VYTAUTO
DIDŽIOJO
UNIVERSITETAS
M C M X X I I

Modernizuojamos LKSSAIS lietuvių kalbos teksto morfologinės analizės komponento administratoriaus instrukcija

2020

TURINYS

1. Įžanga 3

1.1. Dokumento paskirtis 3

1.2. Sutrumpinimai 3

*1.3. Panaudotų dokumentų sąrašas ..... **Error! Bookmark not defined.***

1.4. Lietuvių kalbos teksto morfologinės analizės įrankio prototipo paskirtis ir tikslai 3

2. Lietuvių kalbos teksto morfologinės analizės įrankio prototipo valdymas 4

2.1. Lietuvių kalbos teksto morfologinės analizės įrankio sandara 4

2.2. Lietuvių kalbos teksto morfologinės analizės įrankio diegimas į sistemą 5

2.3. Lietuvių kalbos teksto morfologinės analizės įrankio veikimas 6

2.3.1. Įėjimas (:8090/morphology) 6

2.3.2. Išėjimas 6

2.4. Lietuvių kalbos teksto morfologinės analizės įrankio konteinerizavimas 8

2.4.1. Konteinerizavimo scenarijus (Dockerfile) 8

2.4.2. Konteinerizuoto komponento naudojimas 9

3. Dokumento istorija ..... Error! Bookmark not defined.

1. Įžanga

1.1. Dokumento paskirtis

Šio dokumento paskirtis – pateikti modernizuojamos LKSSAIS Lietuvių kalbos teksto morfologinės analizės 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. Lietuvių kalbos teksto morfologinės analizės įrankio prototipo paskirtis ir tikslai

LKSSAIS lietuvių kalbos morfologinės analizės įrankio pagrindiniu tikslu yra atlikti teksto morfologinę analizę. Komponentas įgyvendina šiuos uždavinius ir reikalavimus:

- Lietuvių k. morfologinės analizės įrankis nustato kiekvieno analizuojamo teksto lietuvių kalbos žodžio ar samplaikos morfologinę informaciją.
- Įrankis anotuoja tekstą, pažymėdamas žodžių morfologinę informaciją (kalbos dalį ir kalbos dalių morfologines kategorijas) ir nežodinę informaciją (skyrybos ženklus, skaičius, simbolius ir kt.).
- Įrankis siūlo pagal kontekstą ir statistiką labiausiai tikėtiną variantą esant daugiareikšmiam analizės rezultatui.
- Komponentas vystosi tobulinant Semantika 1 sukurtą analogą bei remiantis jo kūrimo metu parengta dokumentacija.
- Įrankio funkcijų vykdymas užtikrinamas lietuvių k. morfologinio žodyno duomenų baze.
- Komponentas yra kuriamas remiantis jau sukurtais LKSSAIS komponentais ir prototipais.
- Komponentas sugeba analizuoti *UTF-8* arba lygiavertės koduotės tekstus.
- Komponentas sugeba analizuoti pateiktus *TEI P5* ar lygiavertės struktūros dokumentu.
- Komponentas sugeba gražinti analizės rezultatą (teksto anotacijas) *TEI P5* ar lygiavertės struktūros *JSON* ar lygiaverčiu formatu.
- Komponento duomenų išvesties ir įvesties formatai remiasi architektūrinės LKSSAIS dokumentacijos specifikacija.
- Komponentas užtikrina ne mažesnį nei 95% tikslumo (angl. *precision*) ir 95% išsamumo (angl. *recall*) žodžio pagrindinės formos ir morfologinių kategorijų nustatymo rodiklius

2. Lietuvių kalbos teksto morfologinės analizės įrankio prototipo valdymas

2.1. Lietuvių kalbos teksto morfologinės analizės įrankio sandara

Lietuvių kalbos teksto morfologinės analizės įrankį sudaro:

1. Linux terpei skirtas vykdomasis serviso failas

`/usr/local/bin/hmorphd2`

2. Nustatymų ir veikimo režimų failas

`/etc/hmorph2.ini`

3. Duomenų failai:

`/usr/local/share/hs_dictionaries/lt-LT_morphology.aff`

`/usr/local/share/hs_dictionaries/lt-LT_morphology.dic`

`/usr/local/share/hs_dictionaries/model_1.dat`

`/usr/local/share/hs_dictionaries/model_10.dat`

`/usr/local/share/hs_dictionaries/model_11.dat`

`/usr/local/share/hs_dictionaries/model_12.dat`

`/usr/local/share/hs_dictionaries/model_13.dat`

`/usr/local/share/hs_dictionaries/model_14.dat`

`/usr/local/share/hs_dictionaries/model_15.dat`

`/usr/local/share/hs_dictionaries/model_2.dat`

`/usr/local/share/hs_dictionaries/model_3.dat`

`/usr/local/share/hs_dictionaries/model_4.dat`

`/usr/local/share/hs_dictionaries/model_5.dat`

`/usr/local/share/hs_dictionaries/model_6.dat`

`/usr/local/share/hs_dictionaries/model_7.dat`

`/usr/local/share/hs_dictionaries/model_8.dat`

`/usr/local/share/hs_dictionaries/model_9.dat`

`/usr/local/share/hs_dictionaries/stabilios_frazes.dat`

5. Protokolavimo failas

`/var/log/hmorphd2.log`

6. Automatinio serviso paleidimo failas

`/etc/init.d/hmorphd2`

7. Išeities tekstai kataloge

`/home/fotonija/hmorphd2/src`

2.2. Lietuvių kalbos teksto morfologinės analizės įrankio diegimas į sistemą

1. Sukompiliuojam vykdomąjį failą komanda `make` būnant kataloge `/home/fotonija/hmorphd2`

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

3. Sukuriam nustatymų failą `hmorph2.ini` ir nukopijuojam jį į katalogą `/etc`

`hmorph2.ini` failo pavyzdys su save paaiškinančiais parametrų pavadinimais:

```
[listener]
;host=158.129.51.163
port=8090
minThreads=4
maxThreads=16
cleanupInterval=60000
readTimeout=60000
maxRequestSize=1024000
maxMultiPartSize=10000000
pathDic=/usr/local/share/hs_dictionaries/lt-LT_morphology.dic
pathAff=/usr/local/share/hs_dictionaries/lt-LT_morphology.aff
pathStat=/usr/local/share/hs_dictionaries/stat.dat
UseModel = true
model1=/usr/local/share/hs_dictionaries/model_1.dat
model2=/usr/local/share/hs_dictionaries/model_2.dat
model3=/usr/local/share/hs_dictionaries/model_3.dat
model4=/usr/local/share/hs_dictionaries/model_4.dat
model5=/usr/local/share/hs_dictionaries/model_5.dat
model6=/usr/local/share/hs_dictionaries/model_6.dat
model7=/usr/local/share/hs_dictionaries/model_7.dat
model8=/usr/local/share/hs_dictionaries/model_8.dat
model9=/usr/local/share/hs_dictionaries/model_9.dat
model10=/usr/local/share/hs_dictionaries/model_10.dat
model11=/usr/local/share/hs_dictionaries/model_11.dat
model12=/usr/local/share/hs_dictionaries/model_12.dat
model13=/usr/local/share/hs_dictionaries/model_13.dat
model14=/usr/local/share/hs_dictionaries/model_14.dat
model15=/usr/local/share/hs_dictionaries/model_15.dat
StablePhrasePath=/usr/local/share/hs_dictionaries/stabilios_frazes.dat

[logging]
fileName=/var/log/hmorphd2.log
minLevel=1
bufferSize=100
maxSize=100000
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/hmorphd2` , pvz.:

```
#! /bin/sh
### BEGIN INIT INFO
# Provides: hmorphd2
# Required-Start: $syslog
# Required-Stop: $syslog
# Default-Start: 2 3 4 5
# Default-Stop: 0 1 6
# Short-Description: Morphology service
# Description: This file starts and stops Morphology service
#
### END INIT INFO

START_DIR=/usr/local/bin

case "$1" in
 start)
   $START_DIR/hmorphd2
   ;;
 stop)
   $START_DIR/hmorphd2 -t
   sleep 1
```

```
    ;;
  restart)
    $START_DIR/hmorphd2 -t
    sleep 2
    $START_DIR/hmorphd2
    ;;
  *)
    echo "Usage: hmorphd2 {start|stop|restart}" >&2
    exit 3
    ;;
esac
```

6. Paleidžiam / stabdom / perstartuojam servisą:

```
sudo service hmorphd2 start|stop|restart
```

2.3. Lietuvių kalbos teksto morfologinės analizės įrankio veikimas

2.3.1. Įėjimas (:8090/morphology)

Komponento įėjimas turi būti JSON dokumentas dviem galimais formatais.

1. Įprastinis veikimas tekstų apdorojimo grandinėlėje, kai norima gauti pilną morfologinės analizės ir kamieninimo rezultatą:

```
{
 "body":"Penki kandidatai vakare susirungs BBC televizijos debatuose.",
 "annotations":{
 "lex":{"seg":[[0,5],[6,10],[17,6],[24,9],[34,3],[38,11],[50,9],[59,1]],"
s":[[0,60]],"p":[[0,60]]}
 }
}
```

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. Parametrinis veikimas, reguliuojant pateikiamų rezultatų apimtį:

```
{
   "scope":"lemmas",
   "body":"Penki kandidatai vakare susirungs BBC televizijos debatuose.",
   "annotations":{
   "lex":{"seg":[[0,5],[6,10],[17,6],[24,9],[34,3],[38,11],[50,9],[59,1]],
"s":[[0,60]],"p":[[0,60]]}
   }
}
```

Galimos `"scope"` reikšmės yra `"all"` (grąžinami visi analizės rezultatai), `"morphology"` (grąžinami tik morfologinės analizės rezultatai), `"stems"` (grąžinami tiktai kamienai), `"lemmas"` (grąžinamos tiktai labiausiai tikėtinos lemos), `"all_lemmas"` (grąžinamos visos lemos).

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
```

Įprastiniu atveju komponentas grąžina pagal tikėtinumą surūšiuotus morfologinės analizės rezultatus (tikėtiniausias – pirmas) ir teksto kamienų sąrašą:


```
{
    "msd": [
        [
            [
                "penki",
                "Mcm-nl-"
            ],
            [
                "penki",
                "Mcm-vl-"
            ]
        ],
        [
            [
                "kandidatas",
                "Ncmpnn-"
            ],
            [
                "kandidatas",
                "Ncmpvn-"
            ]
        ],
        [
            [
                "vakaras",
                "Ncmsln-"
            ],
            [
                "vakaris",
                "Agpfsin"
            ],
            [
                "vakaras",
                "Ncmsvn-"
            ],
            [
                "vakaris",
                "Agpfsvn"
            ]
        ],
        [
            [
                "susirungti",
                "Vgmf3---n--yi-"
            ]
        ],
        [
            [
                "BBC",
                "Ya"
            ]
        ],
        [
            [
                "televizija",
                "Ncfsgn-"
            ],
            [
                "televizija",
                "Ncfpnn-"
            ],
            [
                "televizija",
                "Ncfpvn-"
            ]
```

```
            ],
            [
                [
                    "debatai",
                    "Ncmpln-"
                ]
            ],
            [
                [
                    ".",
                    "Tp"
                ]
            ]
        ],
        "stem": [
            "Penk",
            "kandidat",
            "vakar",
            "susirung",
            "BBC",
            "televizij",
            "debat",
            "."
        ]
    }
```

Parametrizuoto veikimo atveju grąžinami tik ta informacija, kurios buvo reikalaujama, pvz., jeigu buvo nurodytas parametras `"scope":"lemmas"`, tai atsakymas bus sudarytas tik iš labiausiai tikėtinų lemų:

```
{
    "lemmas": [
        "penki",
        "kandidatas",
        "vakaras",
        "susirungti",
        "BBC",
        "televizija",
        "debatai",
        "."
    ]
}
```

2.4. Lietuvių kalbos teksto morfologinės analizės įrankio 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 hmorphd2 /usr/bin/hmorphd2
COPY etc/hmorph2.ini /etc/hmorph2.ini
COPY hs_dictionaries/lt-LT_morphology.dic /usr/local/share/hs_dictionaries/lt-LT_morphology.dic
COPY hs_dictionaries/lt-LT_morphology.aff /usr/local/share/hs_dictionaries/lt-LT_morphology.aff
COPY hs_dictionaries/stat.dat /usr/local/share/hs_dictionaries/stat.dat
COPY hs_dictionaries/model_1.dat /usr/local/share/hs_dictionaries/model_1.dat
COPY hs_dictionaries/model_2.dat /usr/local/share/hs_dictionaries/model_2.dat
COPY hs_dictionaries/model_3.dat /usr/local/share/hs_dictionaries/model_3.dat
COPY hs_dictionaries/model_4.dat /usr/local/share/hs_dictionaries/model_4.dat
COPY hs_dictionaries/model_5.dat /usr/local/share/hs_dictionaries/model_5.dat
COPY hs_dictionaries/model_6.dat /usr/local/share/hs_dictionaries/model_6.dat
COPY hs_dictionaries/model_7.dat /usr/local/share/hs_dictionaries/model_7.dat
COPY hs_dictionaries/model_8.dat /usr/local/share/hs_dictionaries/model_8.dat
COPY hs_dictionaries/model_9.dat /usr/local/share/hs_dictionaries/model_9.dat
COPY hs_dictionaries/model_10.dat /usr/local/share/hs_dictionaries/model_10.dat
COPY hs_dictionaries/model_11.dat /usr/local/share/hs_dictionaries/model_11.dat
```

```
COPY hs_dictionaries/model_12.dat /usr/local/share/hs_dictionaries/model_12.dat
COPY hs_dictionaries/model_13.dat /usr/local/share/hs_dictionaries/model_13.dat
COPY hs_dictionaries/model_14.dat /usr/local/share/hs_dictionaries/model_14.dat
COPY hs_dictionaries/model_15.dat /usr/local/share/hs_dictionaries/model_15.dat
COPY hs_dictionaries/stabilios_frazes.dat /usr/local/share/hs_dictionaries/stabilios_frazes.dat
CMD ["/usr/bin/hmorphd2", "-e"]
EXPOSE 8090
```

2.4.2. Konteinerizuoto komponento naudojimas

Komponento konteinerinis variantas gali būti sukonfigūruotas darbui naudojant kitokį nei 8090 prievado numerį, todėl, pavyzdžiui, gali būti kviečiamas kiek kitaip:

```
curl -X "POST" "http://158.129.51.163:31473/morphology" \
-H 'Content-Type: application/json' \
-d $'{"body":"Penki kandidatai vakare susirungs BBC televizijos
debatuose.","annotations":{"lex":{"seg":[[0,5],[6,10],[17,6],[24,9],[34,3],[38,11],[50,9],[59,1]],
"s":[[0,60]],"p":[[0,60]]}}}'
```