Lietuvių kalbos teksto sintaksinės-semantinės analizės informacinė sistema

LKSSAIS vystymas

Modernizuojamos LKSSAIS interneto lietuviškų socialinių tekstų kalbos normalizavimo komponento administratoriaus instrukcija

8 puslapiai

2020

TURINYS

LENTELIŲ SĄRAŠAS

1. Įžanga

1.1. Dokumento paskirtis

Šio dokumento paskirtis – pateikti modernizuojamos LKSSAIS Interneto lietuviškų socialinių tekstų kalbos normalizavimo 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.	Interneto lietuviškų socialinių tekstų kalbos normalizavimo komponento detalios analizės ir projektavimo specifikacija
2.	Lietuvių kalbos sintaksinės ir semantinės analizės informacinės sistemos techninė užduotis

1.4. Interneto lietuviškų socialinių tekstų normalizavimo įrankio prototipo paskirtis ir tikslai

LKSSAIS interneto lietuviškų socialinių tekstų normalizavimo įrankio prototipo pagrindinis tikslas yra atlikti teksto normalizavimo procesą. Šis komponentas turi gebėti automatizuotai kiek galima labiau pertvarkyti socialiniams tekstams būdingą netaisyklingą tekstą pagal lietuvių kalbos norminius reikalavimus. Netaisyklingumas čia gali būti suprantamas kaip dažnai kartojamos rašybos klaidos, raidžių be diakritinių ženklų vartojimas („šveplas“ tekstas), neteiktinų žodžių („balkis“, „arenda“ ir pan.), slengo bei įžeidžiančios leksikos vartojimas. Tradiciniais būdais analizuojant tokius tekstus morfologiškai, sintaksiškai ar semantiškai, neišvengiamai gausime prastos kokybės rezultatus, nes tokio tipo žodžiai bus traktuojami kaip „neatpažinti“, o didelis jų kiekis stabdys numatytąją analizės eigą. Šio komponento paskirtis – automatiškai pakeisti nenorminius žodžius į jų artimiausius atitikmenis, kurie toliau sėkmingai galėtų dalyvauti teksto analizės procesuose. Komponento prototipas turi naudoti jam pateiktą teksto segmentavimo informaciją arba, kai tokia nepateikiama, turėti autonomiškai veikiantį teksto dalintuvą į žodžius.

2. Interneto lietuviškų socialinių tekstų normalizavimo įrankio prototipo administratoriaus vadovas

2.1. Interneto lietuviškų socialinių tekstų normalizavimo įrankio sandara

Interneto lietuviškų socialinių tekstų normalizavimo įrankį sudaro:

1. Linux terpei skirtas vykdomasis serviso failas
`/usr/local/bin/norma`

2. Nustatymų ir veikimo režimų failas
`/etc/norma.ini`

3. Norminės, submorminės ir necenzūrinės lietuvių kalbos morfologijos Hunspell failai:
`/usr/local/share/hs_dictionaries/lt-LT.dic`
`/usr/local/share/hs_dictionaries/lt-LT.aff`
`/usr/local/share/hs_dictionaries/lt-LT_plus_substandard.dic`
`/usr/local/share/hs_dictionaries/lt-LT_plus_substandard.aff`
`/usr/local/share/hs_dictionaries/lt-LT_plus_obscene.dic`
`/usr/local/share/hs_dictionaries/lt-LT_plus_obscene.aff`

4. Dažninis lietuvių kalbos žodynas:
`/usr/local/share/hs_dictionaries/word_list.txt`

5. Protokolavimo failas
`/var/log/norma.log`

6. Automatinio serviso paleidimo failas
`/etc/init.d/norma`

7. Išeities tekstai kataloge
`/home/fotonija/norma/src`

2.2. Interneto lietuviškų socialinių tekstų normalizavimo įrankio diegimas į sistemą

1. Sukompiliuojam vykdomąjį failą komanda `make` būnant kataloge `/home/fotonija/norma`

2. Vykdomąjį failą nukopijuojam į katalogą `/usr/local/bin`

3. Sukuriam nustatymų failą `norma.ini` ir nukopijuojam jį į katalogą `/etc`

`norma.ini` failo pavyzdys su save paaiškinančiais parametrų pavadinimais:

```
[listener]
;host=158.129.51.163
port=8091
minThreads=4
maxThreads=100
cleanupInterval=60000
readTimeout=60000
maxRequestSize=16000
maxMultiPartSize=10000000
pathDic=/usr/local/share/hs_dictionaries/lt-LT.dic
pathAff=/usr/local/share/hs_dictionaries/lt-LT.aff
pathDicSlang=/usr/local/share/hs_dictionaries/lt-LT_plus_obscene.dic
pathAffSlang=/usr/local/share/hs_dictionaries/lt-LT_plus_obscene.aff
pathDicUnapproved=/usr/local/share/hs_dictionaries/lt-LT_plus_substandard.dic
pathAffUnapproved=/usr/local/share/hs_dictionaries/lt-LT_plus_substandard.aff
pathFrequency=/usr/local/share/hs_dictionaries/word_list.txt
rateStandart=1.5
rateDiacritic=1.5
rateSlang=1.5
rateUnapproved=1.5

[logging]
fileName=/var/log/norma.log
minLevel=1
bufferSize=100
maxSize=1000000
maxBackups=2
timestampFormat=dd.MM.yyyy hh:mm:ss.zzz
msgFormat={timestamp} {typeNr} {type} {thread} {msg}
; QT5 supports: msgFormat={timestamp} {typeNr} {type} {thread} {msg}\n  in {file} line {line}
function {function}
```

4. Hunspell failus ir dažninį žodyną iš katalogo `/home/fotonija/naujienos` kopijuojam į katalogą `/usr/local/share/hs_dictionaries`

5. Sukuriam automatinio paleidimo failą `/etc/init.d/norma` , pvz.:

```
#! /bin/sh
### BEGIN INIT INFO
# Provides: norma
# Required-Start: $syslog
# Required-Stop: $syslog
# Default-Start: 2 3 4 5
# Default-Stop: 0 1 6
# Short-Description: Normalization service
# Description: This file starts and stops Normalization service
#
### END INIT INFO

START_DIR=/usr/local/bin

case "$1" in
 start)
   $START_DIR/norma
   ;;
 stop)
   $START_DIR/norma -t
   sleep 1
   ;;
 restart)
   $START_DIR/norma -t
   sleep 2
   $START_DIR/norma
   ;;
 *)
   echo "Usage: norma {start|stop|restart}" >&2
   exit 3
   ;;
esac
```

6. Paleidžiam / stabdom / perstartuojam servisą:

```
sudo service norma start|stop|restart
```

2.3. Interneto lietuviškų socialinių tekstų normalizavimo įrankio veikimas

2.3.1. Įėjimas (:8091/normalization)

Komponento įėjimas turi būti JSON dokumentas dviem galimais formatais.

1. Autonominis veikimas, nereikalaujantis jokio išankstinio teksto apdorojimo (dalijimo į žodžius, sakinius ir t.t.):

```
{
   "body":"Salia kelio karciamoj mockrusys nusiciaudejo",
   "normalization_mode":"+obscene"
}
```

`"normalization_mode"` galimos reikšmės yra `"standard"`, `"+diacritic"` (jos galima ir nenurodyti, tada tokia reikšmė suveikia automatiškai), `"+substandard"` ir `"+obscene"`.

`"standard"` – tik tipiškų rašybos klaidų koregavimas;
`"+diacritic"` – tipiškų rašybos klaidų koregavimas ir diakritikų atstatymas;
`"+substandard"` – tipiškų rašybos klaidų koregavimas ir diakritikų atstatymas taip pat ir subnorminėje leksikoje;
`"+obscene"` – tipiškų rašybos klaidų koregavimas ir diakritikų atstatymas taip pat ir subnorminėje bei necenzūrinėje leksikoje.

2. Veikimas tekstų apdorojimo grandinėlėje (turi būti paduota dalinimo į žodžius informacija):

```
{
   "normalization_mode":"+obscene",
   "body":"Salia kelio karciamoj mockrusys nusiciaudejo",
   "annotations":{
   "lex":{"seg":[[0,5],[6,5],[12,9],[22,9],[32,12]],"s":[[0,44]],"p":[[0,4
4]]}
   }
}
```

Teksto segmentavimą galima atlikti specializuotu Semantika2 servisu, pvz., pateikę užklausą:

```
curl -X "POST" "http://nagys.vdu.lt:7007" \
     -H 'Content-Type: text/plain' \
     -d "Svečiuose turistas pasakoja apie savo kelionę Afrikos dykumoje."
```

gauname į žodžius, sakinius ir pastraipas sudalinimo atsakymą:

```
{"seg":[[0,9],[10,8],[19,8],[28,4],[33,4],[38,7],[46,7],[54,8],[62,1]],"s"
:[[0,63]],"p":[[0,63]]}
```

2.3.2. Išėjimas

Klaidos atveju komponentas grąžina HTTP atsakymą ne 200 bei trumpą klaidos paaiškinimą, pvz.:

```
HTTP/1.1 204 no found key 'body'
Connection: close
Content-Type: text/html; charset=UTF-8
```

Komponento išėjimas abiem kvietimo atvejais yra JSON dokumentas su normalizuotu tekstu (gali skirtis dalinimas į žodžius):

```
{
```

```
    "normalised_body": "Šalia kelio karčiamoj močkrušys nusičiaudėjo"
}
```

2.4. Interneto lietuviškų socialinių tekstų normalizavimo konteinerizavimas

2.4.1. Konteinerizavimo scenarijus (Dockerfile)

Komponentas paruoštas darbui „cloud-ready“ konteinerinėje Docker platformoje naudojant tokį `Dockerfile` scenarijų:

```
FROM ubuntu:18.04
RUN apt-get update \
    && apt-get install -y \
        g++ \
        qt5-default \
        qtbase5-dev \
        qttools5-dev
COPY norma /usr/bin/norma
COPY etc/norma.ini /etc/norma.ini
COPY hs_dictionaries/lt-LT.dic /usr/local/share/hs_dictionaries/lt-LT.dic
COPY hs_dictionaries/lt-LT.aff /usr/local/share/hs_dictionaries/lt-LT.aff
COPY hs_dictionaries/lt-LT_plus_obscene.dic /usr/local/share/hs_dictionaries/lt-LT_plus_obscene.dic
COPY hs_dictionaries/lt-LT_plus_obscene.aff /usr/local/share/hs_dictionaries/lt-LT_plus_obscene.aff
COPY hs_dictionaries/lt-LT_plus_substandard.dic /usr/local/share/hs_dictionaries/lt-LT_plus_substandard.dic
COPY hs_dictionaries/lt-LT_plus_substandard.aff /usr/local/share/hs_dictionaries/lt-LT_plus_substandard.aff
COPY hs_dictionaries/word_list.txt /usr/local/share/hs_dictionaries/word_list.txt
CMD ["/usr/bin/norma", "-e"]
EXPOSE 8091
```

2.4.2. Konteinerizuoto komponento naudojimas

Komponento konteinerinis variantas gali būti sukonfigūruotas darbui naudojant kitokį nei 8091 prievado numerį, todėl, pavyzdžiui, gali būti kviečiamas kiek kitaip:

```
curl -X "POST" "http://158.129.51.163:31474/normalization" \
     -H 'Content-Type: application/json' \
     -d $'{
  "normalization_mode": "+diacritic",
  "body": "Pirmadieni, kai ten gryzo perzidente Dalia Grybauskaite, dar stafomos ligonines koridoriuosia ir aplinkoje vaikstinejo darbinykai."
}'
```