

A R A N E A : ÑÀÌÅÉÑÒÂÎ ÌÈËËÈÀÐÄÍÛÔ ÂÅÁ-ÊÐÏÓÑÎÂ

© Áëàäèì]ð Áåíéî (Vladimír Benko). Ñëíâàêèÿ, Áðàòèñëàâà. Èíñòèòóò
ÿçûéñïçíàíéÿ èi. Ëþäíâèòà Øðóðà Ñëíâàöéé
Áêàäâàìè è àôóé, þÍÁÑÉ
êàðâåðà ííñäýçû-ñíé è iàæéóëüòðííé èññóíèéàöèè Óíèâåðñèòåòà èi. Bìà
Èññâññâ ã Áðàòèñëàâà

Ëåéöèè iðåäñòàâýò iðåâèò
ñàìåéñòâà âåá-êîðíóñîâ Aranea äëÿ ýçûéñâ èññóíèéüçâàííûô è/èéè
iðåíâàííûô â ñëíâàöéé òíèâåðñèòåòàð, êîòîðûâ iðåäíàçíà-åíû äëÿ iðåíâàâàíéÿ
ðééëéëé-åññéè ò ððâíñëàðíèé-åññéè ðåääàðòíâ, à òàéæå è äëÿ èéíââèñòé-åññéè
èññëäâàíéé.

Âåá-êîðíóñ iðåäñòàâéÿâò ñíáíé iññáûé
âèà èéíââèñòé-åññéâí êîðíóñà, êîòîðûâ ñíçääí iðòâí iññòâíâííé çàäðóçêè òåéñòâ
èç èíòâðíâòà iðè iññùè àâòíâòèçéòíâííûô iðòâðåöð, êîòîðûâ íà èåðò ïðåäââéýþò
ÿçûé è êíæéðíâéò ìòâåëüíûô âåá ñòðâíéò, óääëéýþò øàáëíû, ýéâàíòû iàâæâàöèè,
ñññéè è ðåéëàò (ò.í. boilerplate),
iññóùâñòâéýþò ððâíñòðíàöèþ íà òâéñò, ðèëüòðàöèþ, iðòâëèçàöèþ è ååäóïëèéàöèþ
iññó-ñíûô ãíéóíâòíâ, êîòîðûâ çàðâí iññí íáðâáâòàðû ððâæëðééííûè
èíñòðíâíâòíâè êîðíóñíé èéíââèñòéè (òíèâåéçàöèÿ, ièðòññèíòâéñè-åññéàÿ è
ñéíâèñè-åññéay áííòâöèÿ) è áíâåðèòû â iññéñâóþ êîðíóñíóþ ñèñòâí. Níçääíèå
âåá-êîðíóñà íà òíèëüí iàíííâí ååðâåâéå, ií iðåæäå åññâí åäí ðâçìâð iññâòå ãûòü
åâæå íà iññýâíé åíëüøå òðåäâèëøííûô êîðíóñà.

Iðè ñíçääíèè âñåð êîðíóñâ áûëà
iññéñâíâí iññéñâàâéÿ iàðíâíèé è íàíâð iññòðâííûô èíñòðóíâíòíâ: SpiderLing, Onion, Unitok [1] è TreeTagger [2]. Â èâ-å
iññéñâíâí ñèñòâí ñíññéüçóâðñý NoSketchEngine [3] (open-source) èéè SketchEngine [4] (ièàðíâáÿ). O êîðíóñâ iàçâàíéÿ íà «íâéð
(èàðèéññéí) ýçûéâå íáçíà-åþùâí ýçûé è ôíæå ðâçìâð êîðíóñà, íàíðèíâð AraneumAnglicumMinus,
AraneumRussicumMaius, è ò. í. Å íññòðíâå ååðâí ñââñòâéñòâí ñíñââðæèò
18 êîðíóñâ íà 14 ýçûéâå ã ååñòðâåðàò è âñâ êîðíóñû ñâñòâííû á ååññéèòíí
ðâææíâ íà êîðíóñíí iññòðâéå iðåâèòà [5].

Â ìòëè-èå ìò åû-èñëèòåëüíûô èéíââèñòâ,
êîòîðûâ íáðâåâòûâåþò êîðíóñûâ åàííûâ á iâéâòíí ðåâæèíâ, iññòâëüíûâ iññéüçâàòåéè
êîðíóñâ íà ñíññéüçâàòåéè ñíññéüçâàòåéè ñíññéüçâàòåéè ñíññéüçâàòåéè
ñéíâèñè-åññéè ñòðóðéòð ðâñâð-àðâííûô á åèäå êííèíðâæññâ, ÷àññòðíòíû ñíññéñâ è
iññóðééåé íà yéðâíá. Ò-èòûââéÿ ðâçìâðû ñâñâðâæññâû ëîðíóñâ ñòâíâð ÿññíû, ÷òí
ýóðâåâðéâíññòû è óäíâíññòû iññéñââéÿ ÿâëëýðñý iññâí ãâææíû òâéòðíí åëÿ âñýéé
ðââñòû ñ êîðíóñâ.

Ñââñâíâð iññâýùâí iðâéòèéå ðââíòû ñ
ñàìåéñòâí êîðíóñûô iññéñââí ñèñòâí NoSketch
Engine [6] è Sketch Engine [7],
iññéñââæàòè è ñââñâí ëó-øèí à ièðâ èíñòðóíâíòàí åëÿ ðââíòû ñí «ñââððâíëüøèíè»
êîðíóñâ è (ðâçìâðíí á åâñýðéè iññééâðâíí ñíññéñâ). Íâá ñèñòâíû áûëè ñíçääíû á Ëâáíðâòíðè è íâðâáâòíðè è åñòâñòâåâíí ýçûéâ
Óàééðüðâðàò è íóíðâòèéè Óíèâåðñèòåòà èi. Iàñâðèéâ à Áðíí, iðè-åí ôóíéëè
åâññéèòíí (open-source) ñèñòâíû NoSketch
Engine ýâëëýþñý iññâíæññâíí ôóíéëè Sketch Engine è åéëþ-åþþò á iññíí
íáúâíâ ðââíòû ñí iññéñââé (Word List)
è êííèíðââíññâðíí, ò. á. iññéñâ ñíññéñâðíâ, èâííâ, ñí-åðâòâíè è iññòññéíòâéñè-åññéè
iâðâéå à ðâçíû ëîðâíâðéëÿ íà ýçûéâ CQL
(Corpus Query Language), è ôíæå

âû=èñëåíèå èïëëìèàöèé íà áàçå ñòàòèñòè÷åñêèõ íàð ñí÷åòàåíñòè (T-score, MI, MI3, log likelihood, min. sensitivity è logDice). Ñèñòàìà ðàáîòàåò â ðåæèìå ñâðååð/ëëèåíò, ãäå íà ñâðååðå õðàíýòñý âñâ ääííûå è ïñóùâñòâëýþòñý ïèñêíâûå íïåðàöèè è ïëüçîàòåëü ðàáîòàåò ñ êëèåíòí ÷åðâç ååå-ëíòåðôåéñ íðè ìííè ñòàíâåðòííñ áðàóçåðà.

Íëàòíàÿ ñèñòàìà Sketch Engine ñíäåðæèò êðììå âñåð ôóíêöèé NoSketch Engine òðè ñóùâñòâåíûõ ðàñøèðåíèÿ – èïëëìèå ïðîèèè (ñéåò=è) ïñòðàííûå íà áàçå íïëüçîàòåëüñêè ñéåò=÷åðàíàòèè, àèñòðèåóòèåíûé òåçåóðóñ è ôóíêòþ ñðåâåíàíèÿ ñéåò=åé àëý ååóð èåéñè÷åñêèò ååéíèö. Âñâ ýòè ôóíêöèè ðàáîòàþò ñ ääííûìè åû+èñëåííûìè çàðåíåå, ÷òî ååëàåò ñèñòàíò í-åíû áûñòðí è óåííûé. Ñèñòàìà ïðåäñòâåëýåðñý â åèäå ñâðåèñà (íäíèñèè) íà ñâðååðåò ëííàíèè Lexical Computing, íà êîòòðûõ ñðåíýòñý êïðíóñû áîéåå ÷åí íà 80 ýçûèåõ, åéëþ÷àÿ 15-ìèëëèàðäíûé êïðíóñ ðóññéíñí ýçûèå.

Êíêîðääíñ ôîðìàòà KWIC äëý ñëíâîòòíðíû «íâîñèáèðñêèé»

Íðàâñòòíðííûå
èïëëìèàòû ñëíâîòòíðíû «íâîñèáèðñêèé»

Äèñòðèåóòèåíûé òåçåâð äëý èåííû
«íâîñèáèðñêèé»

xàñòòíàÿ äèñòðèåóöèÿ èåííû
«íâîñèáèðñêèé» ïí TLD

Íðâðàíà
êóðñà

«A r a n e a : Ñâìâéñòâî ìèëëèàðäíûõ
ååå-ëíðíóñíâ»

(10 ÷àñâ ëåéöèííûõ çàíýòèé è íðàêòèêóííâ,
16-20 ííýáðý 2015 å.)

Ëåéöèÿ 1. Âååäåíèå. Ëèíâèñòè÷åñêèé êïðíóñ êàâ èñòòí÷íè
èíòòíàòèè í ýçûèå

Íñíâíûå ííýòèÿ: âèäû ýëåéòðííûõ
èïëëåéöèé òåéñòâ

Èñòòðèÿ ñïçäàíèÿ êïðíóñíâ, ååíâðàöèè, íðèíâíåíèÿ

Iðiñâêòû ìàöèñìàëüíûõ êiðiññâ

Êiðiññâiay eëíññâèñòèèà êàê ìåòiä èëë ññiñâiay âåòêà
ýçûêñçìàë

Iðiññâiay ìåòiä ëiðiññâiay eëíññâèñòèèà â ñeíññâiay è
ëæàôññâiay èññëåññâiay è ýçûêà

Iñññâiay âåá-êiðiññâ, ññiñâiññòè è ìòëë÷ëy ìò
òðâæëñññâiûõ êiðiññâ

Éâêëy 2 Aranea – Ññiññòâi ìëëëëàðäiûõ
âåá-êiðiññâ

Iñññâiûõ ðåññâiay iðiññâòà Aranea: ýçûêè, ðàçìåðû è åàðèàíûõ, ìàçâàíay

Èíññòâiay ìåòû äëy iáðâáòèè: êðâóëëíâ, ññiññâiay ýçûêà, óääëåíay
øàëíññâ, äñññëëëàëy, òîëåíèçâëy, ññiññèòàëñè÷åññây ðàçìåðâà, óièòëëàëy
òåññâòâiâ

Ióáëëëàëy è èññëüçìâàíay èiðiññâ: êiðiññâiûõ ìåññâæåðû

Ôiðiññâòà ëiðiññâ Aranea [8]:
àòðèáòû è ñòðóëòóðû

Iðaëòëëò:

A. Ðàáñòà
ñ êiðiññâiay ññiññâiay ñeññòâiay (No)SketchEngine

Iðaëòëëò

1. ññiññâiay ñeññòâiay, ññiññâiay è
ññiññâiay ñeññòâiay

Ðåæëèiû èçìáðâæåíay, íàñòðiéè

Ôëëüòðû, êiðiññâ

xàññòðû ñeññòâiay

Òåññâòû (tagsets), ññiññâiay ñeññòâiay

2. ßçûê çàëðíñíâ CQL

Ðåâóöëýðíûâ âúðàæåíèý

Íõèìåíáíèå CQL:

ííèñê ñeíòàêñè÷åñêèõ ñòðóêòóð

Êíëëíéàöëè, íåðû àññíöèàòèâññòè

Âû÷èñëåíèå êíëëíéàöëííûõ êàíäèàòâíâ

B. Ðàáîòà
ñ êíðíóñííé ííèñêíâíé ñeñòåííé SketchEngine

Íõàêòèêóí 1: Êíðíóñíûé íåíâäæåð Sketch Enginå (<https://www.sketchengine.co.uk/>)

Êíëëíéàöëííûâ íðîòèëè (ñêåò÷è)

Ñêåò÷-åðàíàòèëè: ñeíòàêñè÷åñêèé è êíëëíéàöëííûé ííäõíä

Ñêeò÷-åðàíàòèëè äëÿ êíðíóñíâ Aranea

Ñõàâíáíèå ñêåò÷åé (Sketch-diff)
è äeñòðèáóòèâíûé òåçàâð

Ãâóöçû÷íûâ ñêåò÷è

Íõàêòèêóí 2: Ðåñóðñû è éíòðóìåíòû ñàéòà Sketch Engine

Êíðíóñû ñâíâéñòâà TenTen

Íàðàëëåëüíûâ êíðíóñû

Èíñòðóìåíòû äëÿ ñïçäàíèÿ ííëüçâàòåëüñêèõ êíðíóñíâ: Corpus Architect è WebBootCaT

Ýêñòðàêëëè òåðìèííëè

Â ðâíêàõ

Íõàêòèêóíâ ñòóäåíòû íçíàêííýòñý ñ ðàáîòíé ñ íáâèíè ñeñòåíàíè è ííëó÷àò
íâíâðàíè÷åíûé äíñòóí è ííðíóñàí íà êíðíóñííí ííðòàëä íðíâéòà Aranea (<http://ucts.uniba.sk/>) è âðåíåííûé 3-õ
íâñý÷íûé áâñíëàòíûé àééàòíò äëÿ Sketch Engine.

Webography

<https://savba.academia.edu/VladimirBenko>

http://ucts.uniba.sk/aranea_about/

[1]

<http://corpus.tools/>

[2]

<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

[3]

<http://nlp.fi.muni.cz/trac/noske>

[4]

<https://www.sketchengine.co.uk/>

[5]

http://ucts.uniba.sk/aranea_about/

[6] <http://nlp.fi.muni.cz/trac/noske>

[7] <https://www.sketchengine.co.uk/>

[8] Ñàéò ïõîâéòà Aranea: http://ucts.uniba.sk/aranea_about/

Êîðïóñíûå ïîðòàëû: Aranea: <http://ucts.uniba.sk/>
; <http://ella.juls.savba.sk/>

