Краткое введение
Регулярные выражения

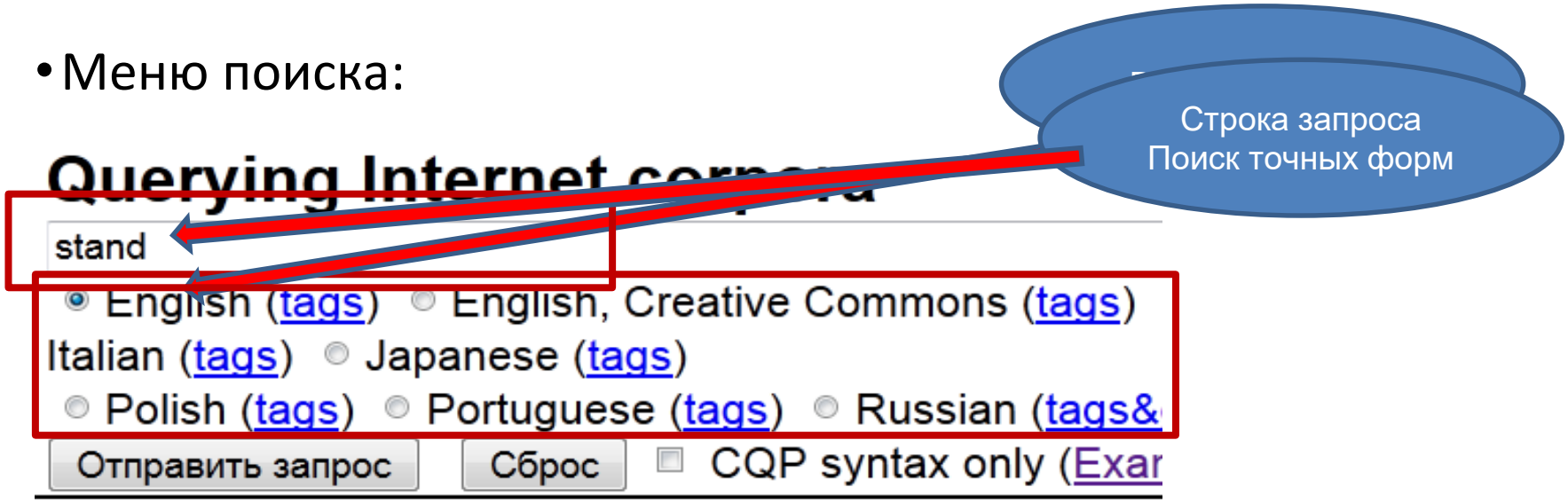
Помощь

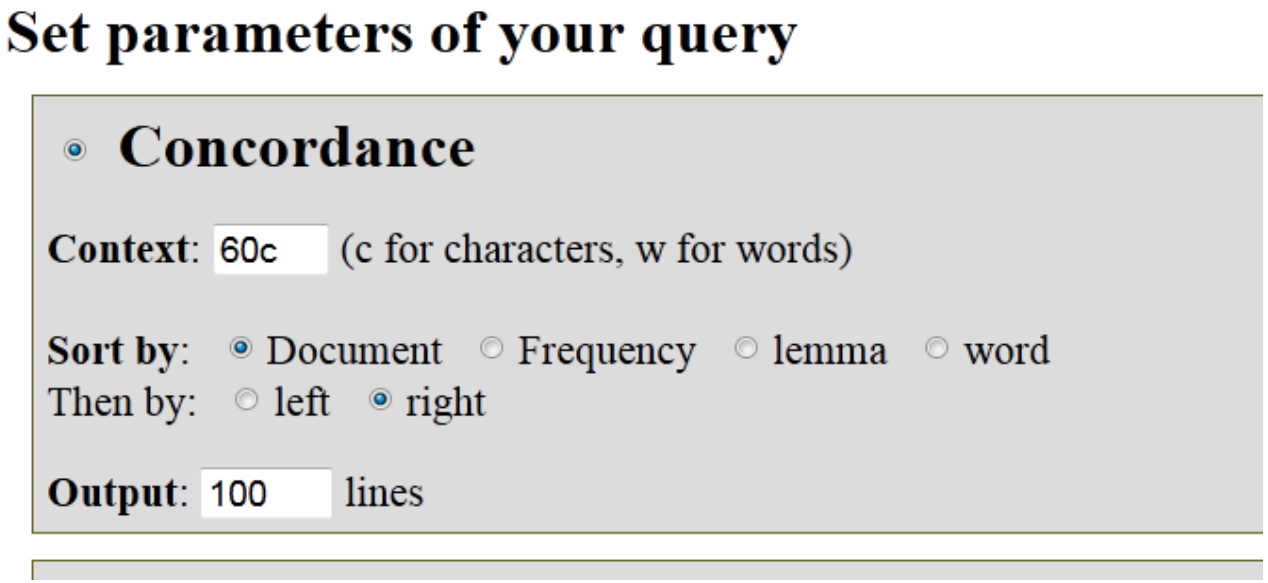
- http://www.regular-expressions.info/
- http://corp.hum.sdu.dk/cqp_help.html
- http://cwb.sourceforge.net/documentation.php
- http://ru.wikibooks.org/wiki/%D0%A0%D0%B5%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D1%8B%D0%B5_%D0%B2%D1%8B%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D1%8F

Часть 1
Поиск в корпусах университета Лидса

1. Простой поиск в корпусе

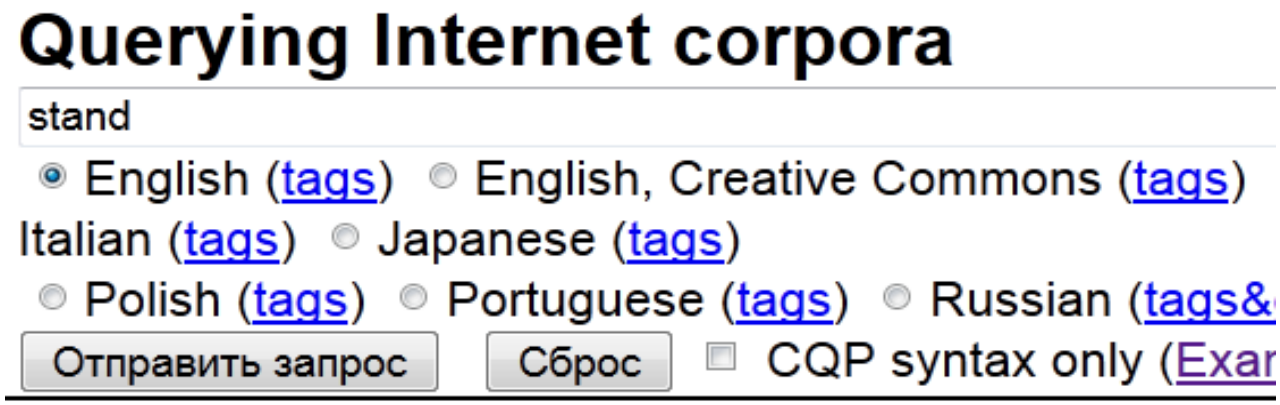
- Меню поиска:

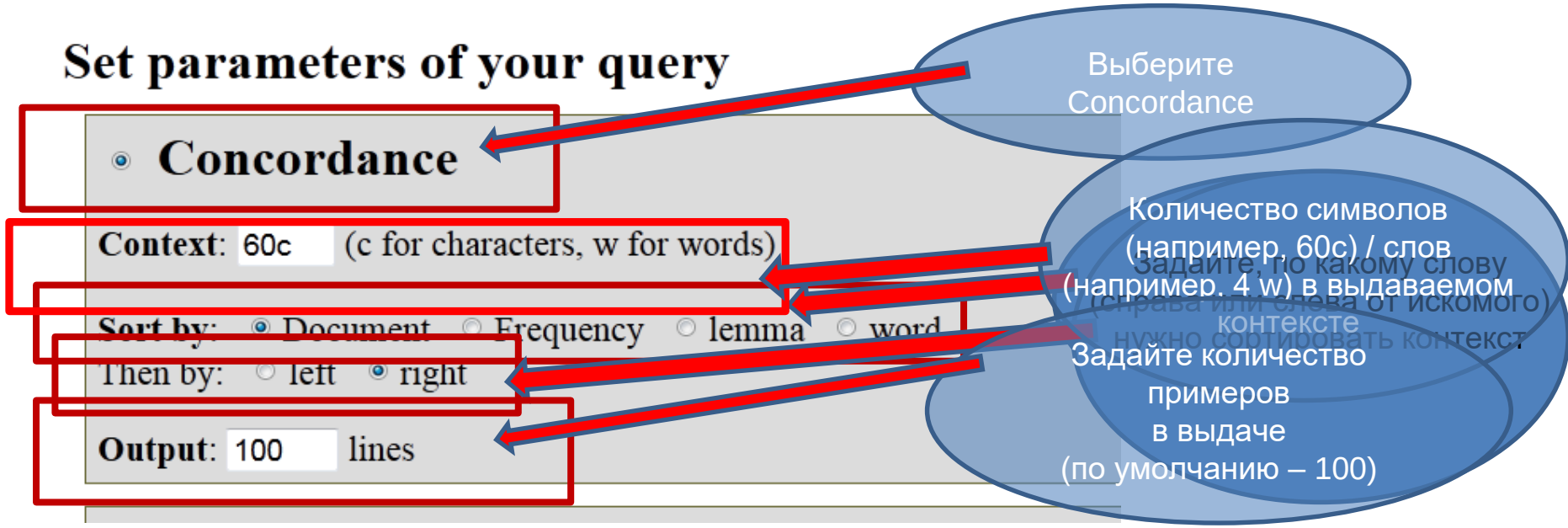




1. Поиск в корпусе
Управление выдачи

- Меню поиска:





1. Поиск в корпусе

See 100 examples of '[word="stand"] cut 100' in INTERNET-EN

>> on the floor and use furniture to pull themselves up to **stand** . Baby walkers are unsafe and not recommended. They can tip
>> agree. 690 (iii) GO from me. Yet I feel that I shall **stand** Henceforward in thy shadow. Nevermore Alone upon the
>> on, through loves eternity. 692 (v) WHEN our two souls **stand** up erect and strong, Face to face, silent, drawing nigh and
>> recoil away And isolate pure spirits, and permit A place to **stand** and love in for a day, With darkness and the death-hour
>> . You 'll find us at the Department of Human Services **stand** where our chef will be demonstrating some of the healthy
>> thou, provd, much enduring, Wave-tossd Wanderer! Who can **stand** still? Ye fade, ye swim, ye waver before me. The cup again!
>> become producers once more rather than gamblers. We could **stand** on our own two feet and take care of our own citizens, and
>> magnificent and highly functional buildings that currently **stand** to the glory of God on our site at Oxford Falls. Click on
>> are experiencing financial hardship. 15 October 1998 CDP 's **stand** against Hansonism Reverend the Hon. F. J. NILE [11.42 a.m.

Упрощенный поиск в корпусе

- Lemma vs. word
- % - поиск леммы
- | - или

Найти: 1) все формы *stand up*

- stand% up

2) все формы глаголов stand и come

- come|stand%

>>	recoil away And isolate pure spirits, and permit A place to	**stand**	and love in for a day, With darkness and the death-hour
>>	their blankets, Asleep on the hills. What forms are these	**coming**	So white through the gloom? What garment out-glistening The
>>	What voices enrapture The nights balmy prime? Tis Apollo	**comes**	leading His choir, The Nine. The Leader is fairest, But
>>	need? Results of the Breakout Session: The recommendations	**coming**	out of our breakout session will cover priorities for
>>	seen Cross and recross the strips of moon-blanchd green;	**Come**	, Shepherd, and again begin the quest. Here, where the
>>	the sun all morning binds the sheaves, Then here, at noon,	**comes**	back his stores to use; Here will I sit and wait, While to
>>	? On cold mornings when you breathe outside, your breathe	**comes**	out as a fog like substance. 2. On some mornings grass and

Регулярные выражения

Используется CQP - POSIX egrep notation of regular expressions (см., например, http://www.regular-expressions.info/)

- . – любой символ
- **[aou] –** символ из списка
- [a-z] – диапазон символов
- **[^...]** любой символ, кроме символов в скобках
- \. – основной (не служебный символ)

Пример 1

Найти глагол ***sing*** и его неправильные формы – ***sing, sang, sung***

Если воспользоваться строкой “**s.ng**”

Получим ***sing, sang, song, sung***

Вариант 1.

“s[iau]ng”

Вариант 2.

“s[^o]ng”

NB: поисковое выражение должно быть в кавычках

Регулярные выражения

Используется CQP - POSIX egrep notation of regular expressions (см., например, http://www.regular-expressions.info/)

- . – любой символ
- **[aou] –** символ из списка
- **[a-z]** – диапазон символов
- **[^…]** любой символ, кроме символов в скобках
- **\.** – основной (не служебный символ)

Пример 2

Найти ***song*** в любом месте предложения, в том числе в самом начале

Если воспользоваться строкой “**song**”

Получим ***song***, но не ***Song***

“[Ss]ong”

NB1: поисковое выражение должно быть в кавычках

NB2: в корпусе Leedsпоиск производится в пределах **одного** слова

Регулярные выражения

Используется CQP - POSIX egrep notation of regular expressions
(см., например, http://www.regular-expressions.info/)

Пример 3
Найти числа, разделенные
точкой: два разряда до точки,
два разряда после точки
(например, ***14.51***)

“[0-9][0-9].[0-9][0-9]”

Нашлось в том числе и 37413

Ответ:

“[0-9][0-9]\.[0-9][0-9]”

Квантификация

Используется CQP - POSIX egrep notation of regular expressions

- Справка:
- * - 0 или более раз
- + - 1 или более раз
- ? – 0 или 1 раз
- {n,m} не менее n и не более m раз
- | - или
- () – группа символов

Пример 4. Найти числа все междометия типа ***hm***, ***hmm*** и т.п.:

“[Hh]mm*”

“[Hh]m+”

Пример 5. Найти прилагательное ***colour*** – британский и американский вариант написания

“colou?r”

Пример 6. Найти все формы глагола ***sing***

“s[iau]ng(s|ing)?”

NB: выбираем SQP syntax only, поисковое выражение должно быть в кавычках

Summary 1

. – любой символ

[aou] – символ из списка

[a-z] – диапазон символов

[^…] любой символ, кроме символов в скобках

\. – основной (не служебный символ)

() – группа символов

- Квантификаторы:
- * - 0 или более раз
- + - 1 или более раз
- ? – 0 или 1 раз
- {n,m} не менее n и не более m раз
- {m} – ровно m раз
- | - или

NB квантификаторы применяются к символу или к группе символов в (), за которыми следует

Пример 6. Найти все формы глагола ***sing***

"s[iau]ng(s|ing)?"

Пример 7. Найти названия, состоящие из 2-х частей типа ***BioMed***, ***AutoStream***, где первая часть из списка: ***Med, Bio, Auto, Tele***

"(Auto|Med|Bio|Tele|Med)[A-Z][a-z]+"

NB: выбираем SQP syntax only, поисковое выражение должно быть в кавычках

Регулярные выражения

- Упражнение:
- 1) найти существительные, оканчивающиеся на -ization – британский и американский вариант написания
- 2) найти все фамилии, начинающиеся на Mac или Mc

NB: выбираем SQP syntax only, поисковое выражение должно быть в кавычках

Упражнения

- http://corpus.leeds.ac.uk/protected/
- Найти все формы глагола *drink,* используя метасимволы

Непечатаемые символы, обозначающие конец строки:

- Win-формат: CR+LF \r\n
  old-Mac-формат: CR \r
  UNIX-формат: LF \n

Классы символов

- \d - цифры
- \w – буквы и цифры
- \s - пробелы, табуляция и перенос строки
- \t - табуляция

Метасимволы

^	Соответствует началу текста (или началу любой строки в мультистроковом режиме).
$	Соответствует концу текста (или концу любой строки в мультистроковом режиме).
\(\)	Объявляет «отмеченное подвыражение», которое может быть использовано позже (см. следующий элемент: \n). «Отмеченное подвыражение» также является «блоком». В отличие от других операторов, этот (в традиционном синтаксисе) требует бэкслеша.

Квантификация

- Квантификация
- *Квантификатор* после символа или группы определяет, сколько раз предшествующее выражение может встречаться.
- {m,n} общее выражение, повторений может быть **от m до n включительно**.
- {m,} общее выражение, **m и более повторений**.
- {,n} общее выражение, **не более n повторений**.
- ? Знак вопроса означает **0 или 1** раз, то же самое, что и **{0,1}**. Например, «colou?r» соответствует и *color*, и *colour*.
- * Звёздочка означает **0, 1 или любое число** раз (**{0,}**). Например, «go*gle» соответствует *ggle*, *gogle*, *google* и др.
- + Плюс означает **хотя бы 1** раз (**{1,}**). Например, «go+gle» соответствует *gogle*, *google* и т. д. (но не *ggle*).

Перечисление

- Вертикальная черта разделяет допустимые варианты. Например, «gray|grey» соответствует *gray* или *grey*.

Группировка

- Круглые скобки используются для задания группы символов.
- Например, «gray|grey» и «gr(a|e)y» являются разными образцами, но они оба описывают множество, содержащее *gray* и *grey*.

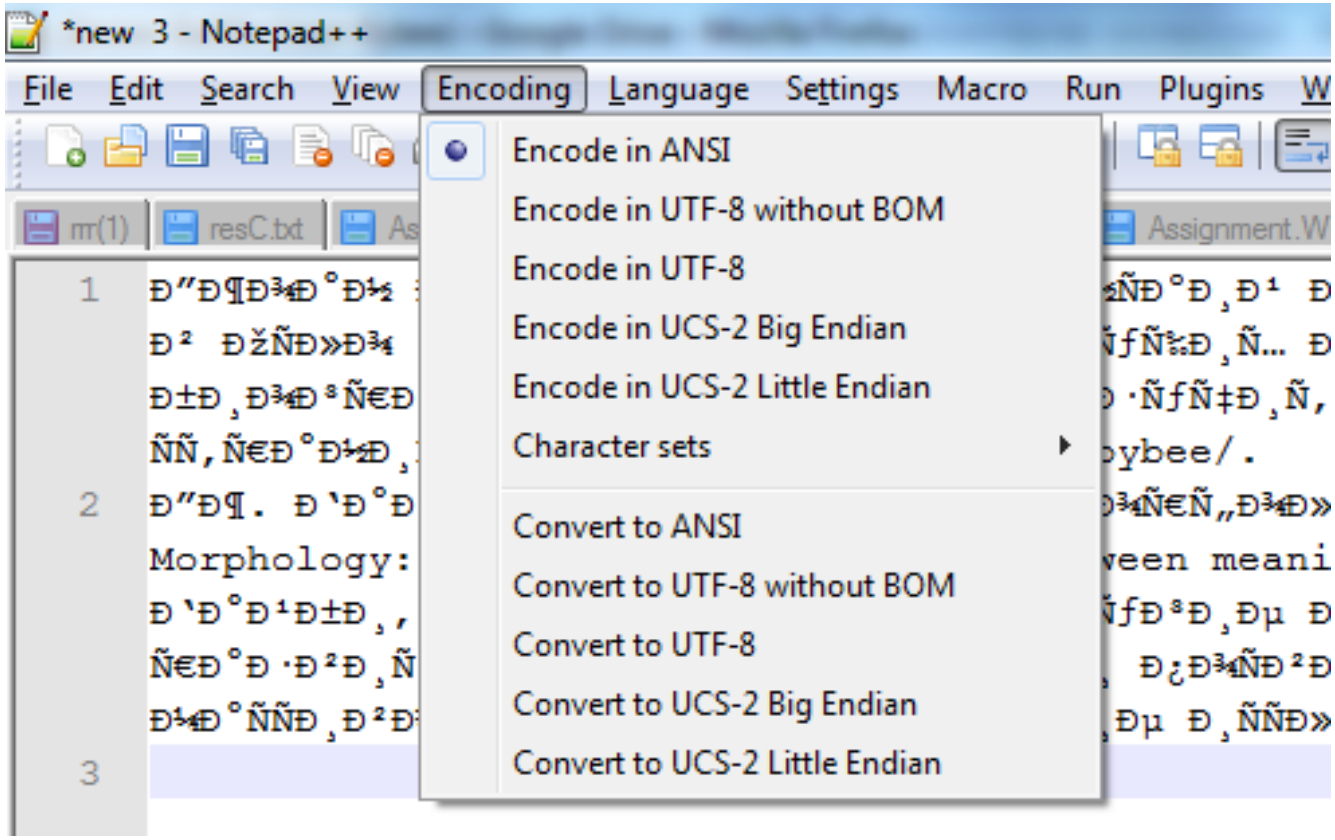
Кодировка

- Типы кодировок текста для русского языка: DOS, KOI8-R, Cyrillic Windows cp1251, Unicode (UTF-8 с/без BOM, цифровой подписи)

Упражнение в Notepad++

- В меню "**Кодировки**" замените текущую на Windows-1251 (оцените эффект:), затем обратно на UTF-8 без BOM.
  Если вы открываете файл и видите "кракозябры", то с помощью меню "Кодировки" можно подобрать правильную.
  "Преобразовать в...": переводит файл в другую кодировку (затем файл нужно сохранить). Windows-1251 называется ANSI.

Кодировка


*new 3 - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins W
Encode in ANSI
Encode in UTF-8 without BOM
Encode in UTF-8
Encode in UCS-2 Big Endian
Encode in UCS-2 Little Endian
Character sets
Convert to ANSI
Convert to UTF-8 without BOM
Convert to UTF-8
Convert to UCS-2 Big Endian
Convert to UCS-2 Little Endian

Непечатаемые символы, обозначающие конец строки:

- Win-формат: CR+LF \r\n
  old-Mac-формат: CR \r
  UNIX-формат: LF \n