Pretraga korpusa

1 Regularni izrazi

Metakarakter	Objašnjenje	Primer	Rezultat
.	bilo koji karakter	ka.a	kada, kafa, kava...
?	0 ili 1 ponavljanje prethodnog izraza	kah?va	kava, kahva
*	0 ili više ponavljanja prethodnog izraza	joo*j	joj, jooj, joooj...
+	1 ili više ponavljanja prethodnog izraza	jo+j	joj, jooj, joooj...
{n}	n ponavljanja prethodnog izraza	jo{3}j	joooj
{n,m}	n-m ponavljanja prethodnog izraza	jo{2,3}j	jooj, joooj
{n,}	najmanje n ponavljanja prethodnog izraza	jo{2,}j	jooj, joooj, jooooj...
[...]	bilo koji od nabrojanih karaktera	ka[fv]a	kafa, kava
[...-...]	bilo koji iz niza karaktera	[A-Z]	A,B,C...Z
[^...]	bilo koji karakter koji nije nabrojan	[^AEIOU]	B, C, D, F...
[^...-...]	bilo koji karakter koji nije u nizu	[^A-Z]	a, b, 0, 1...
...|...	ili	kafa|kava	kafa, kava
(...)	grupisanje	ka(f|hv)a	kafa, kahva
\	doslovna upotreba metakaraktera	dr\.	dr.

Napomena:

- Cela SR/HR latinica: [A-ž], velika slova: [A-ZĆČĐŠŽ], mala slova: [a-zćčđšž]

2 CQL (*Corpus Query Language*)

Osnovna sintaksa: **[atribut = "vrednost"]** (za predefinisani atribut može se zadati samo **"vrednost"**)

Atribut	Objašnjenje	Primer	Rezultat
word	oblik reči	[word = "korpus"]	korpus
lemma	svi oblici reči	[lemma = "korpus"]	korpus, korpusi, korpusa...
pos/tag	oznaka vrste reči	[tag = "N"]	korpus, računar, anotacija...
Sintaksa	**Objašnjenje**	**Primer**	**Rezultat**
[... & ...]	više atributa	[word = "kose" & tag = "V"]	kose (kao glagol)
[...] [...]	više tokena	[word = "veb"] [lemma = "korpus"]	veb korpus, veb korpusi...
[ ]	bilo koji token	"novi" [ ] "korpus"	novi veb korpus, novi ReLDI korpus...
!	negacija	[word = "kose" & tag != "V"]	kose (kao ne-glagol)

Napomene:

- NoSke za oznake vrsta reči koristi atribut *tag*
- Razmaci unutar i između zagrada nisu obavezni, pojedini interfejsi ih ni ne dozvoljavaju
- Unutar vrednosti atributa mogu se koristiti regularni izrazi (npr. [lemma = "korpus.+"], [tag = "N.*"]); neki metakarakteri mogu se primenjivati i na cele tokene (npr. "novi" ([tag = "A.*"]|[lemma = "veb"])? "korpus", "novi" [ ]{2} "korpus")
- Za svaki korpus potrebno je pronaći skup oznaka vrsta reči (eng. *tagset*) koji koristi