Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema

Lietuviškų dokumentų indeksavimo ir pažangios paieškos komponento administratoriaus instrukcija

Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema

Lietuviškų dokumentų indeksavimo ir pažangios paieškos komponento administratoriaus instrukcija

TURINYS

1. Įžanga

1.1. Dokumento paskirtis

Šiame dokumente aprašoma:

1. Lietuviškų dokumentų indeksavimo ir pažangios paieškos komponento prototipo diegimas į *Debian OS*.
2. Duomenų indeksavimas ir atnaujinimas.

1.2. Lietuviškų dokumentų indeksavimo ir pažangios paieškos komponento prototipo aprašymas

Lietuviškų dokumentų indeksavimo ir pažangios paieškos komponento prototipas veikia Linux OS, naudojant standartinį *SOLR* paieškos variklį, papildytą lietuvišku morfologiniu žodynu. Prototipe iliustracijai naudojami iš Užsakovo gauti saugyklos dokumentai JSON formatu (2018-05-27 atsiuntė Darius Amilevičius failu export_anno 20190527.zip). Šie duomenys indeksuoti ir pateikiami *SOLR* branduolyje *Semantika2*.

2. Lietuviškų dokumentų indeksavimo ir pažangios paieškos komponento prototipo administravimas

2.1. Pradiniai reikalavimai

Administravimui, diegimui naudojamas kompiuteris su *Debian* 4.9.144-3.1 (2019-02-19) operacine sistema. Prototipui realizuoti čia naudojama *SOLR* 8.1.0 versija. Iš terminalinio lango vykdoma:

```
sudo bash ./install_solr_service.sh solr-8.1.0.tgz -i /opt -d /var/solr -u solr -s solr -p 8983
```

Jeigu vėliau norima pakeisti prievadą, pirmiausia pasitikriname naudojamą prievadą:

```
sudo service solr status
```

Tada sustabdomas SOLR servisas:

```
/opt/solr-8.1.0/bin$ sudo service solr stop
```

Atliekama:

```
cd to /opt/solr-8.1.0/server/solr/
```

tada redaguojamas failas /opt/solr-8.1.0/server/solr/**solr.xml**, jame surandama ${jetty.port:8983}, ir čia prievadas 8983 pakeičiamas į norimą.

Analogiškai redaguojamas failas /var/solr/data/**solr.xml**, jame taip pat surandama ${jetty.port:8983}, čia prievadas 8983 pakeičiamas į norimą.

Po to redaguojamas failas /etc/default/**solr.in.sh**, jame surandama SOLR_PORT="8983", čia prievadas 8983 pakeičiamas į norimą. Failas **solr.in.sh** išsaugomas ir uždaromas, tada iš terminalinio lango jis perkraunamas:

```
/etc/default# source solr.in.sh
```

Tada iš naujo paleidžiame servisą ir pasitikriname, jog prievadas pakeistas į norimą:

```
/etc/default# sudo service solr start
/etc/default# sudo service solr status
```

2.2. Papildomas lietuviško kamieninimo, lietuviškos morfologijos ir sinonimų žodynų diegimas

Po standartiniu būdu atlikto *SOLR* serviso diegimo būtina atlikti papildomą lietuviško kamieninimo, lietuviškos morfologijos ir sinonimų žodynų diegimą.

2.2.1. Lietuviško kamieninimo filtro atnaujinimas

Standartinėje *SOLR* versijoje reikia pakeisti ten esantį lietuvišką kamieninimą patobulintu kamieninimu. Tuo tikslu iš terminalinio lango vykdoma komanda:

```
sudo cp ~/solr_stemmer_lt/lucene-analyzers-common-8.1.0.jar /opt/solr-8.1.0/server/solr-webapp/webapp/WEB-INF/lib/lucene-analyzers-common-8.1.0.jar
```

2.2.2. Lietuviškos morfologijos žodyno diegimas

Iš terminalinio lango vykdomos komandos:

```
sudo cp ~/lr_sem2_core/Semantika2/lt.dict /opt/solr-8.1.0/server/solr-webapp/webapp/WEB-INF/lib/lt.dict
```

```
sudo cp ~/lr_sem2_core/Semantika2/lt.info /opt/solr-8.1.0/server/solr-webapp/webapp/WEB-INF/lib/lt.info
```

2.2.3. Lietuviškų sinonimų žodynų diegimas

Iš terminalinio lango vykdomos komandos, pvz.:

```
sudo cp ~/lr_sem2_core/Semantika2/conf/synonyms.txt /var/solr/data/Semantika2/conf/synonyms.txt
```

2.3. Komponento veikimo patikrinimas

Sprendimą realizuojantis prototipas kaip servisas veikia eksperimentinėje VDU Semantika 2 terpėje. Pasiekiamas adresu http://158.129.51.163:8983/solr/#/

Šiuo metu paieškos serveryje yra suindeksuoti iš Užsakovo gauti saugyklos dokumentai JSON formatu (2018-05-27 atsiuntė Darius Amilevičius failu export_anno 20190527.zip).

Terminaliniame lange pateikiame testinę užklausą:

```
curl "http://158.129.51.163:8983/solr/Semantika2/select?q=gyva%C4%8Di%C5%B3"
```

Gaunamame atsakyme turi būti tokie fragmentai:


```
{
  "responseHeader":{
    "status":0,
    "QTime":165,
    "params":{
      "q":"gyvačių"}},
  "response":{"numFound":53,"start":0,"maxScore":857.4379,"docs":[
      {
        "estampille_temporelle":"2019-03-18T23:12:24Z",
        "titre":"Kaip elgtis įkandus gyvatei?",
        "url":"https://www.delfi.lt/sveikata/sveikatos-naujienos/kaip-elgtis-ikandus-gyvatei.d?id=65175291",
        "id":"f1891f87-cc7e-4b35-8d40-213b5d595e09",
        "score":857.4379},
      {
        "estampille_temporelle":"2019-03-18T23:12:48Z",
        "titre":"Lietuvos bioįvairovė. Paprastoji angis",
        "url":"https://www.delfi.lt/grynas/gamta/lietuvos-bioivairove-paprastoji-angis.d?id=60187283",
        "id":"4fb80202-bb3f-4ac5-94d9-fb13df15f2a2",
        "score":857.4379},

...

      {
        "estampille_temporelle":"2019-03-19T12:43:40Z",
        "titre":"Obuolių actas - nuo antsvorio ir ligų",
```

```
            "url":"https://www.delfi.lt/gyvenimas/grozis_ir_sveikata/obuoliu-actas-nuo-antsvorio-ir-ligu.d?id=61924689",
            "id":"e7211eeb-a5e5-48f7-be81-dfd2d5bcb94b",
            "score":26.301422}]
      },
      "highlighting":{
        "f1891f87-cc7e-4b35-8d40-213b5d595e09":{
          "titre_lemmatized":["Kaip elgtis įkandus <em>gyvatei</em>?"],
          "contenu":["", sukurti iš <em>gyvačių</em> nuodų, buvo naudojami padidėjusiam kraujospūdžiui gydyti.\nTepalai, į kurių sudėtį įeina <em>gyvačių</em> nuodų, puikiai malšina sąnarių ir raumenų skausmus. Tačiau gydymą <em>gyvačių</em> nuodais arba jų turinčiais medikamentais turi skirti tik gydytojas.\nĮdomu tai, kad kremai su <em>gyvačių</em> nuodais gali pakeisti"],
          "contenu_lemmatized":[" Maišelienę prieš šešetą metų įkando vienintelė Lietuvoje gyvenanti nuodinga <em>gyvatė</em> – <em>angis</em>. Datos, kada tai"]},
        "4fb80202-bb3f-4ac5-94d9-fb13df15f2a2":{
          "titre_lemmatized":["Lietuvos bioįvairovė. Paprastoji <em>angis</em>"],
          "contenu":["", šių <em>gyvačių</em> labai sumažėjo, vietomis tapo retos ar net išnyko.\nGyvatė atsiveda jauniklius rugpjūčio"],
          "contenu_lemmatized":[". <em>Gyvatė</em> gelia dantimis, o kaišiodama dvišaką liežuvį uodžia kvapus.\nPaprastosios <em>angies</em> priešai gamtoje"]},

    ...

        "e7211eeb-a5e5-48f7-be81-dfd2d5bcb94b":{
          "contenu_lemmatized":[", stabdo kraujavimą, gangreną, padeda įkandus <em>gyvatei</em>, apalpus ir netgi gydo auglius – fibromas ir kitus"]}},
      "spellcheck":{
        "suggestions":[],
        "correctlySpelled":true,
        "collations":[]}}
```


2.3.1. SOLR serviso būsenos patikrinimas

Terminaliniame lange pateikiame būsenos testavimo užklausą:

```
curl –I http://158.129.51.163:8983/solr/#/
```

Turime gauti atsakymą su HTTP būsenos kodu:

```
HTTP/1.1 200 OK.
```

Jeigu tokios būsenos nepavyksta gauti, lokaliame kompiuteryje turėdami administratoriaus teises terminaliniame lange turite restartuoti SOLR servisą:

```
service solr restart
```

2.3.2. SOLR branduolio Semantika2 veikimo patikrinimas

Terminaliniame lange pateikiame testinę užklausą:

```
curl "http://158.129.51.163:8983/solr/Semantika2/select?q=gyva%C4%8Di%C5%B3"
```

Gaunamame atsakyme turi būti tokie fragmentai:


```
{
  "responseHeader":{
```

```
        "status":0,
        "QTime":165,
        "params":{
          "q":"gyvačių"}},
      "response":{"numFound":53,"start":0,"maxScore":857.4379,"docs":[
          {
            "estampille_temporelle":"2019-03-18T23:12:24Z",
            "titre":"Kaip elgtis įkandus gyvatei?",
            "url":"https://www.delfi.lt/sveikata/sveikatos-naujienos/kaip-elgtis-ikandus-
gyvatei.d?id=65175291",
            "id":"f1891f87-cc7e-4b35-8d40-213b5d595e09",
            "score":857.4379},
          {
            "estampille_temporelle":"2019-03-18T23:12:48Z",
            "titre":"Lietuvos bioįvairovė. Paprastoji angis",
            "url":"https://www.delfi.lt/grynas/gamta/lietuvos-bioivairove-paprastoji-
angis.d?id=60187283",
            "id":"4fb80202-bb3f-4ac5-94d9-fb13df15f2a2",
            "score":857.4379},

    ...

          {
            "estampille_temporelle":"2019-03-19T12:43:40Z",
            "titre":"Obuolių actas - nuo antsvorio ir ligų",
            "url":"https://www.delfi.lt/gyvenimas/grozis_ir_sveikata/obuoliu-actas-nuo-antsvorio-
ir-ligu.d?id=61924689",
            "id":"e7211eeb-a5e5-48f7-be81-dfd2d5bcb94b",
            "score":26.301422}]
      },
      "highlighting":{
        "f1891f87-cc7e-4b35-8d40-213b5d595e09":{
          "titre_lemmatized":["Kaip elgtis įkandus <em>gyvatei</em>?"],
          "contenu":["", sukurti iš <em>gyvačių</em> nuodų, buvo naudojami padidėjusiam
kraujospūdžiui gydyti.\nTepalai, į kurių sudėtį įeina <em>gyvačių</em> nuodų, puikiai malšina
sąnarių ir raumenų skausmus. Tačiau gydymą <em>gyvačių</em> nuodais arba jų turinčiais
medikamentais turi skirti tik gydytojas.\nĮdomu tai, kad kremai su <em>gyvačių</em> nuodais gali
pakeisti"],
          "contenu_lemmatized":[" Maišelienę prieš šešetą metų įkando vienintelė Lietuvoje
gyvenanti nuodinga <em>gyvatė</em> – <em>angis</em>. Datos, kada tai"]},
        "4fb80202-bb3f-4ac5-94d9-fb13df15f2a2":{
          "titre_lemmatized":["Lietuvos bioįvairovė. Paprastoji <em>angis</em>"],
          "contenu":["", šių <em>gyvačių</em> labai sumažėjo, vietomis tapo retos ar net
išnyko.\nGyvatė atsiveda jauniklius rugpjūčio"],
          "contenu_lemmatized":[". <em>Gyvatė</em> gelia dantimis, o kaišiodama dvišaką liežuvį
uodžia kvapus.\nPaprastosios <em>angies</em> priešai gamtoje"]},

    ...

        "e7211eeb-a5e5-48f7-be81-dfd2d5bcb94b":{
          "contenu_lemmatized":["", stabdo kraujavimą, gangreną, padeda įkandus <em>gyvatei</em>,
apalpus ir netgi gydo auglius – fibromas ir kitus"]}},
```

```
"spellcheck":{
  "suggestions":[],
  "correctlySpelled":true,
  "collations":[]}}
```


2.4. Duomenų indeksavimas ir atnaujinimas

Gautieji saugyklos dokumentai JSON formatu buvo indeksuojami iš terminalinio lango, pvz.:

sudo curl 'http://158.129.51.163:8983/solr/Semantika2/update/json/docs'
'?split=/'
'&f=id:/_id'
'&f=contenu:/body'
'&f=titre:/title'
'&f=estampille_temporelle:/date'
'&f=url:/url'
-H 'Content-type:application/json' --data-binary "@/home/virginijus/indeksavimui_skirti_failai/export1.json"

Naujesnis dokumentas automatiškai pakeičia senesnę dokumento versiją su tuo pačiu ID naujausio indeksavimo metu.

2.5. Konteinerizavimo scenarijus (Dockerfile)

2.5.1. SOLR konteinerio sukūrimas

Darbiniame aplanke padedame čia žemiau aprašytą **Dockerfile**. Jame startuojame nuo **SOLR 8.1** versijos. Modifikuojame lietuvišką kamieninimą pakeisdami failą **lucene-analyzers-common-8.1.1.jar** į failą su patobulintu lietuvišku kamieninimu **lucene-analyzers-common-8.1.0.jar**, esantį aplanke **./jared**. Į atvaizdą įkeliame ***tuščią*** branduolį ***Semantika2*** (su tuščiu indeksu), padėtą aplanke **./sansindexconfig**.

Kuriant *darbinį atvaizdą* gali būti naudojamas toks **Dockerfile** tekstas :

```
FROM solr:8.1
WORKDIR /var/solr/data/
ARG USER_ID=1000
ARG GROUP_ID=1000
ARG HOME_DIR=/var/solr/data
USER root
RUN rm /opt/solr-8.1.1/server/solr-webapp/webapp/WEB-INF/lib/lucene-analyzers-common-8.1.1.jar
COPY ./jared/lucene-analyzers-common-8.1.0.jar /opt/solr-8.1.1/server/solr-webapp/webapp/WEB-INF/lib/
COPY --chown=solr:solr ./sansindexconfig /var/solr/data
RUN chown -R ${USER_ID}:${GROUP_ID} ${HOME_DIR}
USER solr
```

Sukuriame SOLR *atvaizdą* **solr_image_ready_core:8.15** iš darbinio aplanko vykdydami tokią komandą:

docker build -t solr_image_ready_core:8.15 --build-arg USER_ID=$(id -u) --build-arg GROUP_ID=$(id -g) .

Šis *atvaizdas* naudojamas *konteinerio* sukūrimui.

Sukuriame SOLR *darbinį konteinerį* **jupiter-solare_sema2**, kurio branduolys ***Semantika2*** su indeksu saugomas išorinėje vardinėje laikmenoje (*named volume*), šiame pavyzdyje – **jupiteris**. Pasirnkame prievadą, per kurį bus komunikuojama su konteineriu, šiame pavyzdyje – **8999**. Iš darbinio aplanko vykdydama tokia komanda:

docker run -d --name jupiter-solare_sema2 -p 8999:8983 -v jupiteris:/var/solr/data solr_image_ready_core:8.15

Su aplinka konteineris komunikuoja per prievadą (*port*) **8999**. SOLR konteineris paruoštas darbui. Konteinerio funkcionalumą patikriname per **SOLR Admin**, kurį pasijungiame per http://xxx.xxx.xx.xxx:8999 .

2.5.2. Konteinerizuoto komponento naudojimas

Komponento konteinerinis variantas gali būti sukonfigūruotas darbui naudojant kitokį nei 8983 prievado numerį.Pvz., aukščiau aprašytame pavyzdyje naudojamas prievadas 8999. Konteineris valdomas standartinėmis docker komandomis, pvz., jeigu konteineris yra nepaleistas, jį paleidžiame iš terminalo vykdydami tokią komandą:

$ docker start jupiter-solare_sema2

Paieška gali būti vykdoma iš terminalinio lango formuojant užklausas (žr. pavyzdžius aukščiau), pasijungus **SOLR Admin** (http://xxx.xxx.xx.xxx:8999) arba komunikuojant užklausomis su konteineriu iš turinio valdymo sistemos ar kitokios vartotojo sąsajos.

2.5.3. Indeksavimo procedūra

Pavyzdys. Sakykim, naudodami ką tik sukurtą SOLR konteinerį (sukūrimo procedūra aprašyta aukščiau), norime suindeksuoti failą **small.json** su 727 dokumentų individualizuotais JSON duomenimis (*custom JSON data*), padėtą *lokalinio* kompiuterio aplanke, pvz. **/home/vardenis/indeksuojamifailai/**. Turi būti vykdoma tokia komanda:

sudo curl 'http://localhost:*8999*/solr/Semantika2/update/json/docs'

'?split=/'

'&f=id:/_id'

'&f=contenu:/body'

'&f=titre:/title'

'&f=estampille_temporelle:/date'

'&f=url:/url'

-H 'Content-type:application/json' --data-binary "@/home/vardenis/indeksuojamifailai/small.json"

Kaip buvo aukščiau minėta, šiame pavyzdyje ką tik sukurtas 727 dokumentų indeksas yra saugomas išorinėje vardinėje laikmenoje **jupiteris**, todėl yra tvarus (*persistent*), t.y. sunaikinus konteinerį **jupiter-solare_sema2**, SOLR branduolys ***Semantika2*** su indeksavimo duomenimis visą laiką išlieka išorinėje vardinėje laikmenoje **jupiteris**, esančioje toje pačioje mašinoje, kurioje buvo sunaikintasis konteineris. Atkūrus konteinerį, išlikęs indeksas gali būti naudojamas ir pildomas vardinėje laikmenoje **jupiteris**. Paieškos funkcionalumą ką tik sukurtame indekse galime patikrinti naudodami **SOLR Admin**.