Orodja za delo s korpusi

Tomaž Erjavec
Korpusno jezikoslovje / Jezikovne tehnologije
UNG
2009/2010

1. 3. 2010

Pregled predavanja

1. uvod
2. konkordančniki
3. prikaz: konkordance, frekvenčni seznami, kolokacije
4. iskanje: regularni izrazi, iskanje po oznakah

Kaj lahko jezikoslovno analiziramo?

A. jezikoslovne lastnosti
   1. besedilo (leksikalne lastnosti)
   2. jezikoslovne oznake (gramatične lastnosti) besedna vrsta, spol, druge pregibne lastnosti, skladenjski odnosi, pomenske oznake,…

B. nejezikovne lastnosti
   1. metapodatki (bibliografski podatki o posameznih besedilih): registri, dialekti, časovna obdobja

Primeri posameznih analiz

A. uporaba besede “zgoščenka” (leksikalna lastnost) ali uporaba glagolnikov (gramatična lastnost)
   1. s katerimi besedami se najbolj pogosto uporablja ali katere besedne vrste se pojavljajo v njeni okolici
   2. koliko se uporablja v tehničnih/netehničnih besedilih ali kakšna je distribucija v korpusu po letih

B. v čem se razlikujejo tehnična od netehničnih besedil (vrsta besedila) ali v čem se razlikujejo besedila pred 1991 od tistih po 1991 (časovno obdobje)
   1. kako se razlikuje leksika
   2. kako se razlikuje uporaba slovničnih vzorcev

A.1 Leksikalna okolica: Word Sketches

zgoščenka **Fida PLUS 620m freq = 14595**

a_modifier	3563	1.4
kompilacijski	40	53.47
priložen	99	51.81
nov	1226	48.64
piratski	36	40.21
multimedijski	48	40.01
glasben	154	35.73
spremljajoč	35	34.0
samostojen	61	29.54
avtorski	38	27.85
računalniški	50	26.99

post_z	1342	8.4
naslov	587	62.42
glasba	166	44.15
posnetek	61	35.25
skladba	42	33.1
pesem	61	31.08

is_obj4	2426	6.0
izdati	839	62.56
posneti	267	47.86
predstaviti	260	35.61
snemati	57	30.84
pripravljati	65	21.69
podariti	31	21.55
kupiti	57	20.01
prejeti	39	17.25
predstavljati	33	13.24
pripraviti	39	10.79

prec_na	1077	5.5
skladba	35	31.73
iziti	51	28.9
posneti	45	27.42
pesem	33	24.55
izdati	41	22.83
glasba	31	22.61
biti	163	21.23
najti	40	16.26

prec_z	498	3.1
nagraditi	106	51.51
glasba	35	27.99

gen_2	1529	2.4
predvajalnik	131	61.11

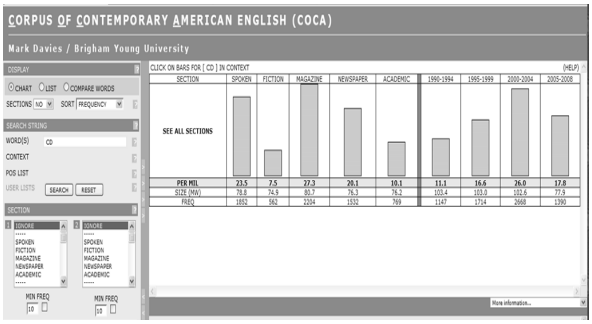
is_obj2	268	2.2
izdati	35	28.26

coord	2463	1.7
kaseta	1153	98.57

is_subj	1046	1.7
iziti	292	56.08

prec_verb:	2195	0.7
izdati	655	58.89

A.2 “CD” v COCA/BYU po registrih in časovnih obdobjih



Orodja za delo s korpusi

Opozorilo:
orodja niso popolna ali vedno intuitivna
- kaj orodje razume kot “besedo”
- ali so vse oznake v korpusu pravilne
- vrednosti statističnih cenilk

zato je rezultate, dobljene iz korpusov, potrebno kritično ovrednotiti

Orodja za poizvedovanje po korpusih

- tipična uporaba:
  iskalni izraz → korpus → prikaz zadetkov
  (→ selekcija → interpretacija → rezultat)
- iskanje:
  - po besedilu, jezikoslovnih oznakah, metapodatkih
  - dobesedno, z izrazi (npr. mehko ujemanje)
- prikaz:
  - besedil, jezikoslovnih oznak, metapodatkov
  - oblikovanje: seznami, tabele, grafi
  - sortiranje

“Konkordančniki”

- najbolj pogosto orodje za raziskovanje korpusov
- poleg samih konkordanc ponavadi ponujajo še druge funkcionalnosti
  - frekvenčni seznami, statistične obdelave
  - sortiranje, filtriranje
  - izbira podkorpusov po metapodatkih
  - hramba, izpis in izvoz najdenega

Vrste konkordančnikov

- nekatere konkordančnike dobimo ali kupimo in namestimo na svoj računalnik
  - sami si moramo zagotoviti korpus(e)
- mrežni konkordančniki
  - ni potrebe po instalaciji, potrebujemo pa mrežno povezavo
  - ponujajo (enega, več) velikih korpusov
  - večina pa ne nudi možnosti za nalaganje lastnih korpusov (izjema: SketchEngine, vendar plačljiv)
- poizvedovalni jeziki, izpis in funkcionalnosti se razlikujejo od orodja do orodja

Bolj pomembni konkordančniki (za nas)

- WordSmith Tools
- FidaPLUS
- SketchEngine
- Vmesniki na IJS:
  - JOS (2 majhna referenčna korpusa)
  - iKorpus (računalništvo in informatika)
  - Dvojezični korpusi (en-sl)

Načini predstavitve podatkov iz korpusa

- konkordance
- frekvenčni seznami in ključne besede
- kolokacije

1. Konkordance

- analiza na osnovi pojavnic
- konkordančno jedro z okoljem: levim in desnim sobesedilom
- ena najstarejših metod za analizo besedi (npr. Konkordance Trubarjevega katekizma, 1983)
- v nasprotju s tiskanimi konkordancami sodobni konkordančniki omogočajo samo izpis želene besede oz. izraza
- dobimo primere uporabe: koristno za določanje pomena
- pri prevelikem številu pojavitev nekateri konkordančniki omogočajo naključno sito
- koristno je lahko tudi sortiranje po jedru ali sobesedilu

iKorpus

da lakti niso oprte na rob mize . Zmotno je mišljenje , da mora biti	miška	oddaljena oziroma ločena od tipkovnice . Sodobne ergonomske ti
) (2) , saj jim standardni računalnik , disketna enota , tipkovnica ,	miška	in monitor tega ne omogočajo . Pojasnimo nekaj težav pri uporab
B , ki je združljiv samo z starejšo oziroma veliko manjšo konzolo ,	miška	, tipkovnico , mrežni vmesnik , VGA-kabel ter programski oprem
odive gibalno oviranim učencem . Čeprav ob besedi tipkovnica ali	miška	večina ljudje pomisli na klasične proizvode osebnih računalnikov
prst . Lahko jih nadomestimo z raznimi stikali . Poznana pa je tudi	miška	katero krmilimo z nogami . Slika 6 : Miška krmiljena s stikali Slika
z nogami . Slika 6 : Miška krmiljena s stikali Slika 7 : " No hands "	miška	(miška krmiljena s nogami) Slika 8 : Klasična miška , kjer premi
ni . Slika 6 : Miška krmiljena s stikali Slika 7 : " No hands " miška (	miška	krmiljena s nogami) Slika 8 : Klasična miška , kjer premik krogle
" No hands " miška (miška krmiljena s nogami) Slika 8 : Klasična	miška	, kjer premik krogle nadomeščajo stikala Slika 9 : Stikalo , ki krm
odive gibalno oviranim učencem . Čeprav ob besedi tipkovnica ali	miška	večina ljudje pomisli na klasične proizvode osebnih računalnikov
. Celoto dopolnjujejo prav tako zanimivi dodatki , kot so prozorna	miška	in zvočniki . Skratka paša za oči , pa tudi zmogljiv računalnik , ki
Pravzaprav je zanimivo , da je moralo preteči toliko časa , da se je	miška	, ki so jo začeli razvijati v šestdesetih letih , tako zelo prijela . Miš
za velike proizvajalce računalnikov (IBM , Siemens) , katerim je	miška	pač le eden od delov strojne opreme , ki jo morajo imeti v progra

Kaj smo iskali?

jih uredimo z nekim znanjem , da poda bogatejšo ,	popolnejšo in uporabnejšo informacijo	, ki ji lahko rečemo vsebina . Sistemi za upravljanje v
na oddaljenem strežniku ali če gre za nezdružljivost	poslovnih in repozitorijskih baz	sistema za upravljanje vsebin , se podatki prenašajo p
anju kakovosti vsebin . Delovni proces naj omogoča	zaporedno in vzporedno sodelovanje	. Zaporedno sodelovanje pomeni , da si uporabniki sl
pošta in spletne strani del poslovnega vsakdanjika v	velikih in manjših podjetjih	, želijo manjša podjetja temu vsakdanjiku dodati tudi
v e-poslovanja ; ni cenovno ugodnih rešitev za vse -	velika in manjša podjetja	; glavni akterji na trgu IT-rešitev veliko obljubljajo , v
ativno izvajanje . Pri tem ne bi smelo biti razlik med	velikimi in manjšimi podjetji	. In tu nastopi standard ebXML s svojim poslanstvom
povežejo , pošljejo sporočila in na njih odgovorijo na	varen ter zanesljiv način	, in postaja - predvsem v ZDA - vse bolj priljubljena
vogim nadzorom države . Njen namen je zagotavljati	individualno in kolektivno varnost	ljudi po naključih izravnavanja nevarnosti . Zavarovaln
aršega gospodarskega vidika in zato predstavljajo vir	politične in družbene moči	(1) . Zavarovalniška dejavnost je zelo primerna za u
na , vendar je zagotovo skrajni čas , da se pristopi k	organiziranemu in sistematičnemu prestrukturiranju	zavarovalniške dejavnosti v skladu s paradigmo elekt
n sektorju . K temu sklepu nas navaja razvoj novih ,	kompleksnih oziroma mejnih produktov	, ki jih ne bi bilo brez sodobnih informacijsko-komun
ovani mejni produkti kot kombinacija zavarovalnih ,	bančnih in skladovskih produktov	(3) . To pomeni , da različne finančne institucije , b
e predstavlja za zavarovalno družbo velik poslovni ,	organizacijski in tehnični izziv	. Nepripravljenost na elektronsko poslovanje bo zara

Pojavnice

- konkordančnik (korpus) razdeli besedilo (niz znakov) na pojavnice (angleško *token)*
- postopku pravimo *tokenizacija*
- pojavnice: besede in ločila
- ponavadi nas dejansko zanimajo samo besede
- koliko pojavnic ima naslednji stavek:
  Jabolke, hruške, itd.
- koliko pojavnic je v:
  “rumeno zelen”, “rumeno-zelen”, “rumenozelen”

2. Frekvenčni seznami

- seznami različnic skupaj s številom pojavitev
- evidentira uporabo besedišča
- pove lahko npr. katere besede so najbolj pogoste v korpusu (jeziku)
- Zipfova distribucija:
  - malo besed je zelo pogostih, dosti besed je zelo redkih
  - približno polovica besed se pojavi samo enkrat

Najbolj pogoste leme v iKorpusu

N°	Hits	Atts	
1	6357	lemma	,
2	4092	lemma	biti
3	4036	lemma	.
4	2691	lemma	in
5	2237	lemma	v
6	1468	lemma	(
7	1421	lemma	)
8	1410	lemma	z
9	1336	lemma	na
10	1294	lemma	za
11	1079	lemma	ki
12	896	lemma	se
13	883	lemma	ta
14	675	lemma	da
15	571	lemma	sistem
16	544	lemma	podjetje
197	68	lemma	zbirka
198	68	lemma	vrsta
199	68	lemma	trije
200	68	lemma	prodaja
201	67	lemma	stanje
202	67	lemma	obdelava
203	67	lemma	kupec
204	67	lemma	internet
205	67	lemma	and
206	66	lemma	vloga
207	66	lemma	treba
208	66	lemma	prednost
209	66	lemma	osnoven
210	65	lemma	namen
211	64	lemma	poleg
212	64	lemma	podatkoven
213	64	lemma	naloga
214	64	lemma	dostop

- napogostejše so funkcijske
- polnopomenske besede vseeno nakazujejo področje, ki ga pokriva korpus
- splošen vtis o korpusu in njegovem besednem zakladu
- koristno kot pripomoček za izbiro posameznih besed za nadaljnjo analizo
- koristno tudi sortiranje (od spredaj ali od zadaj)

Še par primerov

N°	Hits	Atts	Hit
1	9888	lemma	operacijski sistem
2	3196	lemma	informacijski sistem
3	903	lemma	datotečen sistem
4	640	lemma	računalniški sistem
5	514	lemma	nov sistem
6	385	lemma	posloven sistem
7	318	lemma	celoten sistem
8	230	lemma	imeniški sistem
9	226	lemma	različen sistem
10	208	lemma	velik sistem
11	182	lemma	obstoječ sistem
12	150	lemma	zmogljiv sistem
13	146	lemma	varnosten sistem
14	139	lemma	transakcijski sistem
15	121	lemma	nadzoren sistem
16	117	lemma	eksperten sistem
17	105	lemma	pomniliški sistem
18	105	lemma	podoben sistem
19	97	lemma	navigacijski sistem

N°	Hits	Atts	Hit
1	9483	lemma	podpirati
2	1301	lemma	podreti
3	703	lemma	podpreti
4	457	lemma	podajati
5	444	lemma	podpisati
6	396	lemma	podaljšati
7	338	lemma	podvojiti
8	216	lemma	podediti
9	125	lemma	podstavati
10	123	lemma	podeljevati
11	123	lemma	podedovati
12	119	lemma	podrezati
13	106	lemma	podreti
14	82	lemma	podleči
15	81	lemma	podkrepiti
16	79	lemma	podtalniti
17	66	lemma	podpisovati
18	62	lemma	podirati
19	61	lemma	podcenjevati

- kakšna sta bila iskalna izraza?
- čemu bi bili taki seznami korisnti?

Pojavnice in različnice

- angleško *token* in *type*
- pojavnica: kar se pojavi v besedilu (vsebina korpusa → konkordance)
- različnica: različne pojavnice v besedilu (besedišče korpusa → frekvenčni seznami)
- koliko pojavnic/različnic je v stavkih
  - Pri surovem krompirju se barva spremeni zaradi fermentov, pri kuhanem pa zaradi oksidacije.
  - Nova vpadnica za Novo mesto.
  - Gori na gori gori.
- kaj nam pove razmerje različnice/pojavnice?

Problemi z različnicami

- Kdaj sta dve pojavnici dejansko različni?
  - velike in male črke, npr. *Novo, novo, NOVO, NoVo*
  - naglasna znamenja: jêsen/jesén/jesen
  - razlika v besedni obliki, vendar ne v lemi, npr. *miza, mize, mizi,…*
  - razlika v lemi ali pomenu, vendar ne v besedni obliki, npr. *"Hotela je domov"* proti *"Hotela ni več v mestu"* *"Ure so bile pokvarjene"* proti *"Ure so bile poldne"*
- potreba po normalizaciji pojavnic

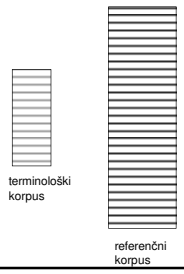
3. Statistične obdelave

- Z uporabo statističnih metod lahko odgovorimo na vprašanja, kot so:
  - katere besede najbolje opišejo neko besedilo?
  - katere besede najbolje razlikujejo dve besedili?
  - katere besede se najraje sopojavljajo z neko določeno besedo?
- večina teh metod primerja neko specifično besedišče s splošnim besediščem
- za vsako nalogo obstaja več konkurenčnih statističnih formul..

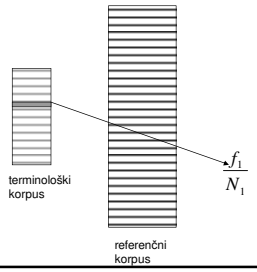
Ključne besede

- besede, ki najbolje opišejo neko besedilo (ali (pod)korpus)
- primerjamo število pojavitev vseh besed v našem besedilu s številom pojavitev teh besed v referenčnem korpusu
- število pojavitev delimo s številom besed v besedilu oz. ref. korpusu
- formula za "ključnost"
- opisno:
  neka beseda pokrije v v ref. korpusu 0.1% pojavnic,
  v besedilu pa 0.11%, ni ključna beseda,
  če pa 10%, pa je.

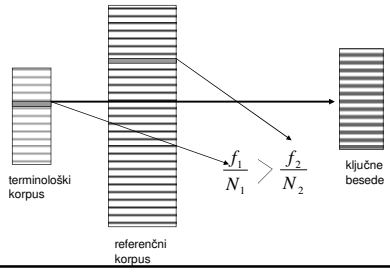
Luščenje ključnih besed



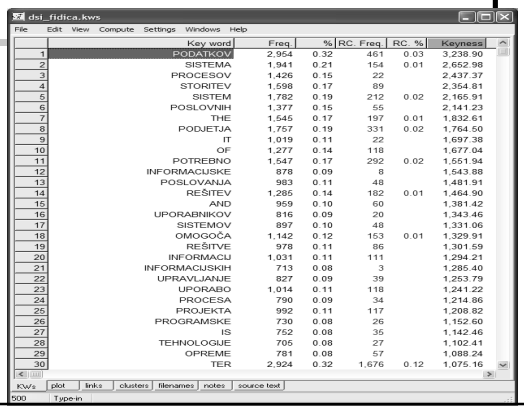
Luščenje ključnih besed



Luščenje ključnih besed



Primer iz Wordsmitha



dsi_fidica.kws

File Edit View Compute Settings Windows Help

	Key word	Freq.	%	RC. Freq.	RC. %	Keyness
1	PODATKOV	2,954	0.32	461	0.03	3,238.90
2	SISTEMA	1,941	0.21	154	0.01	2,652.98
3	PROCESOV	1,426	0.15	22		2,437.37
4	STORITEV	1,598	0.17	89		2,354.81
5	SISTEM	1,782	0.19	212	0.02	2,165.91
6	POSLOVNIH	1,377	0.15	56		2,141.23
7	THE	1,545	0.17	197	0.01	1,832.61
8	PODJETJA	1,757	0.19	331	0.02	1,764.50
9	IT	1,019	0.11	22		1,697.38
10	OF	1,277	0.14	118		1,677.04
11	POTREBNO	1,547	0.17	292	0.02	1,551.94
12	INFORMACIJSKE	878	0.09	8		1,543.88
13	POSLOVANJA	983	0.11	48		1,481.91
14	REŠITEV	1,285	0.14	182	0.01	1,464.90
15	AND	959	0.10	60		1,381.42
16	UPORABNIKOV	816	0.09	20		1,343.46
17	SISTEMOV	897	0.10	48		1,331.06
18	OMOGOČA	1,142	0.12	153	0.01	1,329.91
19	REŠITVE	978	0.11	86		1,301.59
20	INFORMACIJ	1,031	0.11	111		1,294.21
21	INFORMACIJSKIH	713	0.08	3		1,285.40
22	UPRAVLJANJE	827	0.09	39		1,253.79
23	UPORABO	1,014	0.11	118		1,241.22
24	PROCESA	790	0.09	34		1,214.86
25	PROJEKTA	992	0.11	117		1,208.82
26	PROGRAMSKE	730	0.08	26		1,152.60
27	IS	752	0.08	35		1,142.46
28	TEHNOLOGIJE	705	0.08	27		1,102.41
29	OPREME	781	0.08	57		1,088.24
30	TER	2,924	0.32	1,676	0.12	1,075.16

KWs plot links clusters filenames notes source text

500 Type-in

Luščenje “terminov”: TF-IDF

- iskanje podatkov (IR) – indeksiranje dokumentov
- namen: poiskati besede, ki naredijo dokument najbolj prepoznaven v množici in po katerih se najbolj razlikuje od vseh dokumentov v množici
- TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF

slovenski del JRC-Acquis / podkorpus besedil s področja jedrske energije

sevanju	0,19082	cepitve	0,05684
radiološkega	0,17864	nivoji	0,05684
dozimetrijo	0,17052	efektivno	0,05684
sivert	0,13804	medicine	0,05278
radionuklidov	0,13804	fuzije	0,05075
sevanja	0,13195	zaposlitvijo	0,04872
Dana	0,12992	termonuklearni	0,04872
Černobil	0,12180	študentov	0,04872
Izpostavljenost	0,12180	guvernerjev	0,04872
Jedrska	0,11368	prioritete	0,04872
dozo	0,09473	reaktorja	0,04872
prebivalstva	0,09256	jedrske	0,04872
sevanjem	0,08932	delodajalca	0,04669
ITER	0,08120	izpostavljenih	0,04601
Oddelkom	0,07308	ionizirajočemu	0,04466
inovativnosti	0,07308	ekvivalentno	0,04263
študente	0,07308	dosegljive	0,04060
izpostavljenosti	0,07308	ionizirajočega	0,04060
radioaktivne	0,06766	jedrskem	0,04060
SRS	0,06766	nuklearnih	0,04060
doza	0,06496	kontrolirana	0,04060
posameznike	0,06090	radiološki	0,04060
pooblaščenimi	0,05684		

Kolokacije

- statistično pogoste besedne zveze: nekatere besede družijo se rade
- idiomi, fraze, termini…
- več formul, ki primerjajo “naključno” porazdelitev sopojavitve besed z dejansko sopojavitvijo:
  MI, MI3, LL

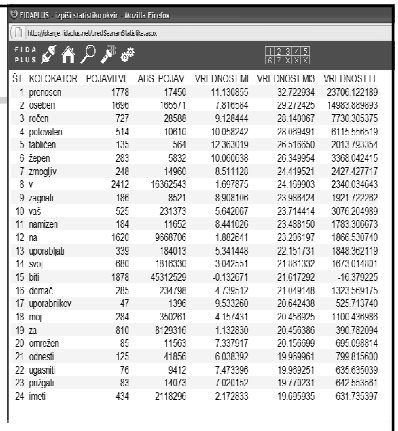
Naivni pristop

Query: IKORPUS; [word=".*"] [lemma="računalnik"]

N°	Hits	Atts	Hit
1	3400	lemma	oseben računalnik
2	3053	lemma	ročen računalnik
3	2983	lemma	v računalnik
4	2155	lemma	prenosen računalnik
5	1629	lemma	z računalnik
6	1439	lemma	. računalnik
7	1272	lemma	na računalnik
8	1219	lemma	biti računalnik
9	1031	lemma	žepen računalnik
10	890	lemma	namizen računalnik
11	708	lemma	za računalnik
12	494	lemma	svoj računalnik
13	418	lemma	drug računalnik
14	414	lemma	domač računalnik
15	408	lemma	omrežen računalnik
16	397	lemma	nov računalnik
17	388	lemma	ves računalnik
18	356	lemma	iz računalnik

"Kolokator + računalnik" v FidaPLUS

Vrednosti MI, MI3 in LL



ŠT	KOLOKATOR	POJAVITVI	ATTS POJAV	VREDNOST MI	VREDNOST MI3	VREDNOST LL
1	pronosen	1778	17450	11.130850	32.722994	23706.122189
2	oseben	1656	16501	7.810364	29.272425	14983.839696
3	ročen	727	26588	9.128444	20.140067	7730.305375
4	prikovljen	514	10610	10.093242	28.998491	6116.[illegible]
5	tabličen	135	561	12.363019	26.315850	2013.[illegible]
6	žepen	283	5832	10.000038	20.349954	3368.042415
7	zmogljiv	246	14960	8.511120	24.419521	2427.427717
8	v	2412	16302543	1.097870	24.109903	2340.034043
9	zapeljal	186	8621	8.908106	23.988424	1821.[illegible]
10	vaš	525	231373	5.642007	23.714414	3076.204989
11	namizen	164	11652	8.441020	23.488150	1783.300673
12	na	1620	9668706	1.862641	23.206197	1866.530740
13	uporabljali	339	184013	5.341446	22.151731	1848.362119
14	svoj	680	1818300	3.042551	21.051302	1672.014101
15	biti	1878	45312526	-0.132071	21.017202	-16.370225
16	domač	205	204098	4.739512	21.049148	1323.569175
17	uporabnikov	47	1396	9.533260	20.642438	525.713740
18	moj	284	360261	4.157451	20.245925	1100.449988
19	za	810	8129316	1.132830	20.450386	390.782094
20	omrežen	85	11563	7.337017	20.159369	695.608814
21	odnesti	125	41856	6.038369	19.999961	796.615600
22	ugasniti	76	9412	7.472396	19.993251	635.635039
23	pavšljali	83	14075	7.000150	19.977031	697.[illegible]
24	imeti	434	2118296	2.172833	19.695935	631.735397

Besedne skice

- Sketch Engine
- kombinacija iskanja po pravilih nad oznakami z iskanjem kolokatorjev

zgoščenka Fida PLUS 620m freq = 14595

a_modifier	3563	1.4	post_z	1342	8.4	is_obj4	2426	6.0
kompilacijski	40	53.47	naslov	587	62.42	izdati	839	62.56
priložen	99	51.81	glasba	166	44.15	posneti	267	47.86
nov	1226	48.64	posnetek	61	35.25	predstaviti	260	35.61
piratski	36	40.21	skladba	42	33.1	snemati	57	30.84
multimedijski	48	40.01	pesem	61	31.08	pripravljati	65	21.69
glasben	154	35.73				podariti	31	21.55
spremljajoč	35	34.0				kupiti	57	20.01
samostojen	61	29.54				prejeti	39	17.25
avtorski	38	27.85				predstavljati	33	13.24
računalniški	50	26.99				pripraviti	39	10.79

Poizvedovalni jeziki

- iskanje v korpusu
- po čem iščemo?
  - po besedilu
  - po jezikoslovnih oznakah
  - po metapodatkih (omejitve iskanja)
- kako iščemo
  - dobesedno
  - z mehkim ujemanjem (regularni izrazi)
  - enostavni oz. kompleksni izrazi
- konkretni poizvedovalni jeziki orodij se med seboj zelo razlikujejo ☹

Primer: JOS

Primer: FidaPLUS

Mehko ujemanje

- večina konkordančnikov dopušča mehko ujemanje:
  po končnici: mizerij*
  po predponi: *gle
  poljubno: *gled*ti
- mehko ujemanje je samo podmnožica t.i. regularnih izrazov, ki dopuščajo tudi bolj kompleksne iskalne pojme

Regularni izrazi

- uporabljajo jih konkordančniki, pa tudi tudi urejevalniki besedil in mnogo programskih jezikov (grep, awk, Perl, Ruby,…)
- regularni izraz prepozna (mogoče neskončno) množico nizov
- sestavljeni so iz literalov in operatorjev:
  literali: npr. *a,b,c,č,d,…,z,ž*
  operatorji: konkatenacija, disjunkcija, ponavljanje, združevanje

Osnovni primeri

- konkatenacija (implicitna):
  /**abc**/ prepozna {*abc}*
- disjunkcija:
  /**ab**|**bc**/ prepozna {*ab, bc}*
- ponavljanje:
  - ničkrat ali enkrat:
    /**ab?**/ prepozna {*a, ab*}
  - ničkrat ali večkrat:
    /**ab***/ prepozna {*a, ab, abb*, …}
  - enkrat ali večkrat:
    /**ab+**/ prepozna {*ab, abb*, abbb,…}
- združevanje:
  /**(ab?)|c** / prepozna {a, *ab, c*}

Razširitve sintakse

- katerikoli literal: "**.**"
  npr. /abc./ prepozna {*abca, abcb, abcc, …*}
- pogosta uporaba: "**.***"
  npr. /abc.*/ prepozna {*abc, abca, abcaa, abcb, …*}
  - dosti programov ".*" okrajša na "*"
- tudi: "**.+**" in "**.?**"
- ponavljanje: "**{n,m}**"
  npr. /**a{2,5}**/ prepozna {*aa, aaa, aaaa, aaaaa*}

Razširitve sintakse

- skupine literalov: "[…]"
  npr. / [fgm]iga/ prepozna {figa, giga, miga}
- negirana množica literalov [^…]
  npr. /abc[^def]ghi/ prepozna {*abcgghi, abchghi, abcighi, …, abczghi, abcžghi*}

Primeri za iKorpus

- miza, miz., miz.?, miz.*
- miz[a,e,i,o], miz(a|e|i|o|ama|ah|ami)
- .*pisati, …pisati
- .*gled.*, pod.*, .*anje
- [aeiou]+

Naloge iz regularnih izrazov

Napišite naslednje iskalne pogoje:

- besede, ki se začnejo na “miš”
- besede, ki vsebujejo “miš”
- besede, ki vsebujejo najmanj tri a-je
- sedanjiške oblike glagola “delati”
- besede, ki vsebujejo najmanj 2 “lj”
- besede, ki vsebujejo dva zaporedna šumnika
- kratice iz najmanj treh velikih črk

Vendar..

- skoraj vsako orodje ima rahlo različno sintakso regularnih izrazov
- vsi ne podpirajo vseh predstavljenih operatorjev
- nekateri jih pa podpirajo še bistveno več

Pa zaključimo

- kaj delajo konkordančniki
- kaj so konkordance in frekvenčni seznami
- pojavnice in različnice
- in še nekaj statističnih metod:
  - ključnost in TF-IDF
  - kolokacije
- regularni izrazi