Linguistica Computazionale

Salvatore Sorce
Dipartimento di Ingegneria
Chimica, Gestionale, Informatica e Meccanica

Ludici Adattati da Alessandro Lenci
Dipartimento di Linguistica "T. Bolelli"

Espressioni Regolari

Espressioni Regolari (RE)

"A language for specifying text search strings

A string is a sequence of symbols

- *Broadly speaking,any sequence of alphanumeric characters (letters, numbers, spaces, tabs, and punctuation)*

Formally, a Regular Expression is an algebraic notation for characterizing a set of strings."

Wildcard

- La RE /./ corrisponde a qualsiasi carattere (eccetto il ritorno-a-capo)
- /./ /.*/

RE	Definizione	Esempi di "matching"
/b.s/	qualsiasi stringa di tre caratteri che inizia con 'b' e termina con 's'	"bas" "bbs" "b3s" "b!s" "b s" "b,s" … "baas"
/b.*s/	qualsiasi stringa che inizia per b e termina per s	"bs" "bas" "bbs" "b3s" "b!s" "b s" "b,s" "baas" "bisogna prendere l'autobus"
/.*/	qualsiasi stringa (compresa quella vuota)	*l'intero corpus, anche se vuoto*

Raggruppamento e memoria

- Le parentesi tonde servono per raggruppare stringhe di caratteri da moltiplicare:

RE	Definizione	Esempi di “matching”
/(ab)+/	una o più stringhe “ab”	“ab” “abab” “ababab”
/ab+/	una “a” seguito da una o più “b”	“ab” “abb” “abbb”

Le parentesi tonde memorizzano la stringa di testo corrispondente al contenuto delle parentesi:

- la stringa viene memorizzata in una variabile temporanea
- Il contenuto della variabile può essere richiamato con \<numero>
- 1 = contenuto della prima coppia di parentesi; 2 = contenuto della seconda coppia di patentesi, ecc.

/(le|gli)(il|lo|la)+\1/ \([bcdfghjklmnpqrstvwxyz])+\1/

Raggruppamento e memoria

RE	Definizione	Esempi di “matching”
/(ab)+\1/	la variabile “\1” corrisponde a qualunque stringa abbia fatto matching con il contenuto delle parentesi	“abab” “abababab” “abababababab”
/(a)+(b)+\1\2/	la variabile “\1” corrisponde a qualsiasi stringa abbia fatto matching con il contenuto della prima coppia di parentesi; la variabile “\2” idem, ma rispetto alla seconda coppia di parentesi	“abab” “abbabb” “aabaab” “abbaab”
/p(.)o p\1o/	la variabile “\1” corrisponde a qualunque stringa abbia fatto matching con il contenuto delle parentesi	“pio pio” “pao pao” “pro pro”, ecc. “pio pao” “pao peo”
/p.o p.o/	la wildcard “.” può essere sostituita da qualsiasi carattere	“pio pio” “pio pao” “pro pso”, “pao pio”, ecc.
/(a)(b)+\1\2/ /a(b)+a\1/		“abab”, “abbabb”, “aabaab”, “abbaab”

Priorità

Ordine	Tipo	Esempi di “matching”
1	parentesi	( )
2	moltiplicatori	? + * {m,n} {m,} {n} ?? +? *?
3	Sequenza e ancore	“cane” ^ $ \b \B
4	alternativa	|

Caratteri speciali

- Alcuni caratteri hanno un significato speciale nel linguaggio delle RE
  - [ ] ? * . ( ) + - / { }
- Se questi caratteri fanno parte del pattern di testo da cercare, devono comparire in una RE con davanti il carattere '\' (carattere di escape)
  - /[a/ ERRORE! '[' è interpretato come classe di carattere e manca la parentesi ']'
  - /\[a/ la stringa "[a"
  - /a./ qualsiasi stringa di due caratteri che inizia con 'a' "ab" "au" as" "a1" "a?" …
  - /a\./ la stringa di testo "a."
  - /cane?/ le stringhe "cane" e "can"
  - /cane\?/ la stringa "cane?"

Ancore

- Le ancore sono caratteri speciali che specificano dove deve comparire il pattern di testo da cercare
  - /^<pattern>/ il <pattern> deve comparire all'inizio di una linea
  - /<pattern>$/ il <pattern> deve comparire alla fine di una linea

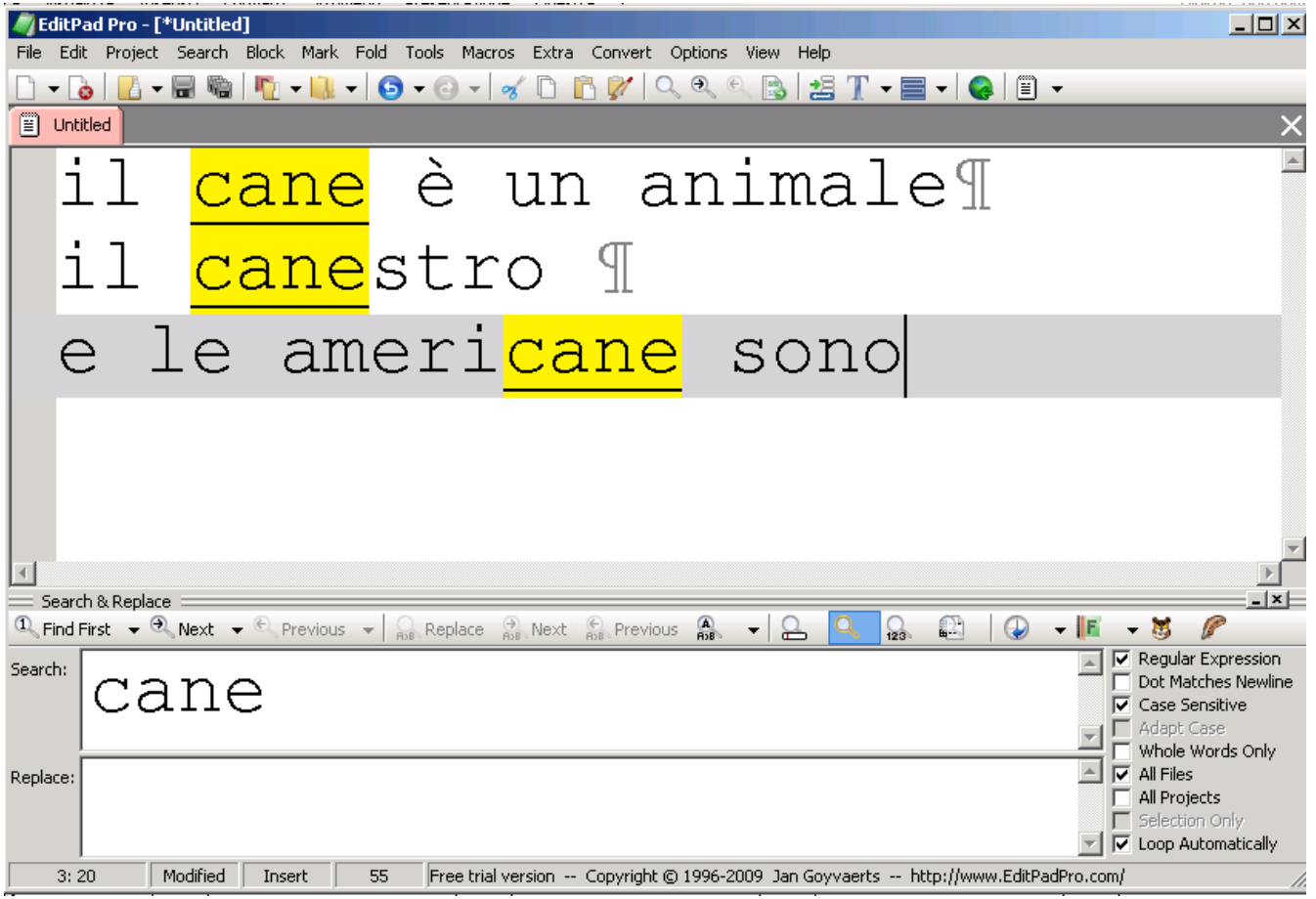
RE	Definizione	Esempi di "matching"
/cane$/	la stringa 'cane' quando compare alla fine di una linea	"…cane¶" "il cane di Mario"
/^La/	la stringa 'La' quando compare all'inizio di una linea	"¶La macchina era guasta" "il treno per La Spezia"
/^La Spezia$/	una riga che contiene solo la stringa "La Spezia"	"¶La Spezia¶" "…a La Spezia per lavoro …"

Ancore

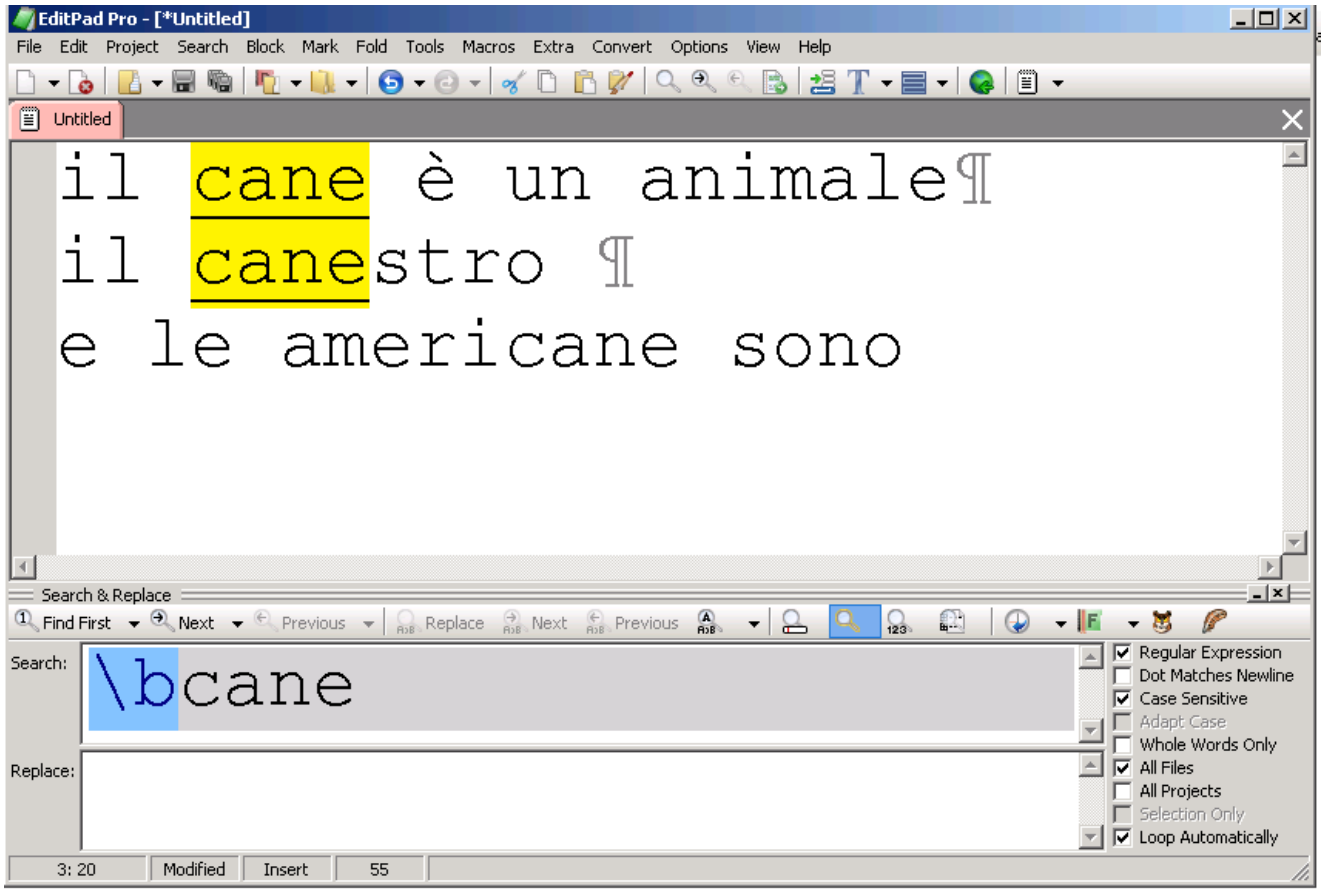
- "\b" è un ancora che indica il confine di una parola ("\B" indica ogni punto non confine di parola)
  - Il confine di una parola è un punto che ha da una parte un carattere di classe \w e dall'altra o un carattere di classe \W o l'inizio (fine) riga
  - **ATTENZIONE**: I caratteri accentati (à, è, è, ì, ò, ù) fanno parte della classe \W (così come lo spazio e gli altri segni di punteggiatura)

RE	Definizione	Esempi di "matching"
/\bcane\b/	la stringa 'cane' deve avere a destra e a sinistra un confine di parola	"il **cane** è …" "il canestro" "le americane sono"
/\Bcane\b/	la stringa 'cane' deve avere a destra (ma non a sinistra) un confine di parola	"il cane è …" "il canestro" "le ameri**cane** sono"
/\bè\b/ **/\Bè\B/**	Trovare 'è' come copula	"Pinocchio **è** stanco" "Pinocchio è stanco"

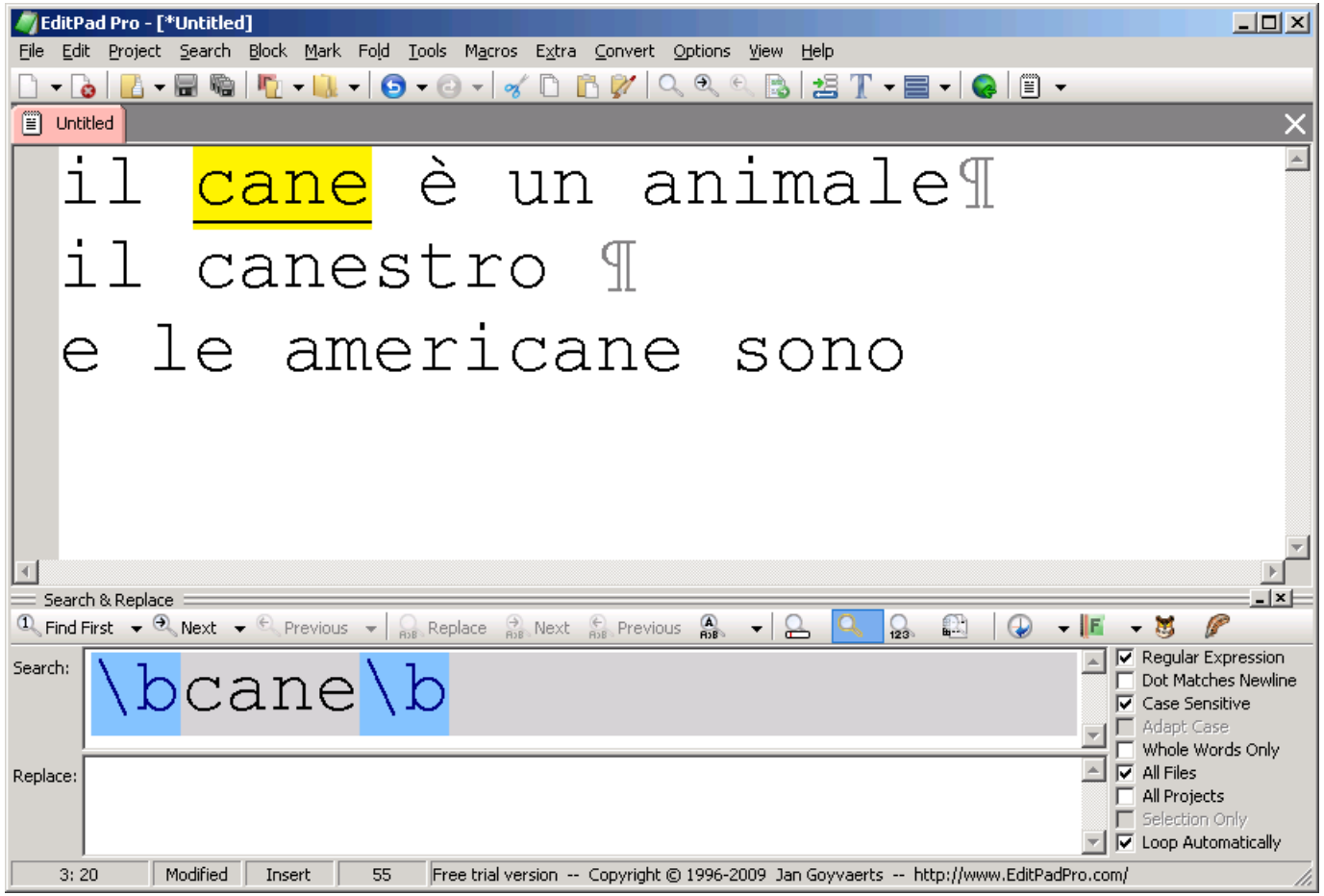
Esempio con EditPad Pro



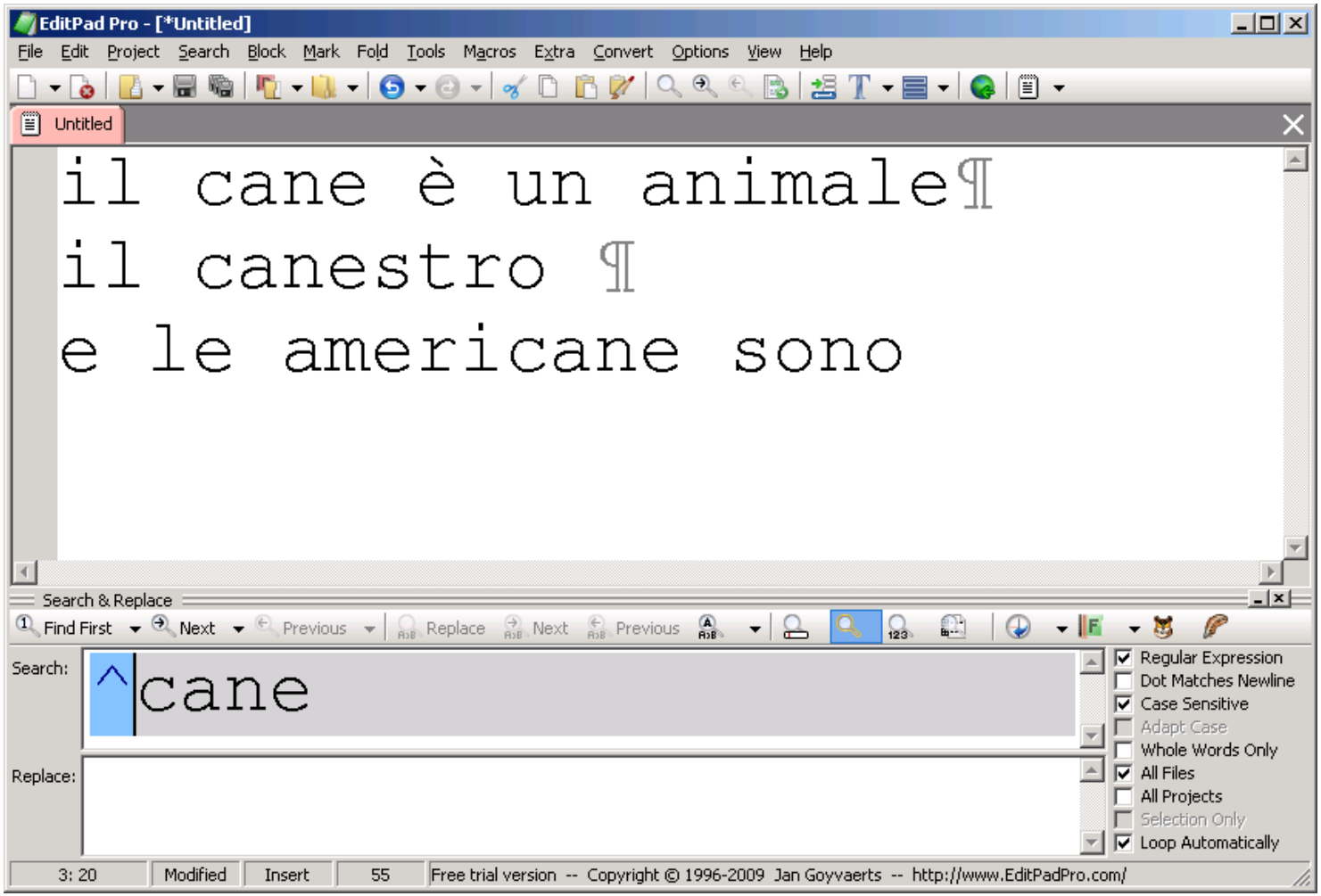
Esempio con EditPad Pro



Esempio con EditPad Pro

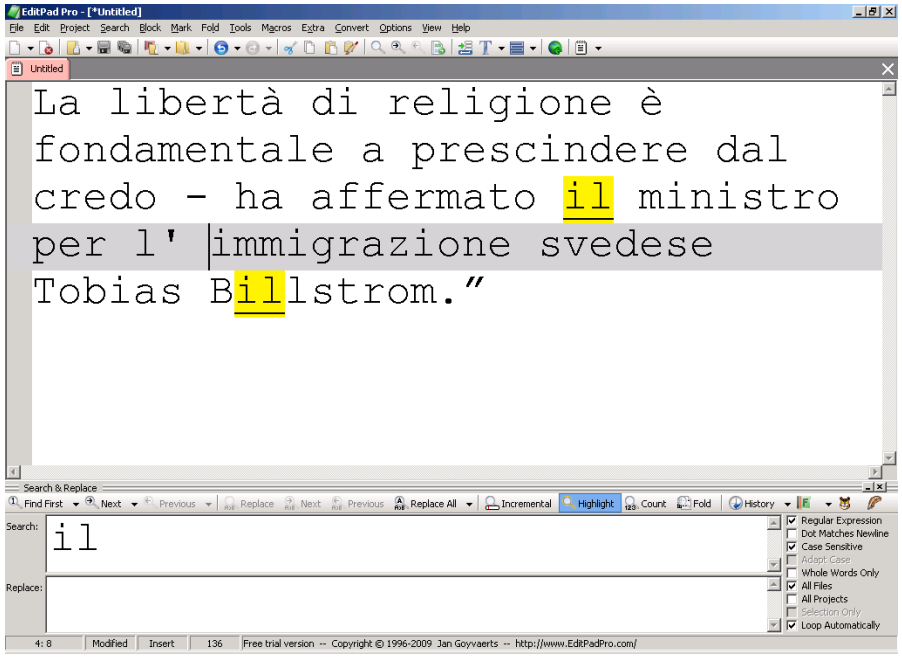


Esempio con EditPad Pro

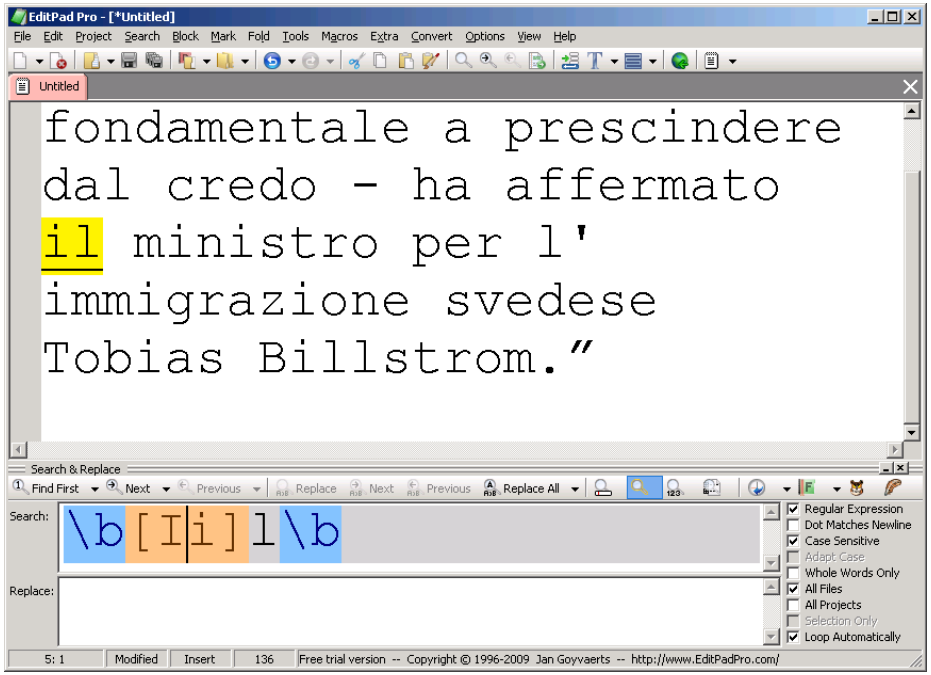


Esempi con EditPad Pro

- Scrivere una espressione regolare per trovare l'articolo "il":



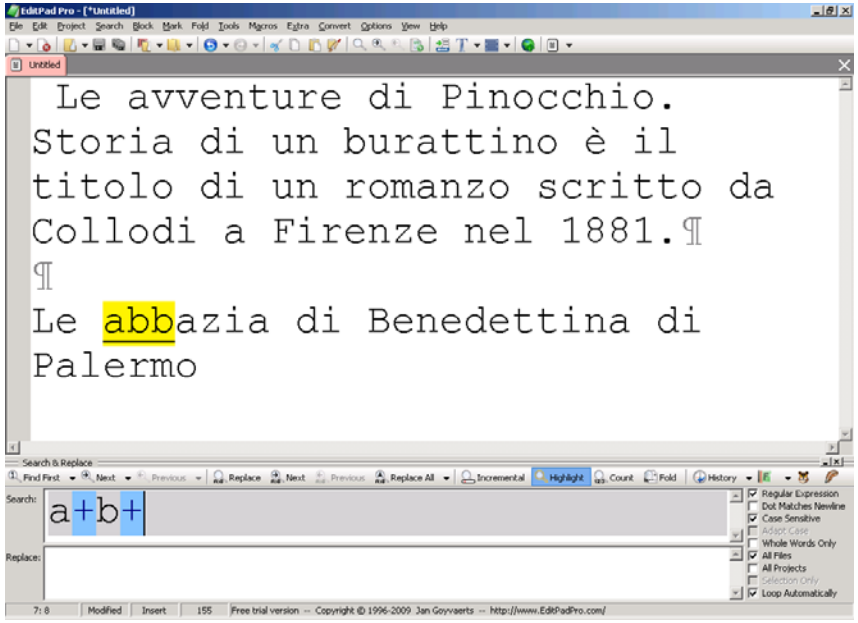
Errata



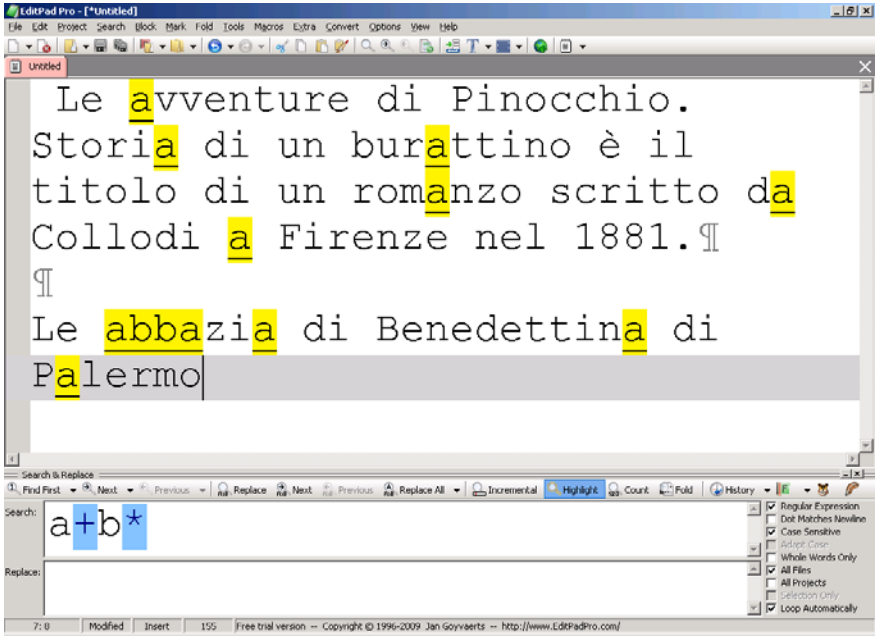
Corretta

Esempi

- Scrivere una espressione regolare che trovi almeno una ‘a’ seguita da un qualunque numero di ‘b’



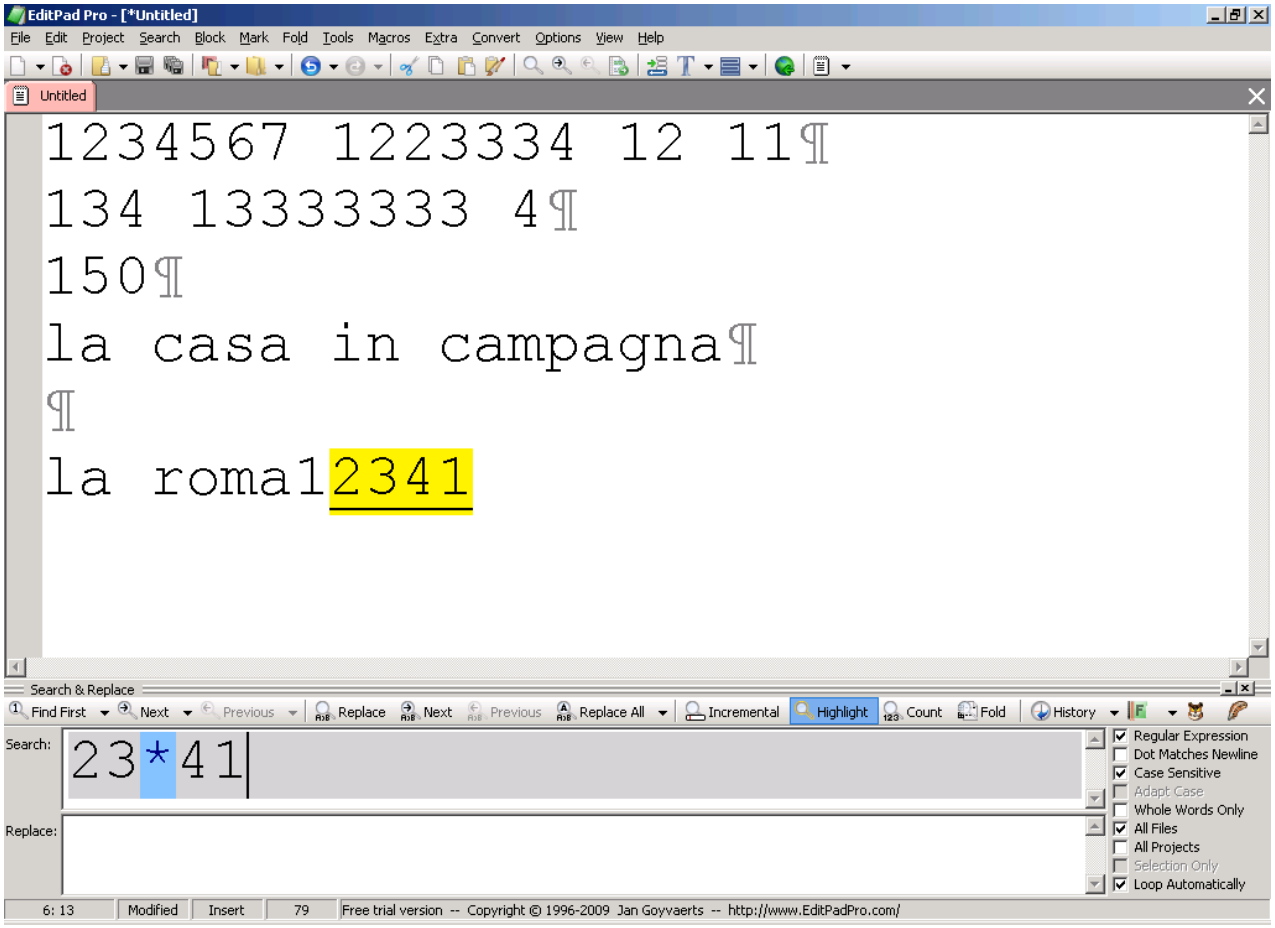
Errata



Corretta

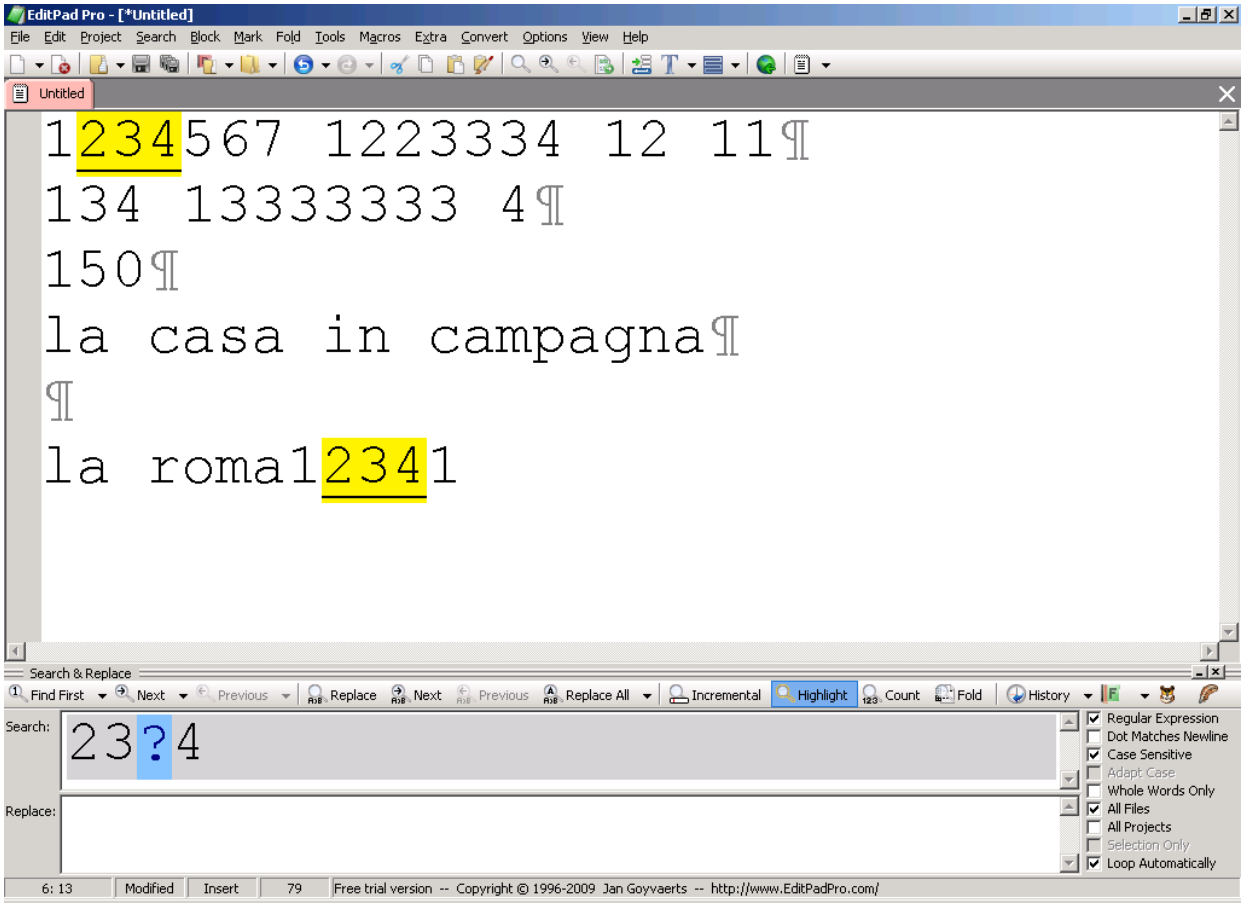
Esempi con EdiPad Pro

- Moltiplicatori di carattere



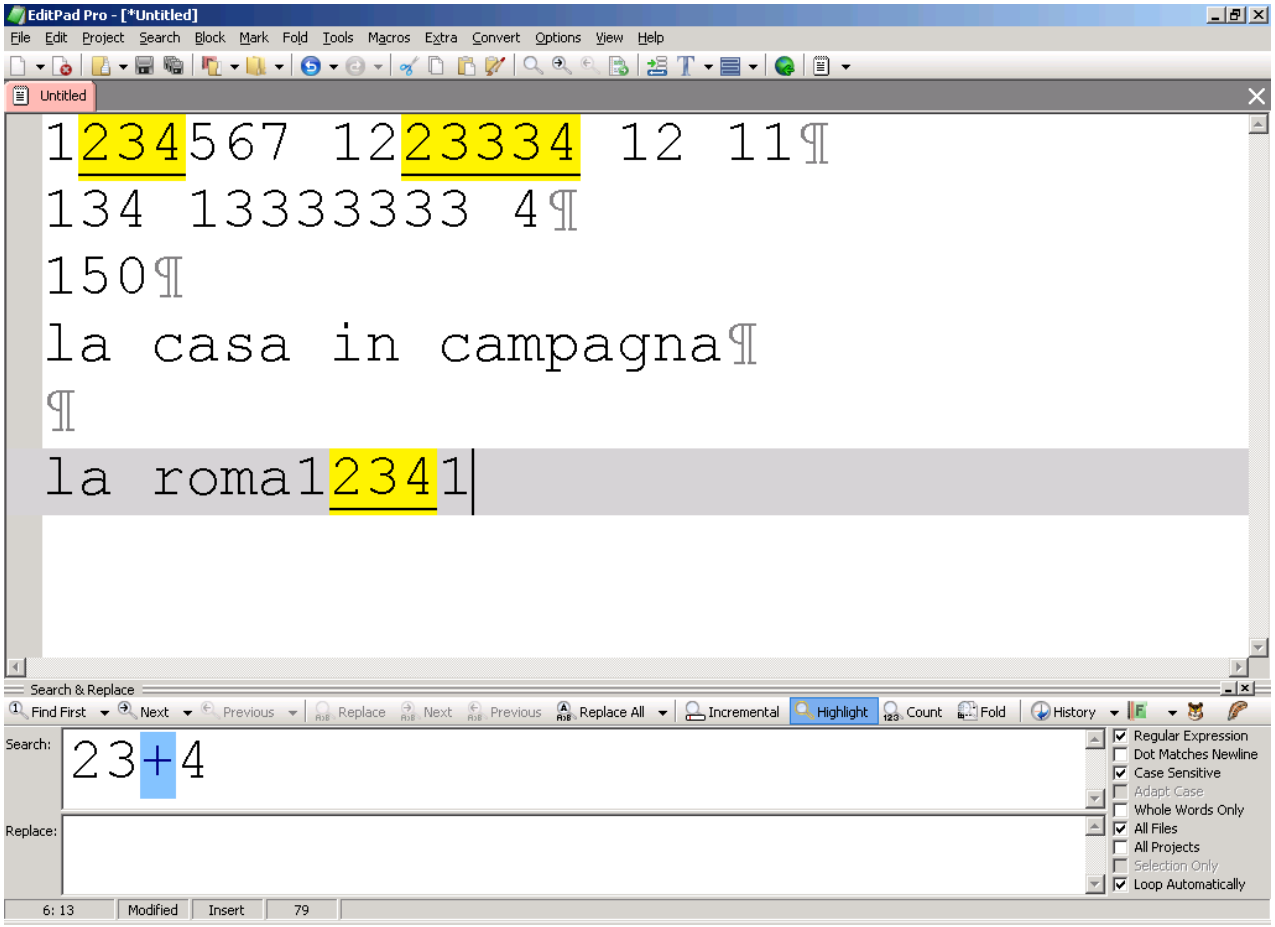
Esempi con EdiPad Pro

- Moltiplicatori di carattere



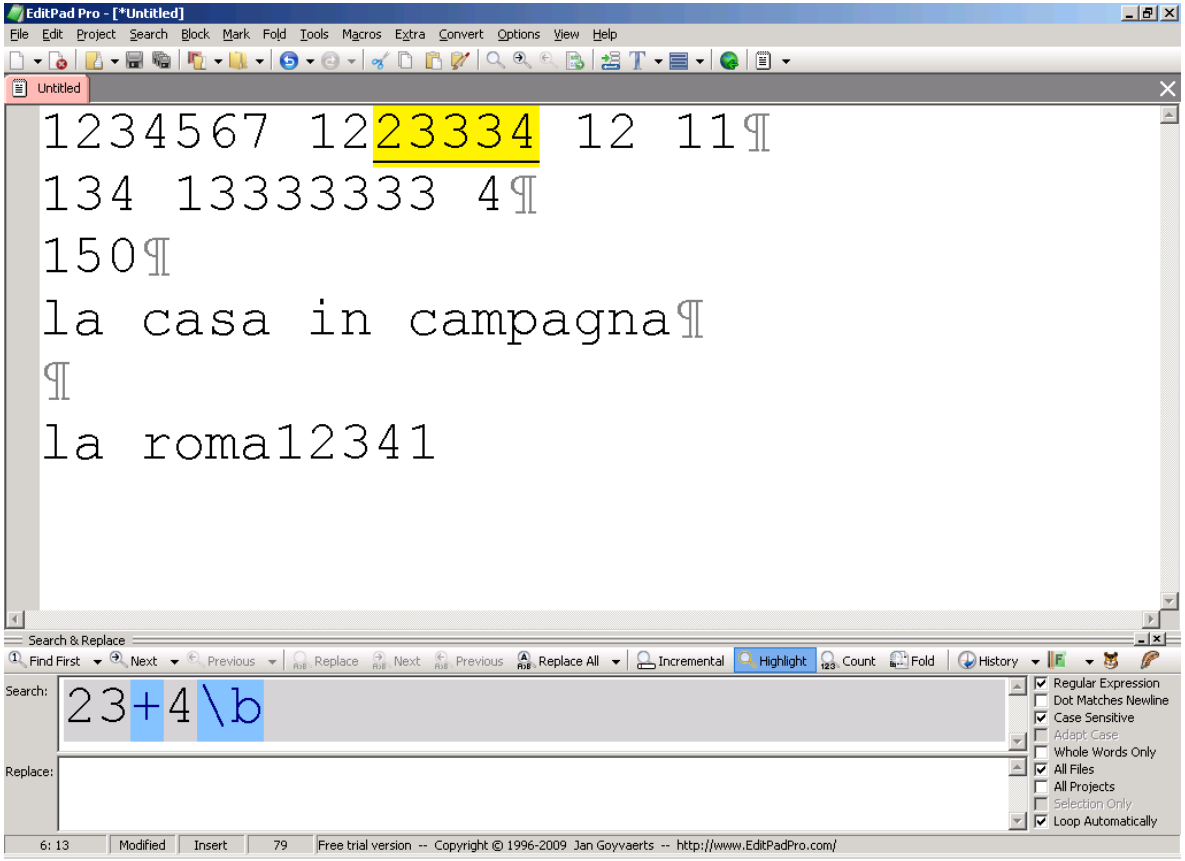
Esempi con EdiPad Pro

- Moltiplicatori di carattere



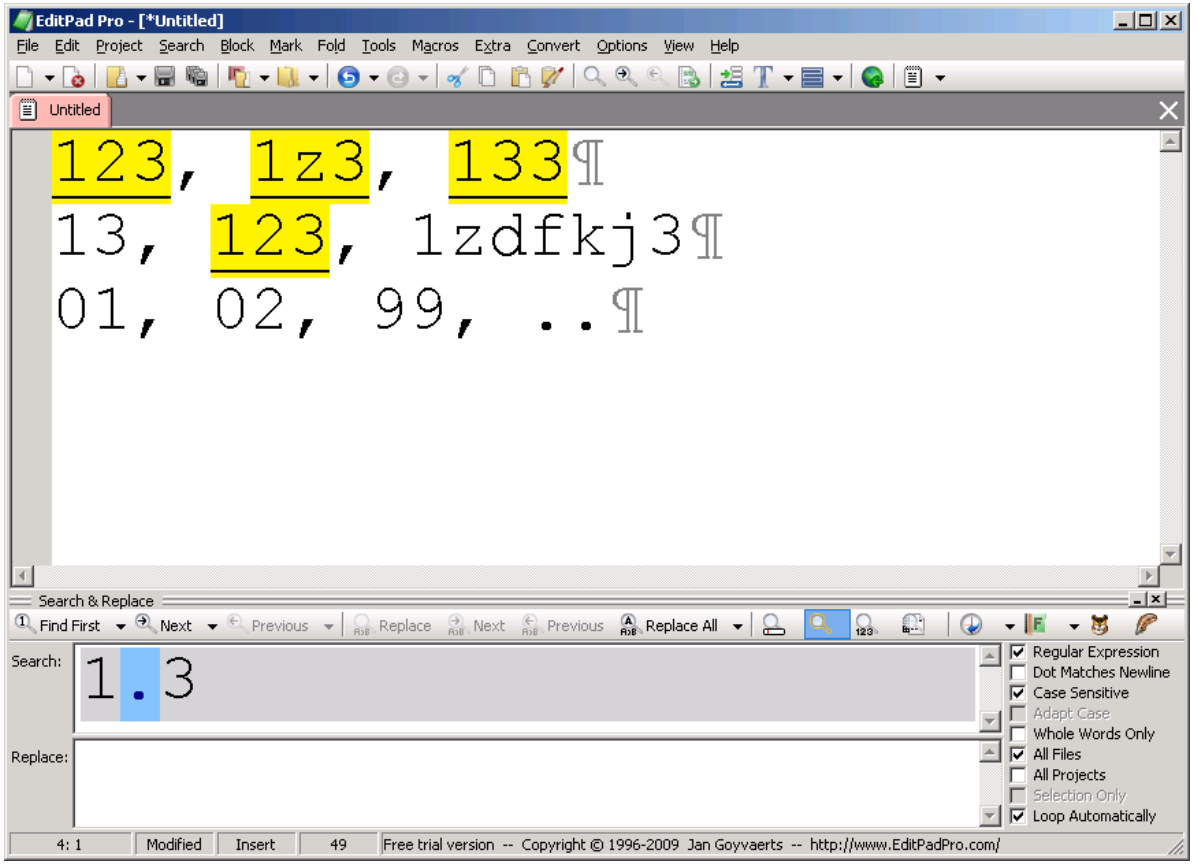
Esempi con EdiPad Pro

- Moltiplicatori di carattere



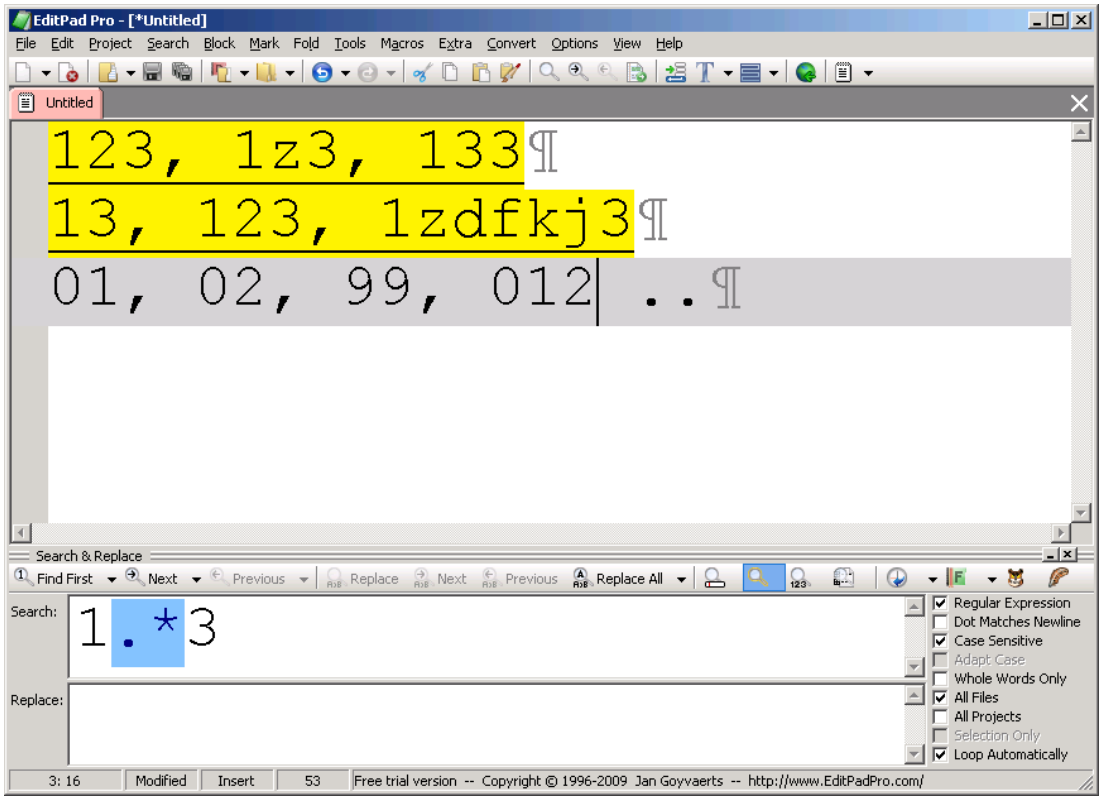
Esempi con EditPad Pro

- ‘.’ qualunque carattere eccetto ritorno a capo



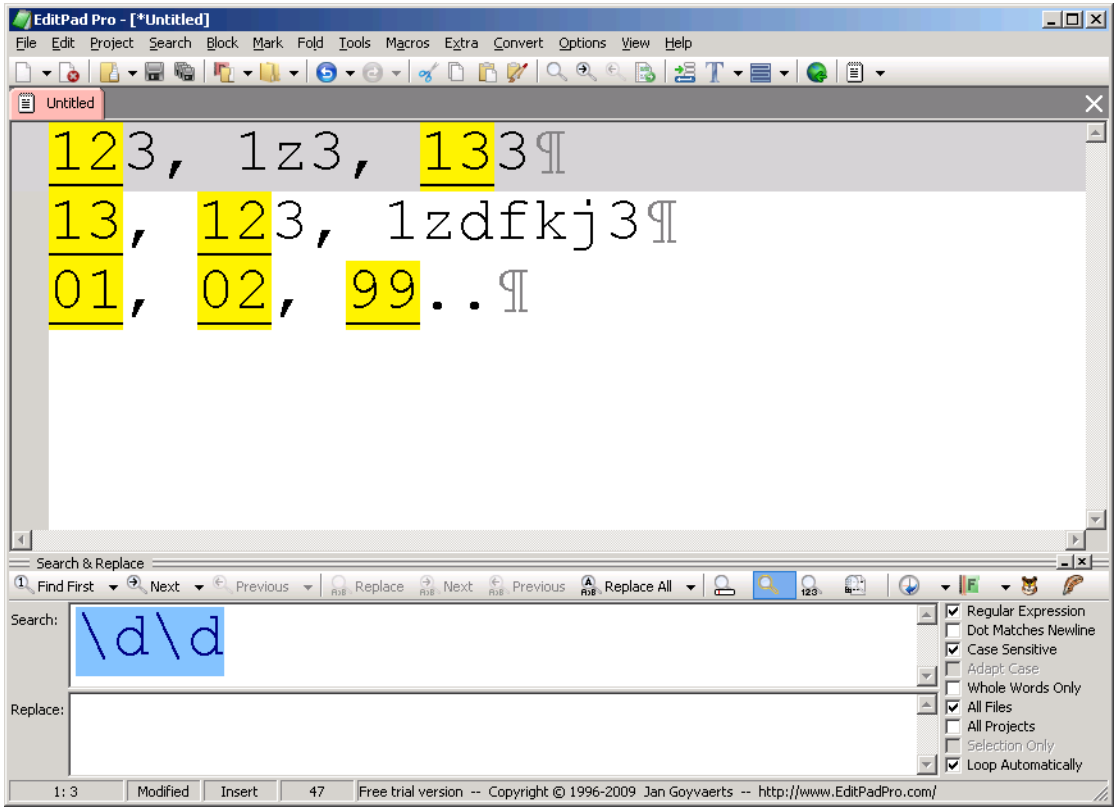
Esempi con EditPad Pro

- ‘.’ qualunque carattere eccetto ritorno a capo



Esempi con EditPad Pro

- due interi tra 0 e 9



Esempi

- Classi di caratteri

RE	Matchs	Does Not Match
/[^ab]/	**c, d, z**	ab
^[1-9][0-9]*$	**Qualunque numero positivo**	Zero, numeri negativi o decimali
[0-9]*[,.]?[0-9]+	**0.1 .8 1 1.2 100,000**	12.

Esercizi

- Formalizzare con le espressioni regolari i patterns per trovare le seguenti stringhe
  - “tutte le vocali minuscole o maiuscole“

Esercizi

- Formalizzare con le espressioni regolari i patterns per trovare le seguenti stringhe
  - “tutte le vocali minuscole o maiuscole“

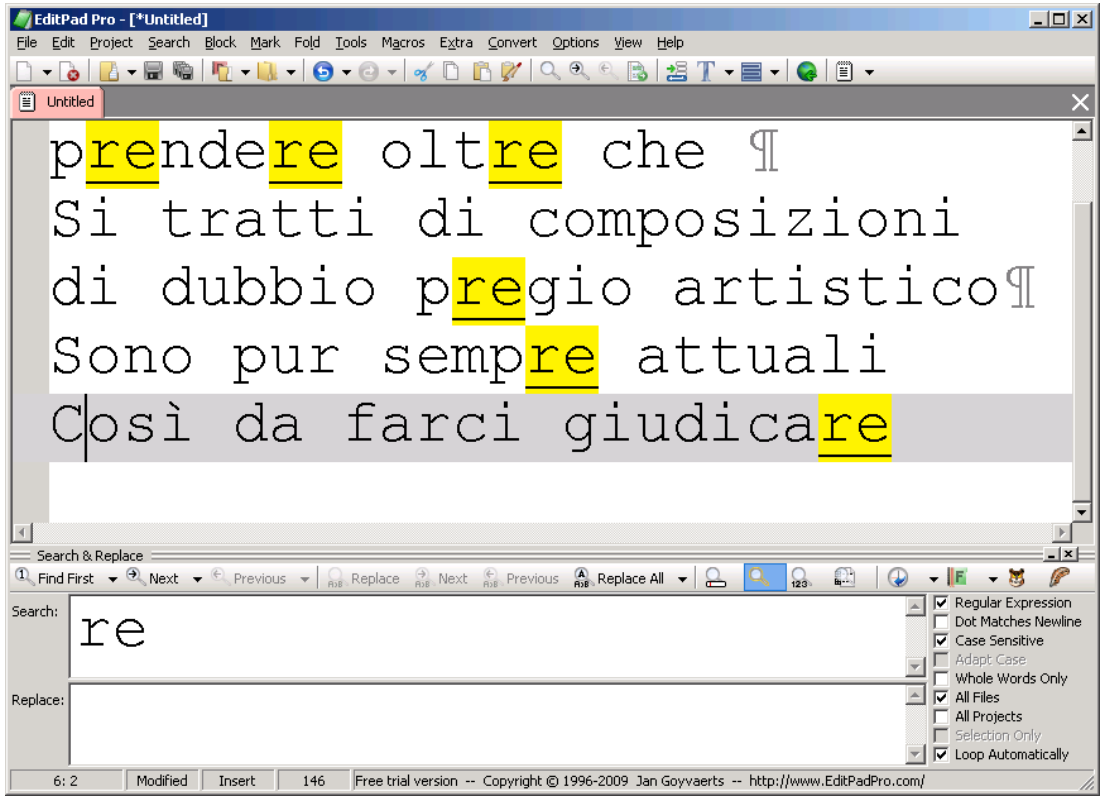
  Sol.: /[AaEeIiOoUu]/

Esercizi EditPad Pro

- Formalizzare con le espressioni regolari i patterns per trovare le seguenti stringhe
  - ‘re’

Esercizi EditPad Pro

- Formalizzare con le espressioni regolari i patterns per trovare le seguenti stringhe
  - ‘re’



Esercizi EditPad Pro

- Formalizzare con le espressioni regolari i patterns per trovare le seguenti stringhe
  - Che terminano in ‘re’

Esercizi EditPad Pro

- Formalizzare con le espressioni regolari i patterns per trovare le seguenti stringhe
  - Che terminano in ‘re’
  - /re/
  - /re\b/
  - /\bre\b/
  - /\Are/
  - /re$/
  - /[^re]/
  - /[re]