Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema

LKSSAIS vystymas

Modernizuojamos LKSSAIS Lietuviškų dokumentų tekstų pagrindinės ir specializuotos statistinės analizės IT sprendimo komponento administratoriaus instrukcija

Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema

TURINYS

LENTELIŲ SĄRAŠAS

1. Įžanga

1.1. Dokumento paskirtis

Šio dokumento paskirtis – pateikti modernizuojamos LKSSAIS Statistikos IT sprendimo komponento administravimo vadovą.

1.2. Sutrumpinimai

1.1 lentelė. Sutrumpinimai

Santrumpa	Paaiškinimas
LKSSAIS	Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema
IT	Informacinės technologijos

1.3. Panaudotų dokumentų sąrašas

1.2 lentelė. Panaudotų dokumentų sąrašas

Eil. Nr.	Pavadinimas
1.	Modernizuojamos LKSSAIS Statistikos IT sprendimo komponento detalios analizės ir projektavimo specifikacija
2.	TEKSTO PILNOS IR SPECIALIZUOTOS STATISTINĖS ANALIZĖS BENDROJO NAUDOJIMO IT SPRENDIMAS, Koncepcijos specifikacija, Kaunas, 2018.

1.4. Statistikos IT sprendimo paskirtis ir tikslai

Lietuviškų dokumentų tekstų pagrindinės ir specializuotos statistinės analizės IT sprendimas (toliau Statistikos IT sprendimas) analizuos dokumento teksto leksinius segmentus ir morfologines leksinių vienetų žymas ir suskaičiuos bazinius teksto statistikos parametrus (angl. *basic text statistics*), dalinę kalbos dalių statistiką, teksto leksinį tankį (angl. *lexical density*). Suskaičiuoti parametrai bus grąžinami IT sprendimo funkcionalumą iškvietusiai sistemai, posistemei ar komponentui.

2. Statistikos IT sprendimo administratoriaus instrukcija

2.1. Statistikos IT sprendimo komponento sandara

Statistikos IT sprendimo komponentas veikia Docker konteineryje.

Komponento vykdomieji failai konteineryje patalpinti kataloge */trm/webservice*.

Komponento tinklinė paslauga konteineryje pasiekiama adresu *http://0.0.0.0:4000/api*. Tinklinės paslaugos kontrakto aprašymą *Swagger* (*OpenAPI*) formatu galima gauti kreipiantis adresu *http://0.0.0.0:4000/swagger/api/swagger.json*.

Komponento tinklinės paslaugos adresas įdiegus Kubernetes aplinkoje priklauso nuo to kokiu adresu Kubernetes išviešinta konteineryje veikianti tinklinė paslauga. Šio adreso parinkimas paliekamas administratoriui. Administratorius gali patikrinti ar parinktas adresas veikia teisingai per jį kviesdamas vidinį konteinerio resursą http://0.0.0.0:4000/api/ping . Jeigu resursas atsako formatu *„PONG: laiko-štampas“*, vidinė paslauga pasiekiama.

Komponento tinklinė paslauga turi du metodus, aprašytus lentelėje 2.1.

2.1 lentelė. Komponento tinklinės paslaugos metodai

Metodas	Paaiškinimas
/api/	Skaičiuoja duoto dokumento statistinius ir skaitomumo įverčius. Kviečiamas per HTTP POST antraštėje *Content-Type* nurodant turinio tipą *application/json* bei teisingą perduodamo teksto koduotę. Įėjimo ir išėjimo formatai aprašyti skyriuje 2.3
/api/ping	Leidžia patikrinti ar paslauga veikia. Kviečiamas per HTTP GET arba HTTP POST. Atsako formatu „PONG: laiko-štampas“.

2.2. Statistikos IT sprendimo komponento diegimas į sistemą

2.2.1. Docker konteinerio sutvėrimas

Komponento Docker konteinerio sutvėrimui reikia turėti komponento išeitinius tekstus bei vykdomąją aplinką tenkinančia tokius parametrus: *Ubuntu Linux v>=18.04*, *.NET Core v>=2.1*, *Docker v>=18*.

Komponento išeitiniuose tekstuose reikia eiti į katalogą *TextReadability/WebService* ir jame įvykdyti komandas:

```
dotnet publish
docker build –t "trm:0.1" ./
```

Lokalioje sistemoje bus sutverta komponento Docker saugykla (angl. *repository*) pavadinta *trm* su žyma (angl. *tag*) *0.1*.

2.2.2. Diegimas į Kubernetes aplinką

Komponento diegimui į Kubernetes aplinką reikia turėti komponento Docker konteinerį patalpintą Kubernetes aplinkai pasiekiamoje Docker saugykloje (angl. *repository*).

Diegiant Kubernetes aplinkoje komponentui sukuriami diegimo (angl. *deployment*) bei paslaugos (angl. *service*) aprašai. Diegimo aprašo pavyzdys:

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: semantika2-trm-webservice
  labels:
    app: semantika2-trm-webservice
spec:
  replicas: 1
  selector:
```

```
    matchLabels:
      app: semantika2-trm-webservice
  template:
    metadata:
      labels:
        app: semantika2-trm-webservice
    spec:
      containers:
      - name: semantika2-trm-webservice
        image: docker-konteinerio-adresas
        ports:
        - containerPort: 4000
```

čia *docker-konteinerio-adresas* yra komponento Docker konteinerio adresas Kubernetes aplinkai pasiekiamoje Docker saugykloje.

Laikant, kad diegimo aprašas patalpintas faile *diegimas.yml*, Kubernetes aplinkoje diegimas sukuriamas naudojant komandą:

```
kubectl apply -f diegimas.yml
```

Paslaugos aprašo pavyzdys:

```
apiVersion: v1
kind: Service
metadata:
  name: semantika2-trm-webservice
spec:
  ports:
  - port: 80
    protocol: TCP
    targetPort: 4000
  selector:
    app: semantika2-trm-webservice
```

šis aprašas Kubernetes aplinkoje komponento tinklinę paslaugą padarys prieinamą per TCP prievadą 80. Prievadą galima pakeisti į bet kurį norimą.

Laikant, kad paslaugos aprašas patalpintas faile *paslauga.yml*, Kubernetes aplinkoje paslauga sukuriama naudojant komandą:

```
kubectl apply -f paslauga.yml
```

2.3. Statistikos IT sprendimo komponento šalinimas

2.3.1. Docker konteinerio pašalinimas

Šiame skyriuje laikoma, kad komponento Docker konteineris lokalioje aplinkoje buvo sutvertas laikantis skyriuje 2.2.1 nurodytų žingsnių.

Iš lokalios aplinkos komponento Docker konteineris pašalinimas naudojant komandą:

```
docker image remove trm:0.1
```

2.3.2. Pašalinimas iš Kubernetes aplinkos

Šiame skyriuje laikoma, kad komponentas Kubernetes aplinkoje buvo įdiegtas laikantis skyriuje 2.2.2 nurodytų žingsnių.

Komponento paslauga pašalinama naudojant komandą:

```
kubectl delete service semantika2-trm-webservice
```

Komponento diegimas pašalinimas naudojant komandą:

```
kubectl delete deployment semantika2-trm-webservice
```

2.4. Statistikos IT sprendimo komponento veikimas

2.4.1. Įėjimas

Komponento įėjimas turi būti JSON dokumentas tokiu formatu:

```
{
   "body": "dokumento tekstas",
   "annotations": {
      "lex" : ... leksikografinės anotacijos dokumentui ...,
      "morph" : ... morfologinės anotacijos dokumentui ...
   }
}
```

Bet kokie papildomi laukai įėjimo dokumente bus ignoruojami.

2.4.2. Išėjimas

Klaidos atveju komponentas grąžina HTTP atsakymo kodą 500 bei išėjimą tokiu formatu:

```
{
   "error": {
      "status" : 500,
      "message" : "klaidos žinutė"
   }
}
```

Komponento išėjimas normaliu atveju yra JSON dokumentas tokiu formatu:

```
{
  "text": { parametrai 1-7, 3.1 lentelėje },
  "lexical": { parametrai 8-13, 3.1 lentelėje },
  "means": { parametrai 14-20, 3.1 lentelėje },
  "min": { parametrai 21-23, 3.1 lentelėje },
  "max": { parametrai 24-26, 3.1 lentelėje },
  "word": { parametrai 27-28, 3.1 lentelėje },
  "freq": { parametrai 29-32, 3.1 lentelėje },
  "morph": { parametrai 33-44, 3.1 lentelėje },
  "morphParameters": { visi parametrai 3.2 lentelėje },
  "readability": { skaitomumo indeksų įverčiai 3.3 lentelėje }
}
```

Čia kiekvienas parametras ir skaitomumo indeksas yra atskiras JSON struktūros laukas, kurio pavadinimas atspindi atitinkamo parametro arba skaitomumo indekso pavadinimą. Detalią dokumento struktūrą galima gauti komponento tinklinės paslaugos *Swagger (OpenAPI)* dokumentacijoje kreipiantis į resursą */swagger/api/swagger.json* , arba tiesiog į */swagger/index.html* .

3. Priedai

3.1. Modernizuojamos LKSSAIS Statistiko IT sprendimo komponento skaičiuojami statistikos parametrai (atkelta iš [2])

3.1 lentelė. Modernizuojamos LKSSAIS Statistikos IT sprendimo komponento skaičiuojami statistikos parametrai (atkelta iš [2])

Nr.	Aprašymas
	Teksto statistika
1.	Simbolių skaičius su tarpais
2.	Simbolių skaičius be tarpų
3.	Skiemenų skaičius
4.	Leksemų skaičius
5.	Žodžių skaičius
6.	Unikalių žodžių (lemų) skaičius
7.	Sakinių skaičius
	Leksinis tankis
8.	Viso teksto Leksinių (turinį perteikiančių) žodžių skaičius
9.	Viso teksto Neleksinių (funkcinių) žodžių skaičius
10.	Viso teksto Leksinis tankis
11.	Sakinio leksinių (turinį perteikiančių) žodžių skaičius
12.	Sakinio neleksinių (funkcinių) žodžių skaičius
13.	Sakinio leksinis tankis
	Vidurkiai
14.	Žodžio ilgio vidurkis simboliais
15.	Žodžio ilgio vidurkis skiemenimis
16.	Žodžio ilgio mediana (simboliais/skiemenimis)
17.	Standartinis žodžio ilgio nuokrypis
18.	Sakinio ilgio vidurkis žodžiais (leksemomis)
19.	Sakinio ilgio mediana
20.	Sakinio ilgio standartinis nuokrypis
	Min
21.	Trumpiausias žodis simboliais
22.	Trumpiausias žodis skiemenimis
23.	Trumpiausias sakinys žodžiais (leksemomis)
	Max
24.	Ilgiausias žodis simboliais
25.	Ilgiausias žodis skiemenimis
26.	Ilgiausias sakinys žodžiais (leksemomis)
	Žodžių ilgio simboliais statistika
27.	Žodžių skaičius ir procentinė dalis kiekvienam žodžio ilgiui simboliais (t. y., vieno simbolio žodžių skaičius ir procentinė dalis, dviejų simbolių žodžių skaičius ir procentinė dalis, trijų simbolių žodžių skaičius ir procentinė dalis ir t.t. iki didžiausio, koks yra analizuotame tekste, ilgio simboliais žodžių skaičiaus ir procentinės dalies)
	Žodžių ilgio skiemenimis statistika
28.	Žodžių skaičius ir procentinė dalis kiekvienam žodžio ilgiui skiemenimis (t. y., vieno skiemens žodžių skaičius ir procentinė dalis, dviejų skiemenų žodžių skaičius ir procentinė dalis, trijų skiemenų žodžių skaičius ir procentinė dalis ir t.t. iki didžiausio, koks yra analizuotame tekste, ilgio skiemenimis žodžių skaičiaus ir procentinės dalies)
	Dažniai
29.	Dažniausių žodžių (funkcinių ir nefunkcinių)
30.	Dažniausių nefunkcinių žodžių pagal dydį ir spalvą (debesis)
31.	Dažniausių žodžių porų
32.	Dažniausių žodžių tripletų
	Teksto morfologinė struktūra

Nr.	Aprašymas
33.	Daiktavardžių skaičius ir procentinė dalis
34.	Asmenuojamų veiksmažodžių skaičius ir procentinė dalis
35.	Bendraties veiksmažodžių skaičius ir procentinė dalis
36.	Dalyvių skaičius ir procentinė dalis
37.	Padalyvių skaičius ir procentinė dalis
38.	Pusdalyvių skaičius ir procentinė dalis
39.	Būdvardžių skaičius ir procentinė dalis
40.	Skaitvardžių skaičius ir procentinė dalis
41.	Įvardžių skaičius ir procentinė dalis
42.	Prieveiksmių skaičius ir procentinė dalis
43.	Kitų kalbos dalių skaičius ir procentinė dalis
44.	Neatpažintų žodžių leksemų skaičius ir procentinė dalis

3.2 lentelė. Modernizuojamos LKSSAIS Statistikos IT sprendimo komponento skaičiuojami morfologinių savybių statistikos parametrai (atkelta iš [2])

Kalbos dalis	Vertinama morfologinė savybė	
	pavadinimas	**galimos reikšmės**
daiktavardis	pobūdis	bendrinis, tikrinis
	giminė	bendroji, moteriška, vyriška
	skaičius	daugiskaita, vienaskaita, dviskaita
	linksnis	vardininkas, kilmininkas, naudininkas, galininkas, įnagininkas, vietininkas, šauksmininkas
	ar sangrąžinis	sangrąžinis, nesangrąžinis
	vardo eilė	vardas, pavardė, vietovė
asmenuojamas veiksmažodis	laikas	esamasis, būtasis kartinis, būtasis dažninis, būsimasis, -
	asmuo	I, II, III
	skaičių	daugiskaita, vienaskaita
	ar neigiamas	teigiamas, neigiamas
	ar sangrąžinis	sangrąžinis, nesangrąžinis
	nuosaka	tiesioginė, tariamoji, liepiamoji
bendraties veiksmažodžiams	ar neigiamas	teigiamas, neigiamas
	ar sangrąžinis	sangrąžinis, nesangrąžinis
dalyviai	laikas	esamasis, būtasis kartinis, būtasis dažninis, būsimasis, -
	giminė	bevardė, moteriška, vyriška
	skaičius	daugiskaita, vienaskaita, -
	rūšis	veikiamoji, neveikiamoji, reikiamoji
	ar neigiamas	teigiamas, neigiamas
	ar įvardžiuotinis	įvardžiuotinis, neįvardžiuotinis
	linksnis	vardininkas, kilmininkas, naudininkas, galininkas, įnagininkas, vietininkas, šauksmininkas
	ar sangrąžinis	sangrąžinis, nesangrąžinis
padalyviai	laikas	esamasis, būtasis kartinis, būtasis dažninis, būsimasis
	ar neigiamas	teigiamas, neigiamas
	ar sangrąžinis	sangrąžinis, nesangrąžinis
pusdalyviai	skaičius	daugiskaita, vienaskaita
	giminė	moteriška, vyriška
	ar neigiamas	teigiamas, neigiamas
	ar sangrąžinis	sangrąžinis, nesangrąžinis
būdvardžiai	laipsnis	nelyginis, aukštesnis, aukščiausias, -
	giminė	bevardė, moteriška, vyriška, -
	skaičius	daugiskaita, vienaskaita, dviskaita, -

Kalbos dalis	**Vertinama morfologinė savybė**	
	pavadinimas	**galimos reikšmės**
	linksnis	vardininkas, kilmininkas, naudininkas, galininkas, įnagininkas, vietininkas, šauksmininkas
	ar įvardžiuotinis	įvardžiuotinis, neįvardžiuotinis, -
skaitvardžiai	grupė	kiekiniai, kuopiniai, -
	giminė	bevardė, moteriška, vyriška, -
	skaičius	daugiskaita, vienaskaita, -
	linksnis	vardininkas, kilmininkas, naudininkas, galininkas, įnagininkas, vietininkas, šauksmininkas, -
	forma	raidinė, skaitmeninė, romėniškas
	ar įvardžiuotinis	įvardžiuotinis, neįvardžiuotinis, -
įvardžiai	giminė	bevardė, moteriška, vyriška, -
	skaičius	daugiskaita, vienaskaita, dviskaita, -
	linksnis	vardininkas, kilmininkas, naudininkas, galininkas, įnagininkas, vietininkas, šauksmininkas
	ar įvardžiuotinis	įvardžiuotinis, neįvardžiuotinis, -
prieveiksmiai	laipsnis	nelyginis, aukštesnis, aukščiausias
kitos kalbos dalys	pobūdis, jei tokį turi	jaustukas, ištiktukas, dalelytė, prielinksnis, jungtukas, trumpinys, skyrybos ženklas, neatpažintas

3.3 lentelė. Modernizuojamos LKSSAIS Statistikos IT sprendimo skaičiuojami skaitomumo indeksai ir susijusios statistikos (atkelta iš [2], papildyta komentarais)

Nr.	**Aprašymas**
	Skaitomumo indeksai
1.	*Flesch-Kincaid* indeksas (skyla į du indeksus: *Flesch-Kincaid Grade Level* ir *Flesch-Kinkaid Reading Ease*)
2.	*Gunning Fog* indeksas
3.	*Coleman-Liau* indeksas
4.	*SMOG* indeksas
5.	*Automated Readability* indeksas
6.	*Indeksų vidurkis* (apima 1-5 eilutes, 8-9 yra metrikos)
7.	*Indeksų mediana* (apima 1-5 eilutes, 8-9 yra metrikos)
8.	*Fry* skaitomumo metrika (X ir Y dedamosios *Fry* grafike)
9.	*Raygor* skaitomumo metrika (X ir Y dedamosios *Raygor* grafike)