
Заменить на:
<< Меньше
Заменить
Заменить все
Найти далее
Отмена
Параметры поиска
Направление:
Везде
Учитывать регистр
Только слово целиком
Подстановочные знаки
Произносится как
Все словоформы
Учитывать префикс
Учитывать суффикс
Не учитывать знаки препинания
Не учитывать пробелы
Заменить
Формат
Специальный
Снять форматирование

Буквы и цифры

- [а-я]
- [А-Я]
- [А-я]
- [a-z]
- [A-Z]
- [A-z]
- [A-zА-я]
- [0-9]

Буквы и их количество

- <[А-я]> - все слова на кириллице
- <[А-я]{5}> - слова из пяти букв
- <[А-я]{5;}> - слова из пяти букв и больше
- <[А-я]{1;}>. – слова, после которых точка
- <[А-я]{1;}>[.,] – или запятая

Замена символов

- {2;} два и более пробелов, заменить на 1
- ^013{2;} на ^013 - два и более символов абзацев, заменить на 1
- ([0-9])-([0-9]) на \1^=\2 – заменить короткое тире на длинное
- ([0-9]{1;})([А-я]) \1 \2 - вставить пробелы между цифрами
- ([А-Я]{1}.)([А-Я]{1}.) \1 \2 – вставить пробелы между инициалами
- ([А-Я]{1}.)([А-Я]{1}.) , \1 \2 – запятая перед инициалами

Изменение форматирования

- Убрать лишние пробелы
- Изменить ссылки на литературу:

  [Rosén 1957, p. 214] → [Rosén 1957: 214]

- Убрать запятые после фамилии автора в библиографии

  Rosén, H. B. → Rosén H. B.

- Заменить дефис на короткое тире между номерами страниц:

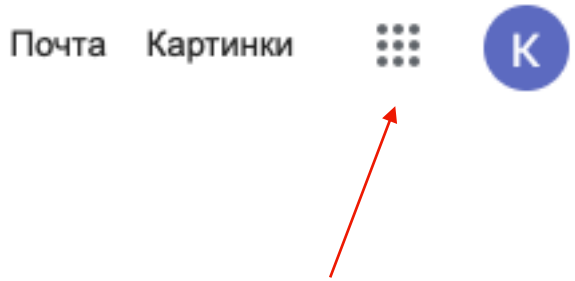
  12-15 → 12–15

regexone.com

abc...	*Letters*
123...	*Digits*
\d	*Any Digit*
\D	*Any Non-digit character*
.	*Any Character*
\.	*Period*
[abc]	*Only a, b, or c*
[^abc]	*Not a, b, nor c*
[a-z]	*Characters a to z*
[0-9]	*Numbers 0 to 9*
\w	*Any Alphanumeric character*
\W	*Any Non-alphanumeric character*
{m}	*m Repetitions*
{m,n}	*m to n Repetitions*
*	*Zero or more repetitions*
+	*One or more repetitions*
?	*Optional character*
\s	*Any Whitespace*
\S	*Any Non-whitespace character*
^...$	*Starts and ends*
(...)	*Capture Group*
(a(bc))	*Capture Sub-group*
(.*)	*Capture all*
(abc|def)	*Matches abc or def*

Google Sheets

- Зайти в свой аккаунт **Google**



- Найти приложение «Таблицы» (**Google.Sheets**)

Полезные ссылки

- **Общая информация и обучающие материалы**
  - https://www.regular-expressions.info
  - https://regexone.com
  - https://regexr.com
- **Регулярные выражения в MS Word**
  - https://www.customguide.com/word/how-to-use-wildcards-in-word
  - https://www.youtube.com/watch?v=xeP9yyg6lF4&t=46s
- **Регулярные выражение в Google Sheets**
  - https://www.youtube.com/watch?v=w5B43VsJqOs