Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema

LKSSAIS vystymas

Automatinio dokumentų santraukų sudarymo IT sprendimo administratoriaus instrukcija

9 puslapiai

2020

TURINYS

1. Įžanga

1.1. Dokumento paskirtis

Šio dokumento paskirtis – pateikti modernizuojamos LKSSAIS automatinio dokumentų santraukų sudarymo (ADSS) IT sprendimo administratoriaus instrukciją. Šiame dokumente aprašoma:

1. ADSS komponento serviso diegimo instrukcija.
2. ADSS komponento grafinės sąsajos diegimo instrukcija.

1.2. Automatinio dokumentų santraukų sudarymo IT sprendimo komponento aprašymas

Automatinio dokumentų santraukų sudarymo IT komponentas susideda iš kelių konteinerizuotų servisų:

- ADSS serviso.
- ADSS grafinės sąsajos.

ADSS komponentui reikalingi papildomi Semantika komponentai:

- *LEX* – segmentuoja tekstus į žodžius ir sakinius nurodydamas jų ribas ir pateikia *json* formate; šio komponento rezultatas būtinas MORPH ir ADSS komponentams;
- *MORPH* – nustato žodžių lemas ir morfologinę informaciją ir pateikia *json* formate; šio komponento rezultatas yra būtinas ADSS komponentui;
- *NER* – nustato įvardytąsias esybes (komponentas nėra privalomas).

1.3. Sutrumpinimai

1.1 lentelė. Sutrumpinimai

Santrumpa	Paaiškinimas
LKSSAIS	Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema
ADSS	Automatinis dokumentų santraukų sudarymas
VDU vykdytojai	Užsakovo (VDU) projekto vykdytojų komanda
Diegėjas	Programavimo ir diegimo paslaugų konkursą laimėjusi įmonė (UAB ATEA)

1.4. Panaudotų dokumentų sąrašas

1.2 lentelė. Panaudotų dokumentų sąrašas

Eil. Nr.	Pavadinimas
1.	ADSS IT sprendimo detalios analizės ir projektavimo specifikacija

1.5. ADSS IT sprendimo paskirtis ir tikslai

Čia aprašomas ADSS IT sprendimas yra skirtas automatiškai sudaryti lietuvių kalbos tekstų santraukas bei nustatyti teksto raktažodžius. ADSS IT sprendimas skirtas žiniasklaidos, administraciniams ar teisės tekstams. Šis IT sprendimas naudodamasis mašininio apmokinimo

algoritmais nustato raktažodžius ir sakinius, kurie geriausiai iliustruoja duotojo teksto turinį. Jo įeigoje turi būti lietuvių kalbos, UTF-8 koduotės tekstas.

ADSS IT sprendimas naudoja projekte modernizuotas interneto tarnybas. Anotatorius: segmentavimo komponento interneto tarnybą (LEX), morfologijos interneto tarnybą kalbos dalių ir antraštinių žodžių formų (lemų) nustatymui (MORPH), įvardytųjų esybių nustatymo komponento tarnybą (NER) ir santraukų nustatymo komponentą (ADSS).

Santraukas galima sudaryti visiems failams, kuriuos palaiko *Apache TIKA* įrankis[1].

[1] https://tika.apache.org/1.24.1/formats.html

2. ADSS serviso administratoriaus instrukcija

2.1. ADSS serviso parengimas darbui ir paleidimas

Administravimui gali būti naudojamas Linux, MacOS arba Windows operacinę sistemą turintis kompiuteris. Kompiuteryje turi būti įdiegtas *Docker* įrankis.

Serveryje arba kompiuteryje išarchyvuojamas failas „**adss-service.zip**“. Atsidarote direktoriją pavadinimu „**adss-service**“.

Komponento paleidimui reikia sukurti *docker* konteinerį:

```
docker build -t adss-service .
```

Konteineris paleidžiamas su komanda:

```
docker container run --detach --name service -p 5000:5000 adss-service
```

Servisas startuoja per kelias minutes.

2.2. ADSS serviso veikimo patikrinimas

Norint įsitikinti ar servisas sėkmingai startavo, galima nueiti adresu http://0.0.0.0:5000/health jei atsakymas yra `{"status": "UP"}`, tai servisas sėkmingai startavo.

2.3. ADSS serviso užklausa

Užklausa turi būti *UTF-8* koduotės ir lietuvių kalba. Užklausa ir atsakymas yra pateikiamas *JSON* formatu.

Užklausos pavyzdys:

```
curl --request POST \
  --url http://localhost:5000/hs \
  --header 'content-type: application/json' \
  --data
'{
 "body": "Tekstas",
 "annotations": {
     "lex": {Lex komponento anotacija},
     "morph": {Morph komponento anotacija},
     "ner": {Ner komponento anotacija}
     },
 "type": "media",
 "options": {
          "debug": false, "min": 3, "max": 10, "keywords": 10
     }
}'
```

Užklausos parametrai pateikiami 2.1 lentelėje.

2.1 lentelė. Užklausos parametrai

Pavadinimas	Tipas	Aprašymas	Kita
body	*String*	Tekstas	Privalomas, UTF-8, lietuvių kalba
lex	*JSON* objektas	LEX komponento anotacija	Privalomas
morph	*JSON* objektas	MORPH komponento anotacija	Privalomas
ner	*JSON* objektas	NER komponento anotacija	Neprivalomas
type	*String*	Dokumento tipas: žiniasklaida (media), teisė (law), administracinis (administration)	Neprivalomas, Jei nėra nurodytas – bus žiniasklaidos tipas
debug	*True/False*	Parametras ar nurodantis ar rodyti ar ne sakinių ir raktažodžių įverčius	Neprivalomas
min	Skaičius nuo 1 - 30	Minimalus santraukos sakinių skaičius	Neprivalomas
max	Skaičius nuo 1 – 30	Maksimalus santraukos sakinių skaičius	Neprivalomas
keywords	Skaičius nuo 1 - 30	Maksimalus raktažodžių skaičius	Neprivalomas

Užklausos atsakymas:


```
{
  "data": [
    {
      "type": "Saskaita_Bankas",
      "value": [
        "Valstybinis bankas"
      ]
    },
    {
      "type": "Produktas",
      "value": [
        "Bliss by"
       ]
    },
    {
      "type": "Asmuo",
      "value": [
        "Violeta",
        "Regimantas Sudžius"
      ]
    },
  ],
  "keywords": [
```

```
    "Invega",
    "karantino"
  ],
  "summary": "COVID paramos verslui laikrodis dar tiksi: kam lėšos išseks po mėnesio…"
}
```


Atsakymo rezultatų paaiškinimai pateikiami 2.2 lentelė. Atsakymo paaiškinimai

2.2 lentelė. Atsakymo paaiškinimai

Pavadinimas	Tipas	Aprašymas
data	*JSON* masyvas	Tekste rastos įvardintos esybės. Pateikiamos masyve, kaip tipas ir reikšmė.
keywords	*JSON* masyvas	Pagrindiniai teksto raktažodžiai.
summary	*String*	Sugeneruota teksto santrauka.

2.4. ADSS serviso konfigūravimas

ADSS serverį galima konfigūruoti redaguojant „**config.ini**“ failą. Galima pakeisti:

1. Prievado numerį (**Port);**
2. Minimalų grąžinamų sakinių skaičių (**MinSentence**);
3. Maksimalų grąžinamų sakinių skaičių (**MaxSentence**);
4. Raktažodžių skaičių (**KeywordCount**).

3. ADSS grafinės sąsajos administratoriaus instrukcija

3.1. ADSS grafinės sąsajos parengimas darbui ir paleidimas

Administravimui gali būti naudojamas Linux, MacOS arba Windows operacinę sistemą turintis kompiuteris. Kompiuteryje turi būti įdiegtas *Docker* įrankis.

Serveryje arba kompiuteryje išarchyvuojamas failas „**adss-gui.zip**“. Atsidarote direktoriją pavadinimu „**adss-gui**“.

Komponento paleidimui reikia sukurti *docker* konteinerį:

```
docker build -t adss-gui .
```

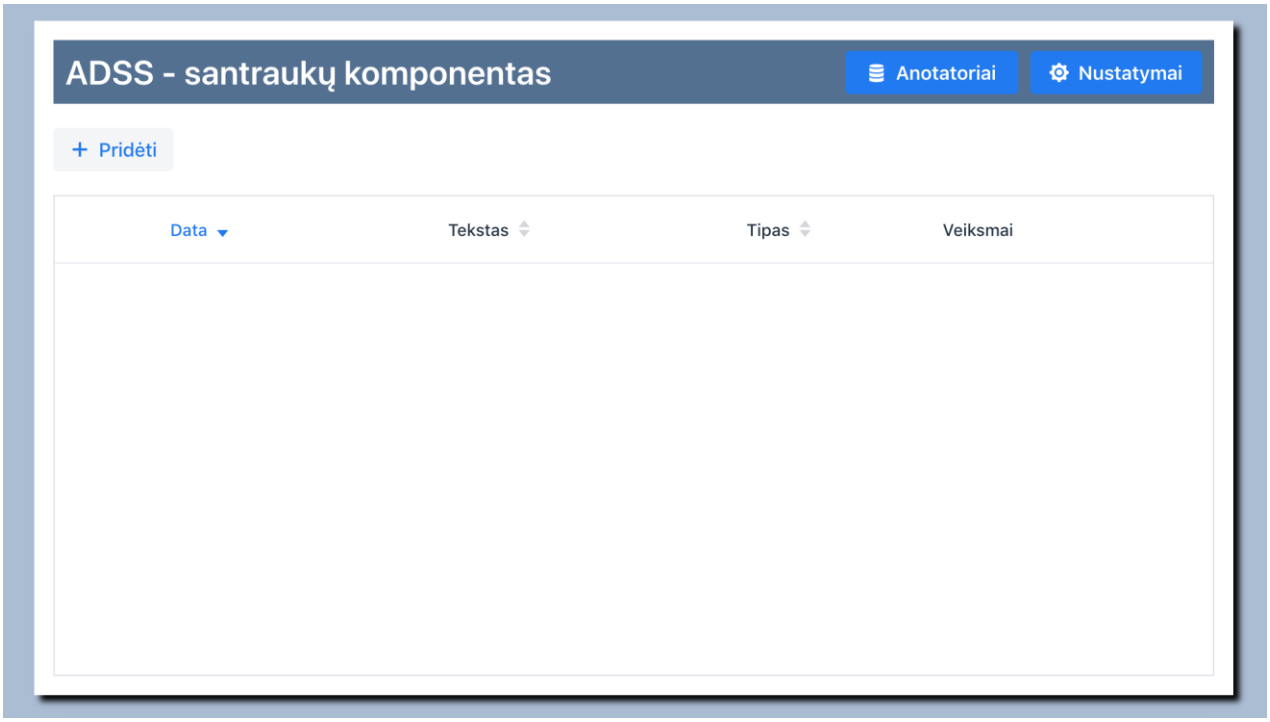
Konteineris paleidžiamas su komanda:

```
docker container run --detach --name gui -p 8080:8080 adss-gui
```

Grafinė sąsaja startuoja per kelias minutes.

3.2. ADSS grafinės sąsajos veikimo patikrinimas

Norint įsitikinti ar grafinė sąsaja veikia, galima nueiti adresu http://0.0.0.0:8080/ atsidarys pagrindinis grafinės sąsajos langas (3.1 pav. Pagrindinis langas)



3.1 pav. Pagrindinis langas

3.3. ADSS grafinės sąsajos paruošimas darbui

Prieš pradedant naudoti grafinę sąsaja, privaloma nurodyti Semantika komponentų adresus.

1. Reikia paspausti mygtuką **[Anotatoriai]**.
2. Atsidaro Semantika komponentų nustatymo langas (3.2 pav.).

3.2 pav. Anotatorių nustatymai

3. Reikia nurodyti Semantika komponentų adresus:
    a. *LEX* komponento.
    b. *MORPH* komponento.
    c. *NER* komponento.
    d. *ADSS (serviso)* komponento.