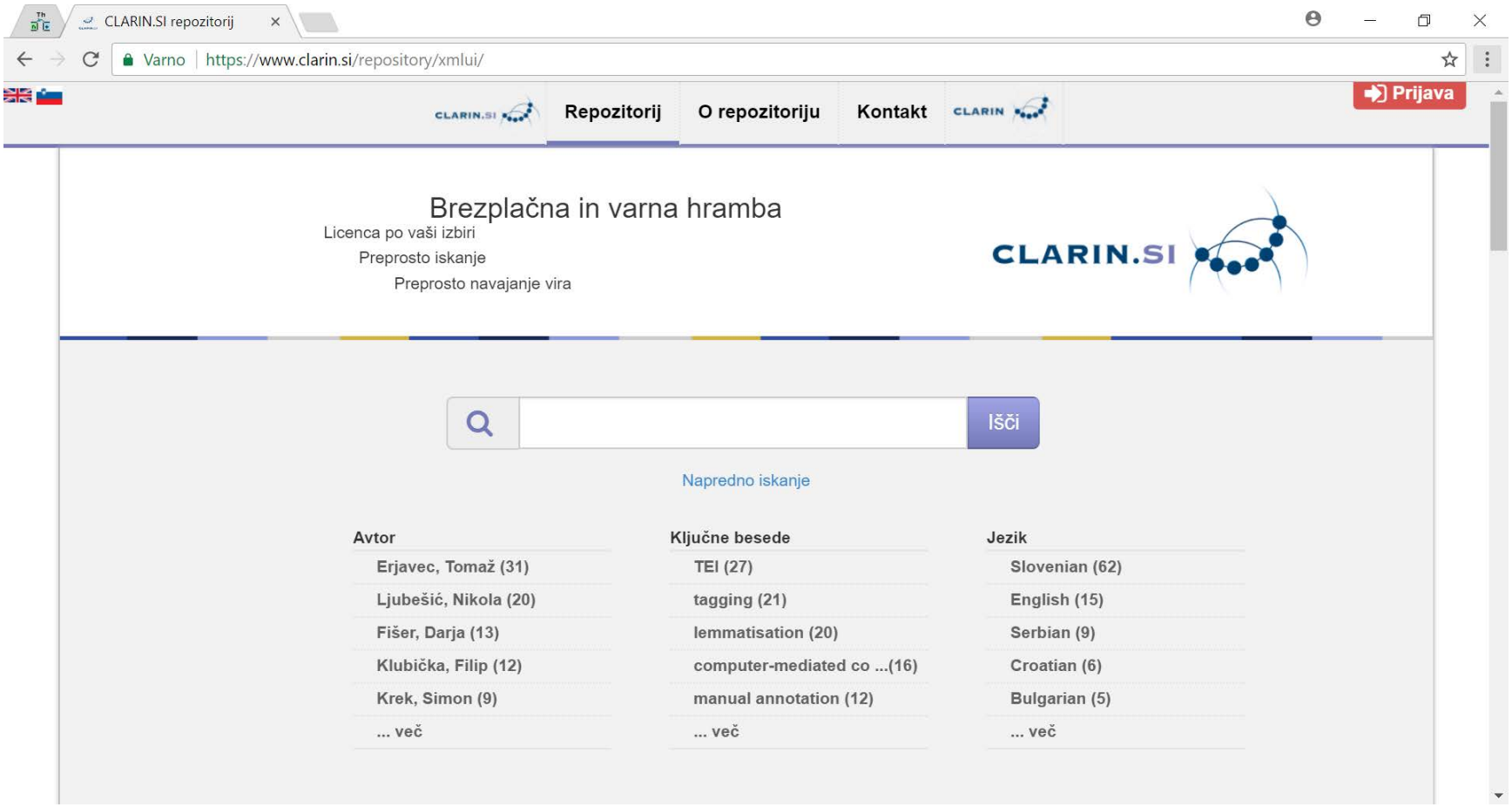
Repozitorij CLARIN.SI: uporaba

Tomaž Erjavec

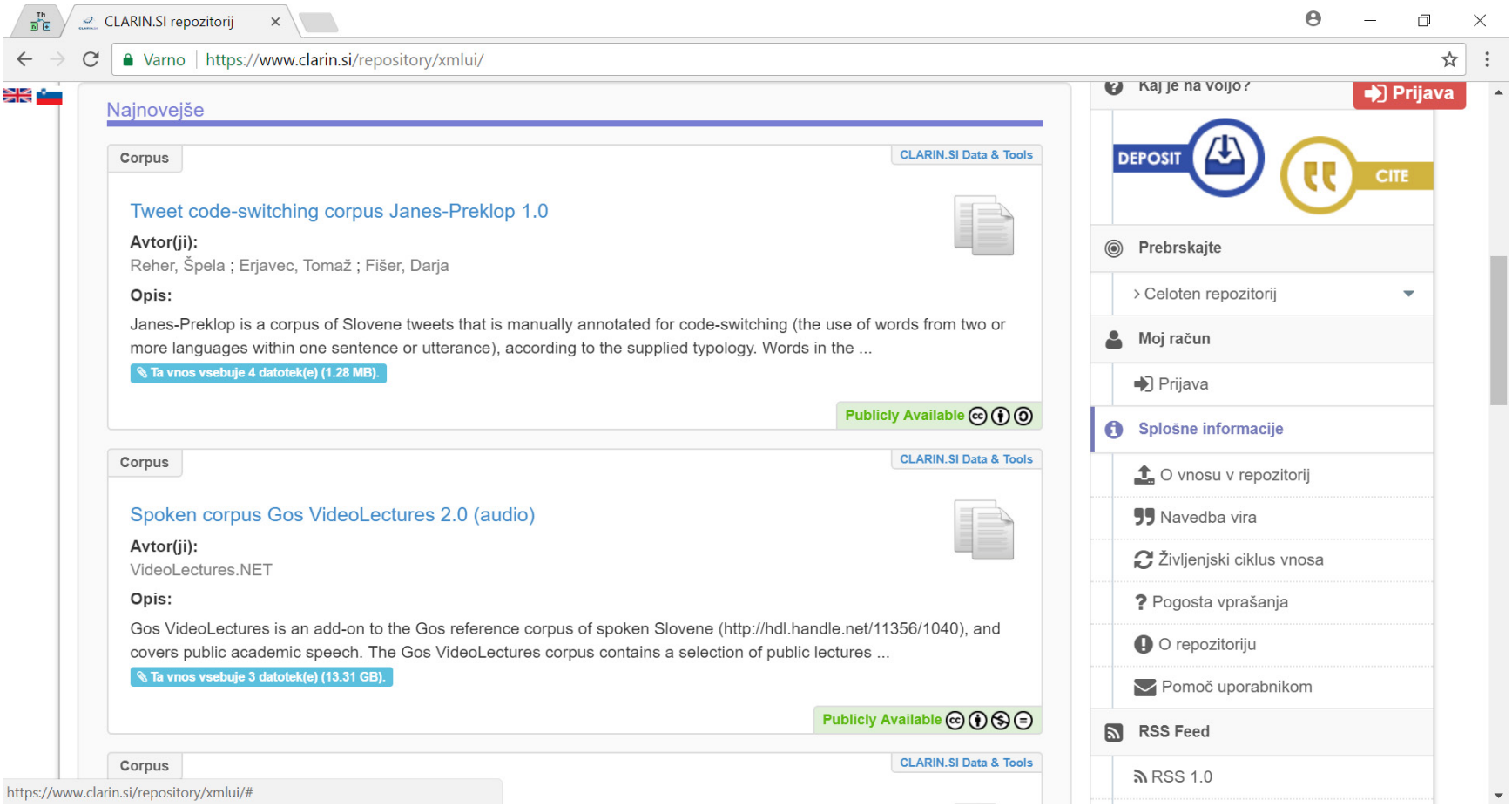
Odsek za tehnologije znanja, Institut „Jožef Stefan“

FRI, 10. 11. 2017

Vstopna stran

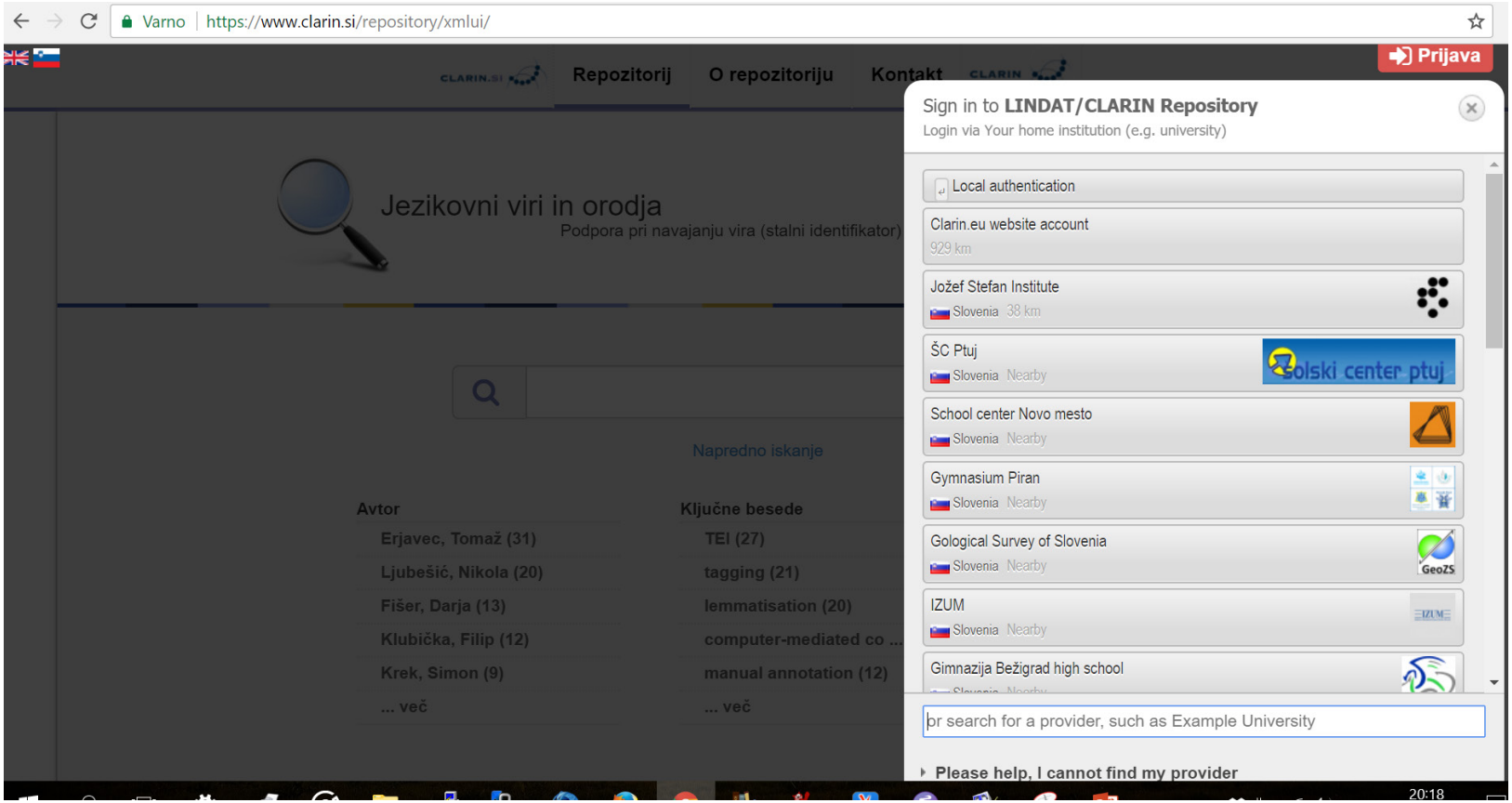


Vstopna stran



Prijava

- potrebna za vnos novega vira, dostop do zaščitenih virov in za uporabo naprednih funkcij



Kako najti zanimive vire

- brskanje po jeziku, vrsti vira, ključnih besedah, avtorju itd.
- mehko iskanje
- ni težko najti, saj je trenutno samo 80 virov

Anatomija vnosa

Emoji Sentiment Ranking 1.0

Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko: BIBTEX CMDI

Kralj Novak, Petra; Smailović, Jasmina; Sluban, Borut and Mozetič, Igor, 2015, *Emoji Sentiment Ranking 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1048.

Delite:

CLARIN.SI Data & Tools

Avtorji	Kralj Novak, Petra ; Smailović, Jasmina ; Sluban, Borut ; Mozetič, Igor
Demo URL	http://kt.ijs.si/data/Emoji_sentiment_ranking/
Referenced by	http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296
Datum izdaje	2015-09-14
Zvrst	lexicalConceptualResource
Velikost	751 entries
Jezik(i)	Albanian , Bulgarian , English , German , Hungarian , Polish , Portuguese , Russian , Serbo-Croatian , Slovak , Slovenian , Spanish , Swedish
Opis	A lexicon of 751 emoji characters with automatically assigned sentiment. The sentiment is computed from 70,000 tweets, labeled by 83 human annotators in 13 European languages. The process and analysis of emoji sentiment ranking is described in the paper: Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. PLoS ONE 10(12): e0144296. doi:10.1371/journal.pone.0144296

Anatomija vnosa 2

EC

Koda projeka: 640772

Ime projekta: DOLFINS

Ključe besede sentiment classification | emojis | unicode

Zbirke CLARIN.SI data & tools

Pokaži vse podatke o v…

Datoteke v tem vnosu — Prevzem vseh datotek v viru (93.95 KB)

Ta vnos je Publicly Available z licenco:
Creative Commons - Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Ime ESR_v1.0_format.txt
Velikost 783 bajtov
Format Text file
Opis Format of the tables.
Prevzem datoteke | Predogled

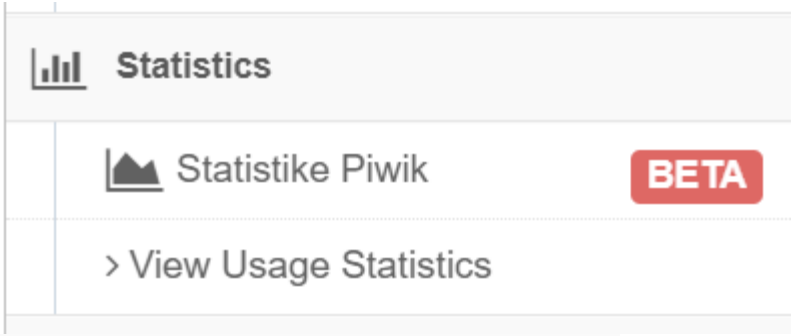
Ime Emoji_Sentiment_Data_v1.0.csv
Velikost 77.03 KB
Format Unknown
Opis CSV table of Emoji Sentiment Ranking
Prevzem datoteke

Ime Emojitracker_20150604.csv

Anatomija vnosa 3

dc.contributor.author	Kralj Novak, Petra
dc.contributor.author	Smailović, Jasmina
dc.contributor.author	Sluban, Borut
dc.contributor.author	Mozetič, Igor
dc.date.accessioned	2015-09-15T17:38:49Z
dc.date.available	2016-04-14T23:00:09Z
dc.date.issued	2015-09-14
dc.identifier.uri	http://hdl.handle.net/11356/1048
dc.description	A lexicon of 751 emoji characters with automatically assigned sentiment. The sentiment is computed from 70,000 tweets, labeled by 83 human annotators in 13 European languages. The process and analysis of emoji sentiment ranking is described in the paper: Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. PLoS ONE 10(12): e0144296. doi:10.1371/journal.pone.0144296
dc.description.provenance	Submitted by Igor Mozetic;Mozetič (igor.mozetic@ijs.si) on 2015-09-15T15:23:10Z No. of bitstreams: 2 Emoji_Sentiment_Ranking_v1.0.csv: 61395 bytes, checksum: fde4f87536275bd7a767339d6fcff486 (MD5) ESR_v1.0_format.txt: 506 bytes, checksum: e1ba7f53f89b74dd3c4e5fcf63a2f09f (MD5)
dc.description.provenance	Approved for entry into archive by Tomaz;Tomaž Erjavec (tomaz.erjavec@ijs.si) on 2015-09-15T17:38:49Z No. of bitstreams: 2 Emoji_Sentiment_Ranking_v1.0.csv: 61395 bytes, checksum: fde4f87536275bd7a767339d6fcff486 (MD5) ESR_v1.0_format.txt: 506 bytes, checksum: e1ba7f53f89b74dd3c4e5fcf63a2f09f (MD5)
dc.description.provenance	Made available in DSpace Tomaz;Tomaž Erjavec (tomaz.erjavec@ijs.si) on 2015-09-15T17:38:49Z Previous issue date: 2015-09-14No. of bitstreams: 2 Emoji_Sentiment_Ranking_v1.0.csv: 61395 bytes, checksum: fde4f87536275bd7a767339d6fcff486 (MD5) ESR_v1.0_format.txt: 506 bytes, checksum: e1ba7f53f89b74dd3c4e5fcf63a2f09f (MD5)

Statistike



Statistike

Showing statistics from 2015-09-14

Skupno št. obiskov

	Ogledi
Emoji Sentiment Ranking 1.0	3130

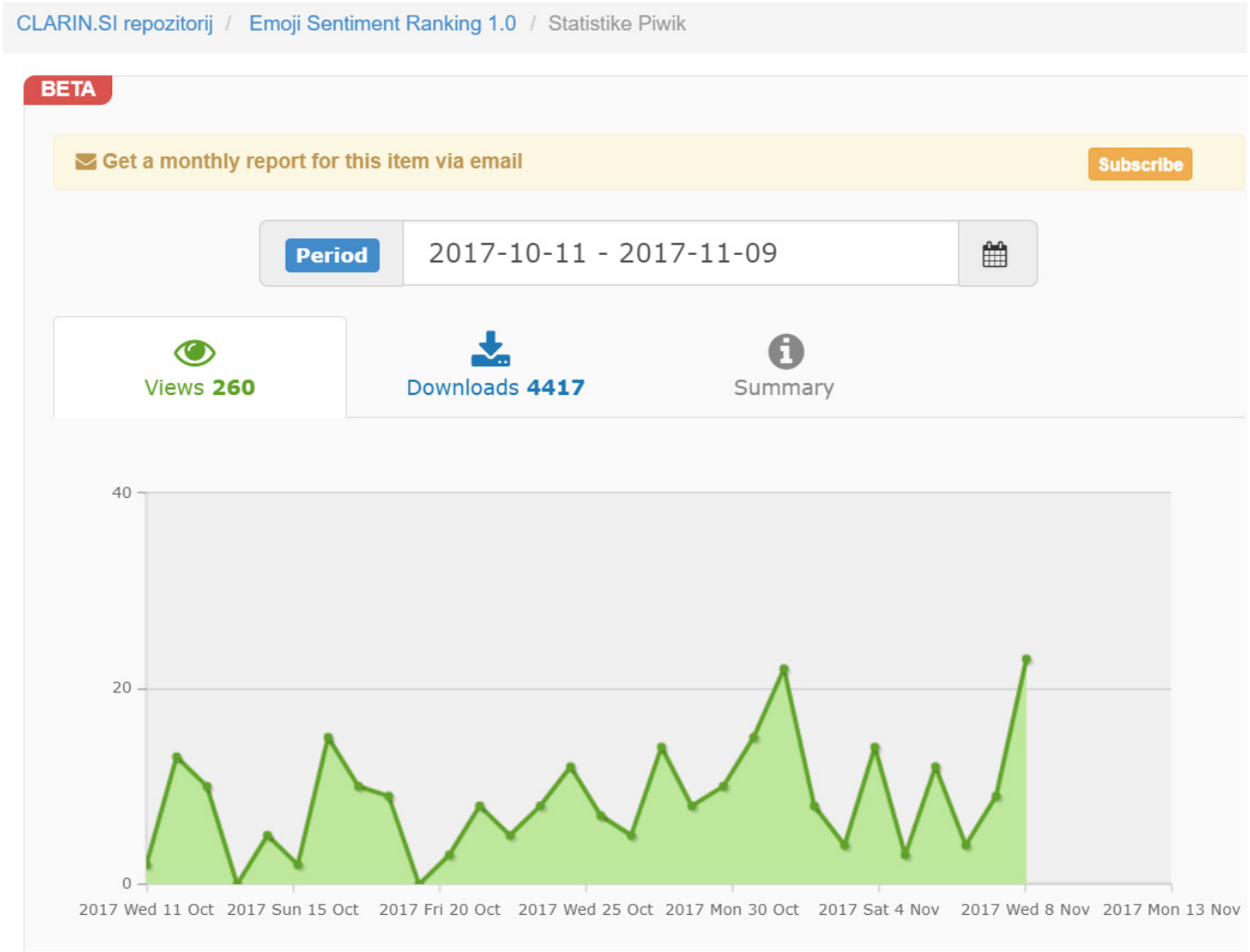
Skupno število obiskov v določenem letu

	2015	2016	2017
Emoji Sentiment Ranking 1.0	339	1193	1598

Top country views

	Ogledi
United States	812
Slovenija	444
Germany	217
United Kingdom	181
China	144

Piwik

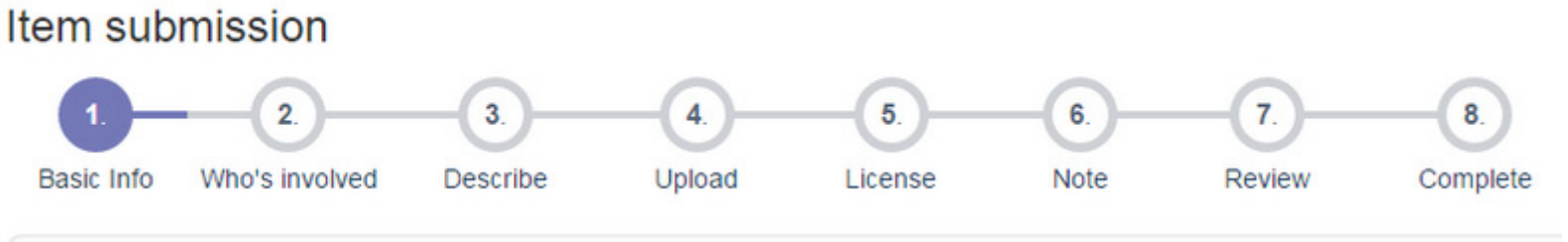

CLARIN.SI repozitorij / Emoji Sentiment Ranking 1.0 / Statistike Piwik
BETA
Get a monthly report for this item via email
Subscribe
Period
2017-10-11 - 2017-11-09
Views 260
Downloads 4417
Summary
40
20
0
2017 Wed 11 Oct
2017 Sun 15 Oct
2017 Fri 20 Oct
2017 Wed 25 Oct
2017 Mon 30 Oct
2017 Sat 4 Nov
2017 Wed 8 Nov
2017 Mon 13 Nov

Repozitorij: prevzem

- Večina vnosov v CLARIN.SI je CC
- Ostali: AAI prijava, strinjanje s pogoji uporabe
- Citiranje!

Repozitorij: vnos

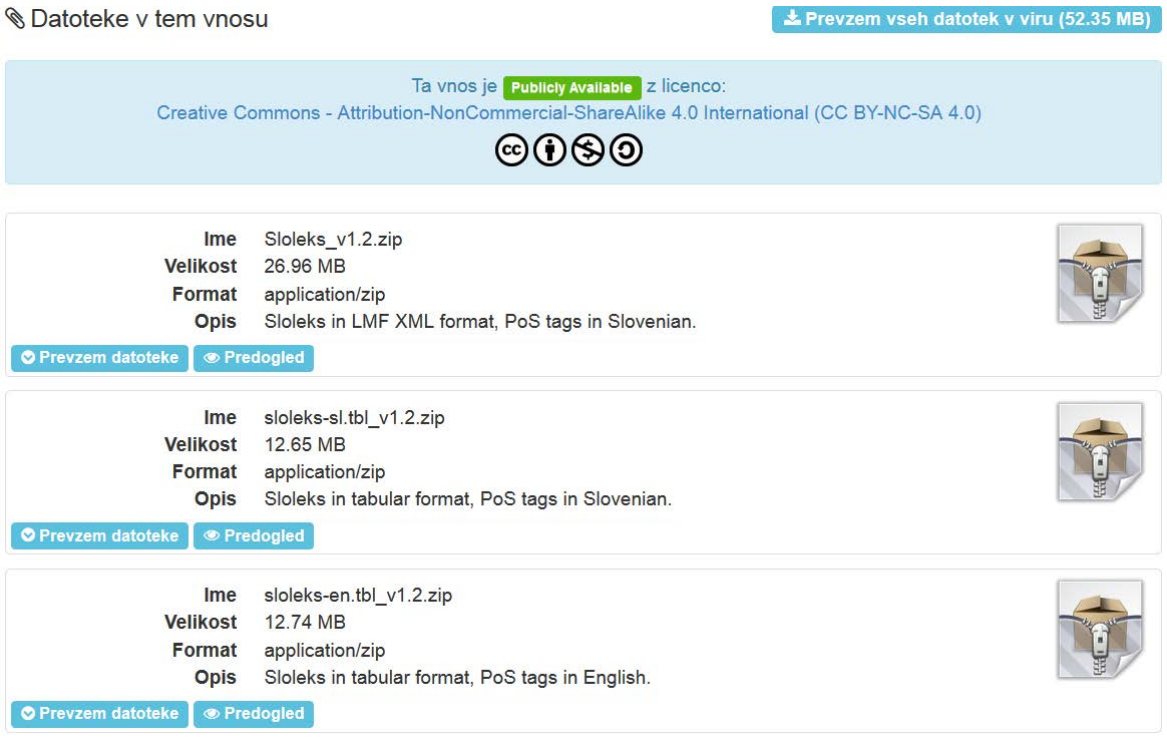
- Prijava AAI
- Razmeroma enostaven delotok



- Izbira licence
- Možnost deljenja vnašanja
- Vnos pregleda eden od CLARIN.SI urednikov
- Možnost embarga podatkov
- Možnost vnosa nove različice vira

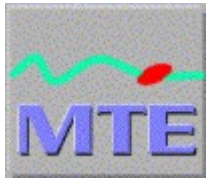
Podatki

- XML po eni od „standardnih“ shem
  - Text Encoding Initiative Guidelines (TEI P5)
  - za enostavno strukturirane podatke zadošča tabela
- + izvedeni formati + dokumentacija
- ZIP



Kaj je zaenkrat v repozitoriju?

- MULTEXT-East
- Viri za učence japonščine
- fi, hr, sr, bs hr-en, fi-en
- Šolar
- sloLeks, hrLex, srLex
- sloWNet
- N-grami
- Emoji
- Slovarji (ZRC SAZU)
- Parlament
- VAYNA



Učne množice

- sodobna slovenščina: **ssj500k**
  MSDji, leme, imena, skladnja, glagolske fraze (2.0)
- starejša slovenščina: **goo300k**
  posodobljene besede, ~MSDji, leme, imena (faksimile)
- spletni mediji:
  - **Janes-Norm**
    standardizirane besede
  - **Janes-Tag**
    standardizirane besede, MSDji, leme, imena
- govor: **GosVL**
  „fonetična“ in standardizirana transkripcija

- zaledni leksikon: **sloLeks, imp25k**

Veliki korpusi

metapodatki + avtomatsko označeni: (normalizirane besede), MSDji, leme

- sodobna slovenščina, referenčni korpusi, projekt SSJ:
  ccGigaFida (100M), ccKRES (10M)
- zgodovinska slovenščina, projekt IMP:
  IMP (18M)
- družbeni mediji, projekt JANES:
  Janes-Tweet (160M), **Janes-Forum** (50M), **Janes-Blog** (35M), **Janes-News** (15M), **Janes-Wiki** (5M)
- splet (WaC: Web as Corpus)
  hrWaC, srWaC, bsWaC, slWaC (po ~1000M)

Govorjeni korpusi

- učenje razpoznavanja govora:
  SNABI (132 govorcev x 200 stavkov)
- govorjene poizvedbe po letalskih letih:
  SOFES (12,536 izjav, 10 ur)
- posneta predavanja:
  GosVL (25 predavanj, 10 ur)
- referenčni:
  Gos (1000.000 besed, ~~120 ur~~)

Zaključki

- CLARIN.SI ponuja odprte podatke o slovenskem (in drugih) jezikih, primernih za pedagoške, raziskovalne in tudi razvojne namene
- uporabo podatkov je potrebno citirati

- CLARIN.SI je zelo vesel novih podatkov, če so le:
  - netrivialni
  - razmeroma kvalitetni
  - primerno kodirani
  - dobro opisani
- uporabo vaših podatkov bo potrebno citirati