Identifikacija spletno specifičnih kolokacij pogostega besedišča

Senja Pollak

Inštitut Jožef Stefan

Slovenščina na spletu in v novih medijih, Ljubljana, 26. 11. 2015

Kolokacije

- Kolokacije: tipične sopojavitve besed
- Pari besed vs. baza in kolokator
- Frekvenčni vs. frazeološki pristopi
- Luščenje in preučevanje kolokacij za različne aplikacije: (leksikografija, terminografija, strojno prevajanje, generiranje naravnega jezika, klasifikacija dokumentov, poučevanje tujega jezika, luščenje informacij, analiza sentimenta)

Spletne uporabniške vsebine

- Korpus JANES: tviti, forumi, blogi, komentarji
- Preučevanje jezika spletnih uporabniških vsebin
  - ortografija (jz, cist, nevem, kwaaa)
  - skladnja (sva šle, bom bil)
  - leksikalna raven (neologizmi, pomenski premiki, aktualne tematike)
    - *všeček, virantovanje*
    - *sledilec, hud (dober), stisniti (datoteko)*
    - *rabiti (potrebovati)*
    - preučevanje na ravni enobesednih in večbesednih izrazov (kolokacije)

Kolokacije v SUV

Leme:

- splošno slov. besedišče (dan, otrok)
- spletno pogostejše besedišče (tvit, blog)
- spletno specifično besedišče (tajmlajn, tviteraš)

Kolokacije v SUV

Leme:

- splošno slov. besedišče (dan, otrok)

  (splošne leme, nove kolokacije, njihovi kolokatorji, pomenski premiki)
- spletno pogostejše besedišče (tvit, blog)
- spletno specifično besedišče (tajmlajn, tviteraš)

avto, beseda, cerkev, človek, dogodek, energija, fant, glava, hiša, igralec, jezik, knjiga, ljubezen, mati, namen, oče, pesem, roka, sistem, šola, telo, ura, vas, zgodba, želja

Obdobja vs. Janes

- samo luščenje kolokatorjev splošnega besedišča
- izdelan API za hitri izvoz kolokacijskih seznamov
- 150 lem
- specifično = ekskluzivno
- okno -1 (+1)
- ...

Cilj - specifično

- korpusa *S* in *R* (v istem jeziku)
- izbor lem *L*
- za vsako lemo *l* iz *L*: najdi K_l (<k(olokator), l(ema)>), specifični za en korpus
- cilj: seta kolokacij K_{LS} (in K_{LR}) <k,l> specifična za S in R

Cilj - specifično

- korpusa *Janes* in *Kres* (v istem jeziku)
- izbor lem *L*
- za vsako lemo *l* iz *L*: najdi K_l (<k(olokator), l(ema)>), specifični za en korpus
- cilj: seta kolokacij K_{LS} (in K_{LR}) <k,l> specifična za S in R

Cilj - specifično

- korpusa *Janes* in *Kres* (v slovenščini)
- izbor lem *L*
- za vsako lemo *l* iz *L*: najdi K_l (<k(olokator), l(ema)>), specifični za en korpus
- cilj: seta kolokacij K_{LS} (in K_{LR}) <k,l> specifična za S in R

Cilj - specifično

- korpusa *Janes* in *Kres* (v slovenščini)
- izbor lem *L* (*150 "splošnih", najpogostejših lem*)
- za vsako lemo *l* iz *L*: najdi K_l (<k(olokator), l(ema)>), specifični za en korpus
- cilj: seta kolokacij K_{LS} (in K_{LR}) <k,l> specifična za S in R

Cilj - specifično

- korpusa *Janes* in *Kres* (v slovenščini)
- izbor lem *L* (*150 "splošnih", najpogostejših lem*)
- za vsako lemo *l* iz *L*: najdi K_l (<k(olokator), l(ema)>), specifični(=ekskluz.) za *Janes*
- cilj: set kolokacij K_{LS}, specifičnen(=ekskluz.) za *Janes*

Metoda

- funkcija *kolokacije* orodja SketchEngine

JANES

Metoda

- funkcija *kolokacije* orodja SketchEngine: slovenščina
- 150 lem
- izvoz kolokacij Kres, Janes: API
- okno -1 (+1), logDice (Rychlý): leks. kolok.
- frekv. k: 5, n: 3 (n: 10)

	Freq	logDice
p | N knjižen	67	8.100
p | N tečaj	79	7.758
p | N profesorica	38	7.490
p | N učiteljica	36	7.134
p | N raba	40	7.129
p | N pravilen	58	6.956
p | N polomljen	24	6.829
p | N pogovoren	24	6.817
p | N znanje	65	6.579
p | N učenje	31	6.557
p | N zboren	16	6.432
p | N pouk	20	6.367
p | N učiti	51	6.323
p | N učitelj	27	6.206
p | N profesor	23	6.023
p | N obvladati	19	5.613
p | N naučiti	29	5.442
p | N translate	8	5.381
p | N poučevanje	7	5.175
p | N inštrukcija	7	5.158
p | N v	2,397	4.920
p | N brezhiben	6	4.866
p | N študirati	9	4.806
p | N uporaba	21	4.734
p | N poznavanje	6	4.652
p | N lektorat	4	4.446
p | N spakedran	4	4.443

Družina

Kolokacije (lema: *družina*)	
Kres	**Janes**
član družine	enostarševska družina
ribiška družina	član družine
kraljeva družina	primarna družina
lovska družina	kraljeva družina
ustvariti družino	mlada družina
rejniška družina	ogrožena družina
mlada družina	cela družina
kmečka družina	ustvariti družino
njegova družina	(*n*) članska družina
svoja družina	romska družina

Družina

Kolokacije (lema: *družina*)	
Kres	**Janes**
član družine	enostarševska družina
ribiška družina	član družine
kraljeva družina	primarna družina
lovska družina	kraljeva družina
ustvariti družino	mlada družina
rejniška družina	ogrožena družina
mlada družina	cela družina
kmečka družina	ustvariti družino
njegova družina	(*n*) članska družina
svoja družina	romska družina

Družina

Kolokacije (lema: *družina*)	
Kres	**Janes**
član družine	enostarševska družina
ribiška družina	član družine
kraljeva družina	primarna družina
lovska družina	kraljeva družina
ustvariti družino	mlada družina
rejniška družina	ogrožena družina
mlada družina	cela družina
kmečka družina	ustvariti družino
njegova družina	(*n*) članska družina
svoja družina	romska družina

Janes:8
199, 1,23/M

Kres:5,7
63; 0,52/M

Metoda

- avtomatska primerjava seznamov
- ekskluzivnost (samo na seznamu Janes, i.e. Kres < 3)
- izvoz kolokacij Kres, Janes: API

Nastavitve:

  - min logDice
  - min frek. 10
  - < razpr. (“idf”): odst. lem, ki imajo kolok. na seznamu oz. št. kolokac. seznamov s kolok./kolok. seznami vseh lem

Metoda

- vrednost *razp.* kot filter
- splošne funkcijske besede*:* (*razp > 0, 9: a, ampak, brez, dober, en, glede, isti, kak, kako, kakšen, moj, morati, nek, nekaj, niti, njihov, noben, oz., reči, svoj, tak, tisti, torej, tvoj, vaš, veliko, vsak, zaradi.*)
- nove rabe jezika, na ravni kolokatorja:
    - pogoste besede z opuščenimi strešicami: *vec*
    - nestand. okrajšave: *slo*,
    - pogosti tematski izrazi (*političen, janšev*)
    - nest. raba (*rabiti: rabiti prostor, hrano*)
    - pom. premiki kolokatorjev (*hud*: *huda slika, hud igralec*)
    - razlike med lematizaroji dveh korpusov (*edini* vs. *edin*)., itd.

Metoda

- vrednost *razp.* kot filter
- splošne funkcijske besede*:* (*razp > 0, 9: a, ampak, brez, dober, en, glede, isti, kak, kako, kakšen, moj, morati, nek, nekaj, niti, njihov, noben, oz., reči, svoj, tak, tisti, torej, tvoj, vaš, veliko, vsak, zaradi.*)
- nove rabe jezika, na ravni kolokatorja:
    - pogoste besede z opuščenimi strešicami: *vec*
    - nestand. okrajšave: *slo,*
    - pogosti tematski izrazi (*političen, janšev*)
    - nest. raba (*rabiti: rabiti prostor, hrano*)
    - pom. premiki (*hud*: *huda slika, hud igralec*)
    - razlike med orodji za predproc. dveh korpusov (*edini* vs. *edin*).

ZANIMIVE BAZE (leme)

Izluščeni kandidati

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Izluščeni kandidati

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Izluščeni kandidati

Vsi pogoji:

- *homoseksualna družina*
- *narcisistična družina*
- *ožja družina*
- *razdreti družino*
- *Festival družin*
- (<razpasti, družina>, *razpadla družina* (oba korpusa))

Izluščeni kandidati

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Izluščeni kandidati

Vsi pogoji:

- *homoseksualna družina*
- *narcisistična družina*
- *ožja družina*
- *razdreti družino*
- *Festival družin*
- (<razpasti, družina>, *razpadla družina* (oba korpusa))

Brez filtra *razp.*:

- *Janševa družina*
- *Praznična Družina*
- *<celje, družina> (cela družina)*

Rezultati luščenja

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Izluščeni kandidati

Le logDice>3 (ni omejitve frek.):

- vse prej naštete +
- *raznospolna družina, partnerjeva družina*
- *črkovna družina* (grafika), *modelna družina* (avtomobilizem), *imenska družina*
- *razdirati družino*
- *časopis Družina*, *priloga Družina*, *Prehrana družine*
- *grožnja družini*

Rezultati luščenja

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Izluščeni kandidati

Brez omejitev:

- vse prej naštete +
- *homoseksualna, oligarhična*, *čefurska*, *gejevska*
- zelo splošni kolokatorji *super* in *ok*
- funkc. besede, ki z lemo ne tvorijo sintaktičnih in semantičnih kolokacij:. *potem*, *kdo*, *pač* in *sploh*

Izluščeni kandidati

Lema	Kolokatorji
država	vitek, zadolžiti, rušiti, pravično, ugrabiti, koruptiven, zavoziti, gnili, pokrasti, ugrabljen, zajebati
jezik	muca, eksotičen, tuja, ang, venetski, šparati
mama	foto, yo, platišče, mreža, odkar, doječ, vreme, kurac, feltna
moški	hetera, napasti, feminizacija, kastracija
zgodovina	servisen, preverljiv, smetisce, spisati, retuširanje, brisanje, odpasti

Izluščeni kandidati

- neformalni kolokatorji *nabaviti avto, nategovati ljudi, frej dan*

- aktualne tematike, novi pojmi
  - *zajebati, zavoziti, pokrasti državo; feminizacija moških, kastracija moških*
  - *nepremičninski zakon, privatizacija vode, varčevalna politika*
  - *transspolna oseba*
  - lastna imena: *Depala vas;* Čezvesoljska *zombi cerkev* blaženega zvonjenja;
  - tujejezične prvina: *gratis knjiga, rimejk filma*
  - *Facebook skupina, startup podjetje, wifi točka*

Izluščeni kandidati

- novi pomeni: *brisanje zgodovine*

  Telefon na tiho, SMS-ji, *brisanje zgodovine* na računalniku: jasni indici, da te vara ...

  Brisanje zgodovine na Googlu je zadnji teden ali dva postal svetovni hit.

  ... and FB doesn't make it easy. A pozna kdo kkšno hitrejšo rešitev za *brisanje zgodovine* ?

- terminologija: *evklidski prostor, rizičen odnos*

- idiomatika
  - *muca jezik papala* (oz. *popapala, popapcala), muca jezik papne*
  - *rana ura*... , pitaj boga (pitati*), jebo vas*

Napake predprocesiranja

- izpuščanje strešic:
  - *smetisce zgodovine, vcerajsnja tekma, koncna cena, soncen dan*
- napačna lematizacija:
  - *hetera moški, tuja jezik*
  - *anything glas*
  - <foto, mama> *Big foot mama*
  - <kurac, mama> *poln kurac mam* oz. *koji kurac mamo*
- nerazpršenost po korpusu:
  - *Pediater je odvetnik otroka* je (300 pojavitev)

Napake predprocesiranja

- izpuščanje strešic: rediakritizacija
  - *smetisce zgodovine, vcerajsnja tekma, koncna cena, soncen dan*
- napačna lematizacija: izboljšanje lematizatorjev, problematična normalizacija
  - *hetera moški, tuja jezik*
  - *anything glas*
  - <foto, mama *Big foot mama*
  - <kurac, mama> *poln kurac mam* oz. *koji kurac mamo*
- nerazpršenost po korpusu: deduplikacija, račun razp. med uporabniki (korak po luščenju)
  - *Pediater je odvetnik otroka je (300 pojavitev)*

Kolokacije SUV

- splošne leme

- specifične leme:
  - drag/ljub/znani/slovenski/pravi tviteraš
  - nova/huda/lepa/dobra profilka
  - meja/število všečkov
  - prava/velika bizarka nova/huda/lepa/dobra profilka

Zaključki

Raziskava:

- luščenje kolokacij “splošnega” besedišča
- aktualne tem., novi pojmi, pomenski premiki
- tujejez. prvine, lastna imena
- napake luščenja

Nadalnje delo:

- primerjava besednih skic (API!), kolokacije spletnega besedišča
- relevantnost kandidatov
  - za vključevanje v slovarje: slovar spletne slovenščine, kolokacijki slovar, SSSJ (Pollak, Arhar Holdt, Gantar)
  - kategorizacija (Pollak, Arhar Holdt, Gantar)
  - aktualne tematike, analiza diskurza (Škrjanec, Sobočan, Pollak)
  - frazeologija