Vybrané nástroje na příkazové řádce

Jana Straková
ÚFAL MFF UK

Výběr nástrojů pro analýzu textu

- UDPipe: nástroj pro analýzu textu
  - 60 jazyků
  - rozdělí text na věty, slova
  - provede morfologickou analýzu - slovní druhy, pády, rody, vidy, ...
  - provede lemmatizaci - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
  - provede závislostní analýzu - gramatika (podmět, přísudek, přívlastky, …)
  - http://ufal.mff.cuni.cz/udpipe

Výběr nástrojů pro analýzu textu

- UDPipe: nástroj pro analýzu textu
  - 60 jazyků
  - rozdělí text na věty, slova
  - provede morfologickou analýzu - slovní druhy, pády, rody, vidy, ...
  - provede lemmatizaci - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
  - provede závislostní analýzu - gramatika (podmět, přísudek, přívlastky, …)
  - http://ufal.mff.cuni.cz/udpipe
- NameTag: hledá vlastní jména v textu (“Praha”, “Jan Novák”, …)
  - http://ufal.mff.cuni.cz/nametag

Výběr nástrojů pro analýzu textu

- UDPipe: nástroj pro analýzu textu
  - 60 jazyků
  - rozdělí text na věty, slova
  - provede morfologickou analýzu - slovní druhy, pády, rody, vidy, ...
  - provede lemmatizaci - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
  - provede závislostní analýzu - gramatika (podmět, přísudek, přívlastky, …)
  - http://ufal.mff.cuni.cz/udpipe
- NameTag: hledá vlastní jména v textu (“Praha”, “Jan Novák”, …)
  - http://ufal.mff.cuni.cz/nametag
- Korektor: spellchecker a oprava gramatiky
  - http://ufal.mff.cuni.cz/korektor

Výběr nástrojů pro analýzu textu

- UDPipe: nástroj pro analýzu textu
  - 60 jazyků
  - rozdělí text na věty, slova
  - provede morfologickou analýzu - slovní druhy, pády, rody, vidy, ...
  - provede lemmatizaci - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
  - provede závislostní analýzu - gramatika (podmět, přísudek, přívlastky, …)
  - http://ufal.mff.cuni.cz/udpipe
- NameTag: hledá vlastní jména v textu (“Praha”, “Jan Novák”, …)
  - http://ufal.mff.cuni.cz/nametag

vzorový nástroj
pro dnešní tutoriál

- Korektor: spellchecker a oprava gramatiky
  - http://ufal.mff.cuni.cz/korektor

Výběr nástrojů pro analýzu textu

- UDPipe: nástroj pro analýzu textu
  - 60 jazyků
  - rozdělí text na věty, slova
  - provede morfologickou analýzu - slovní druhy, pády, rody, vidy, ...
  - provede lemmatizaci - základní tvar slova (“podzimním větrem” -> “podzimní vítr”)
  - provede závislostní analýzu - gramatika (podmět, přísudek, přívlastky, …)
  - http://ufal.mff.cuni.cz/udpipe
- NameTag: hledá vlastní jména v textu (“Praha”, “Jan Novák”, …)
  - http://ufal.mff.cuni.cz/nametag
- Korektor: spellchecker a oprava gramatiky
  - http://ufal.mff.cuni.cz/korektor

příklad na doma
nebo když zbyde čas

Co budeme potřebovat

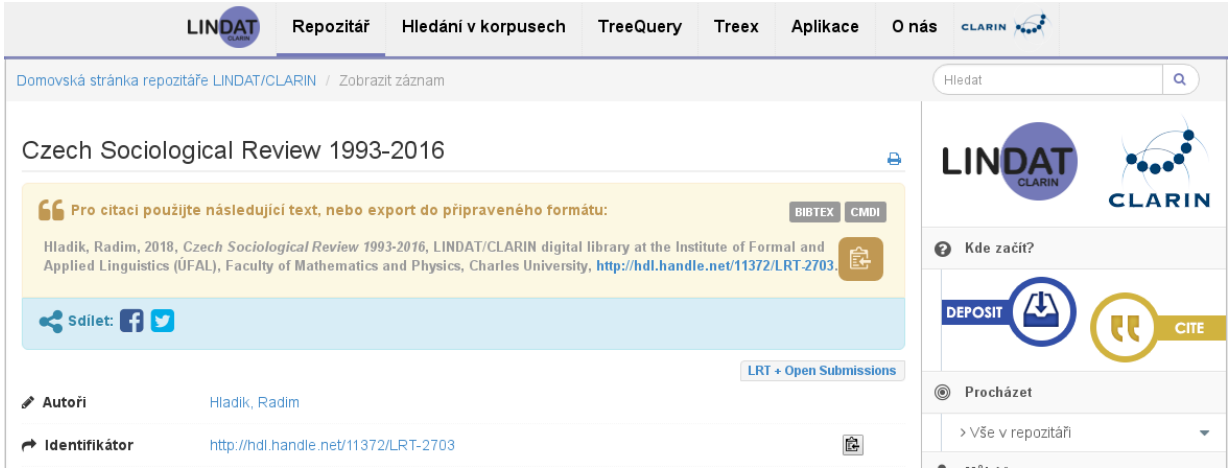
- osobní počítač s operačním systémem Unix / Windows
- webový prohlížeč
- příkazová řádka
  - Linux: Start -> Terminál / Emulátor terminálu
  - Windows:
    - Cygwin - program napodobující chování Unixových systémů ve Windows
    - nebo WSL – Windows Subsystem for Linux

Data

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Korpus: Czech Sociological Review 1993-2016

- Czech Sociological Review 1993-2016
  - Hladik, Radim, 2018, *Czech Sociological Review 1993-2016*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-2703
  - Korpus stáhneme a rozbalíme (nápověda: Předchozí tutoriál)



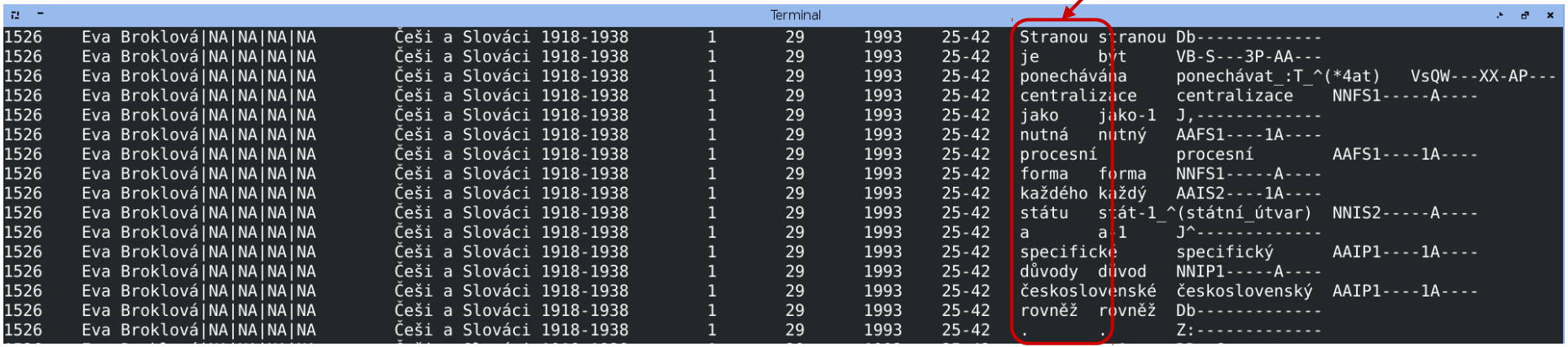
Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.
  - Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ...
  - Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.
  - Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ...
  - Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

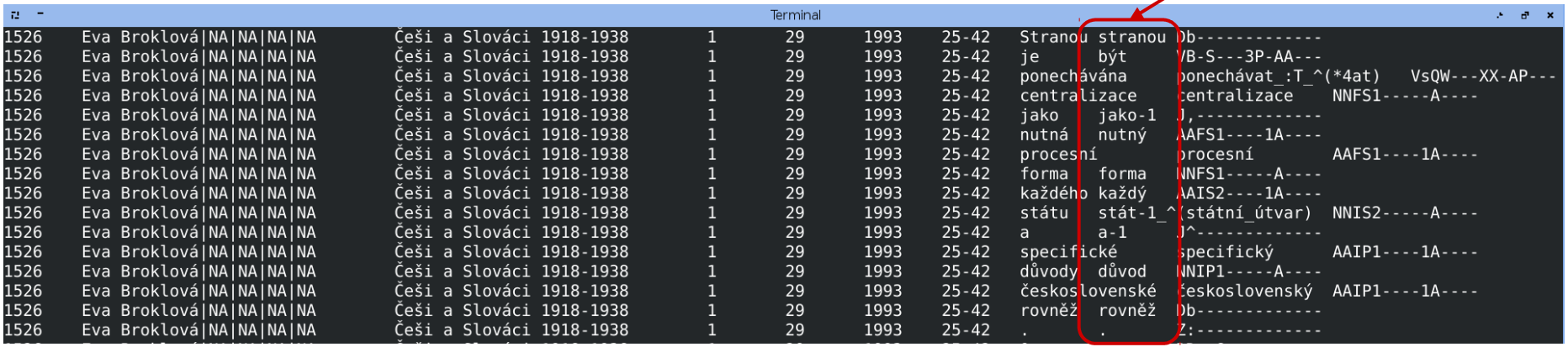
původní text rozdělený na slova



Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.
  - Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ...
  - Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

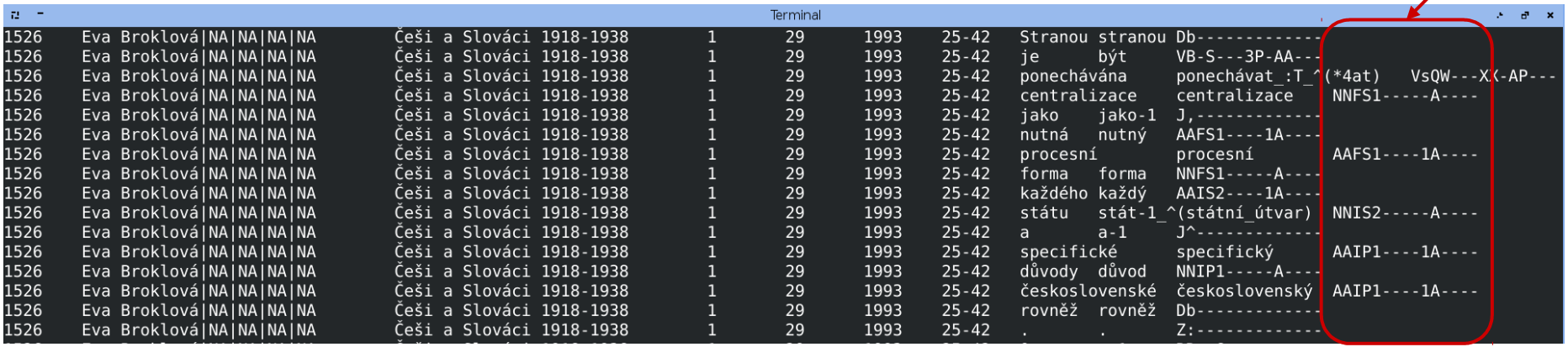
slova v základním tvaru (lemmata)



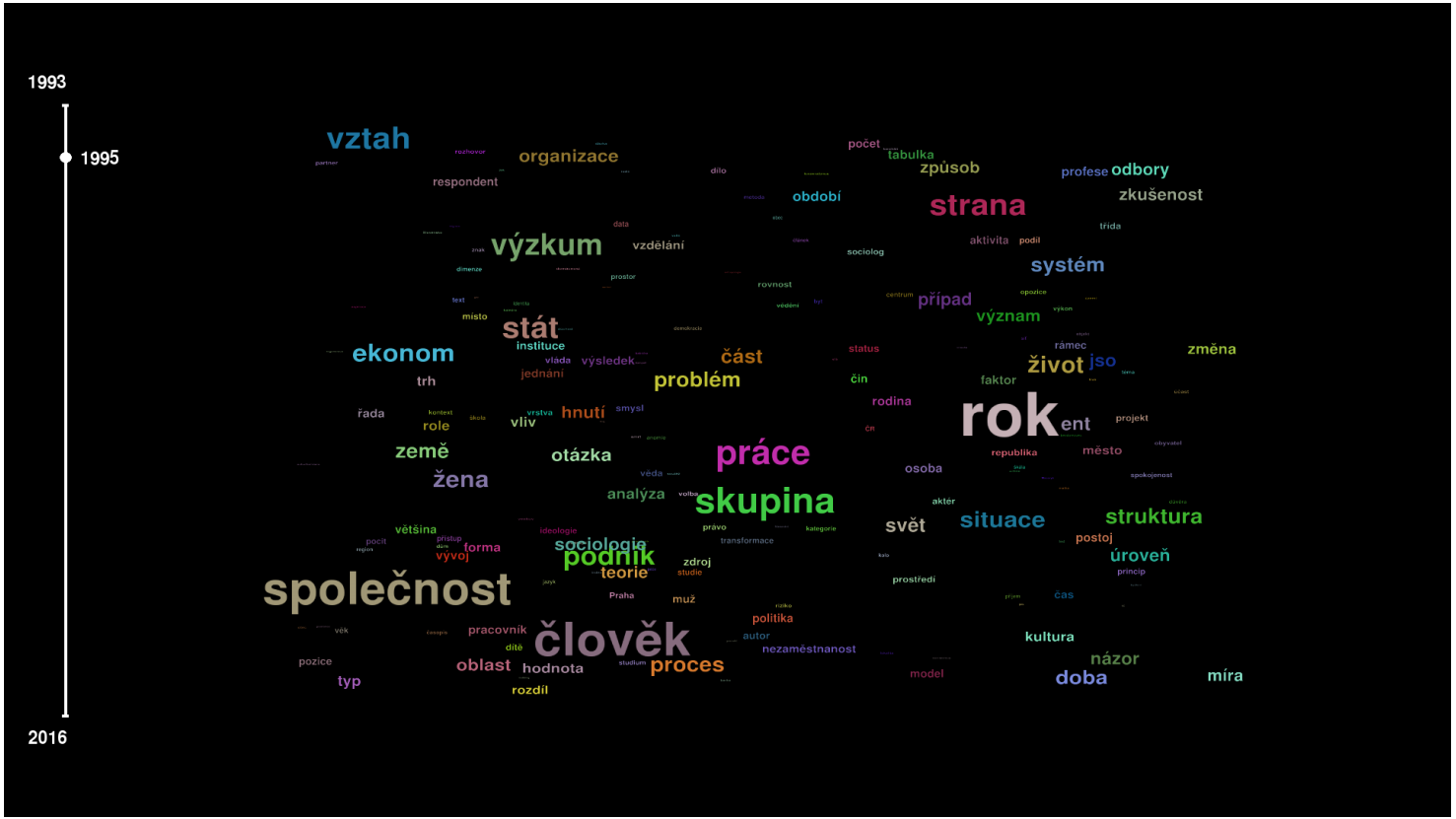
Prohlížení korpusu

- Přípona “tsv” = tab separated value = obsahuje sloupce.
  - Prohlížení v textovém editoru: Poznámkový blok, PSPad, vim, Emacs, ...
  - Prohlížení z příkazové řádky: `less soc_review_corpus_1993-2016.tsv`

morfologické značky



Vizualizace nejčastějších slov



Motivace: O kom se psalo?

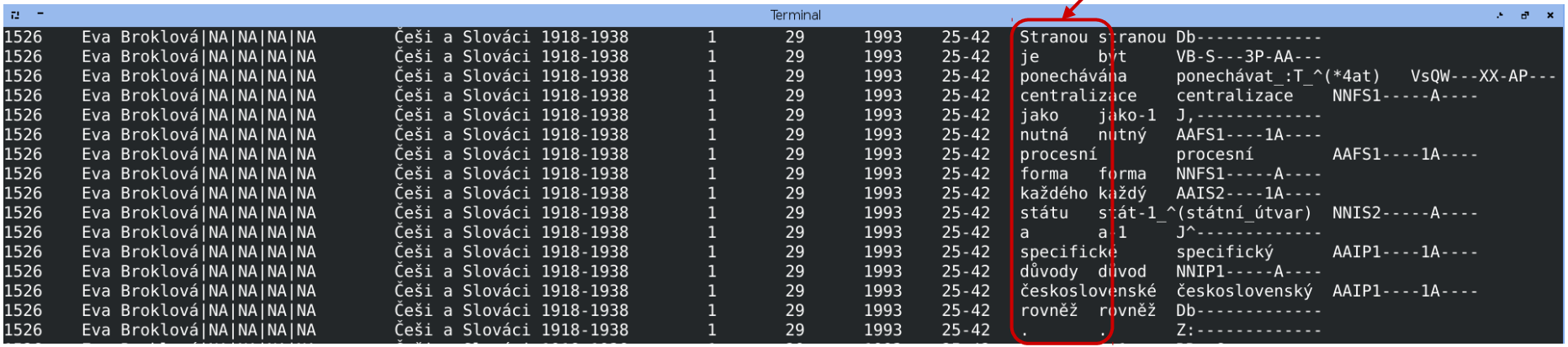
- Chceme analýzu pro vlastní jména -> hledáme pojmenované entity
- NameTag: nástroj pro rozpoznávání pojmenovaných entit

Příprava dat

Vyjmeme vstupní text rozdělený na slova (8. sloupec) a uložíme zvlášť:

```
cut -f8 soc_review_corpus_1993-2016.tsv > slova.txt
```

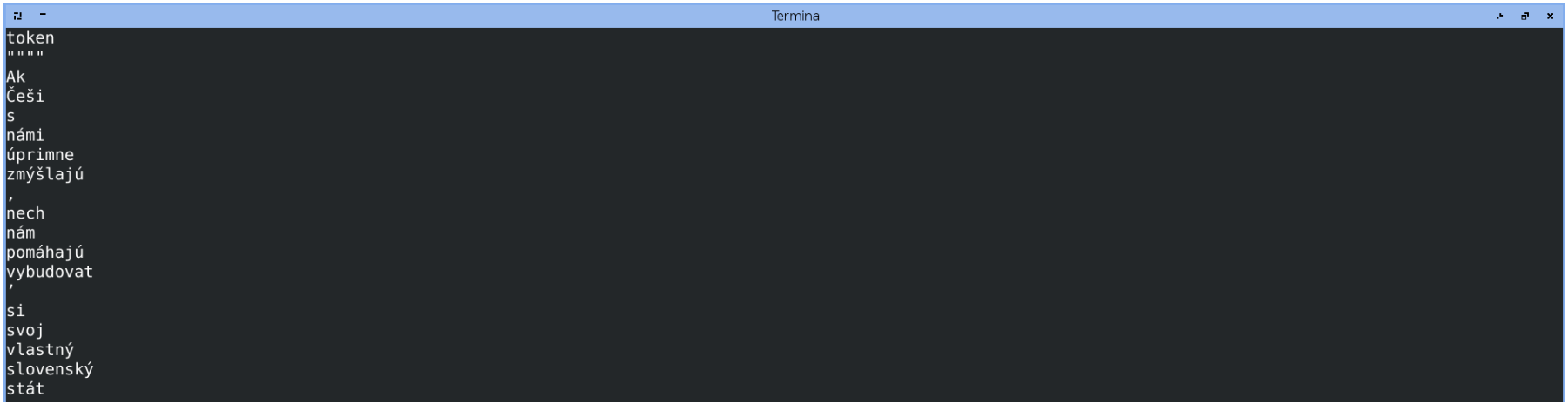
původní text rozdělený na slova



Příprava dat

- Prohlédneme si obsah souboru slova.txt:

```
less slova.txt
```



Způsoby využití nástrojů z LINDAT

Vybrané nástroje na příkazové řádce

Jana Straková, ÚFAL MFF UK

Tři způsoby použití nástrojů

- Demo v prohlížeči
  - pouze pro jednoduché příklady
  - nelze zpracovat větší data

Tři způsoby použití nástrojů

- Demo v prohlížeči
  - pouze pro jednoduché příklady
  - nelze zpracovat větší data
- Instalace nástroje (a modelů) do vlastního počítače
  - Lze zpracovávat velká data
  - Mohu trénovat vlastní modely
  - Nástroj lze zapojit do většího frameworku

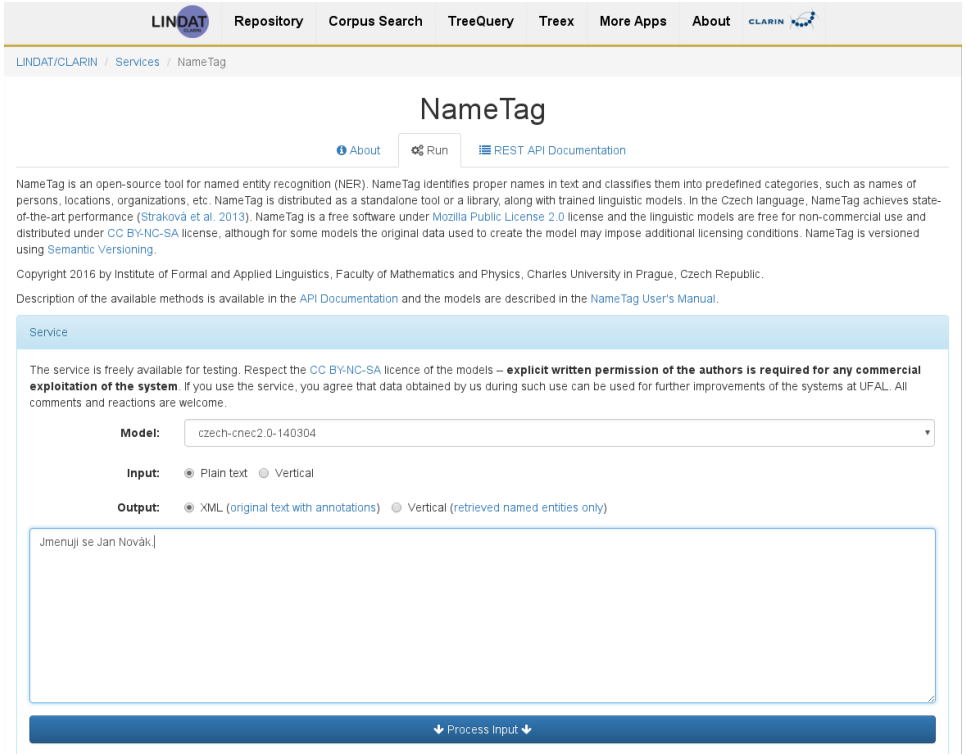
Tři způsoby použití nástrojů

- Demo v prohlížeči
  - pouze pro jednoduché příklady
  - nelze zpracovat větší data
- Instalace nástroje (a modelů) do vlastního počítače
  - Lze zpracovávat velká data
  - Mohu trénovat vlastní modely
  - Nástroj lze zapojit do většího frameworku
- REST API je způsob, jak snadno získávat informace z webové služby
  - Nástroj “žije” na serveru jako webová služba, my (klient) posíláme dotazy a dostáváme zpět odpovědi od serveru
  - Odpadá instalace nástroje a správa modelů

Demo

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Demo



http://lindat.mff.cuni.cz/services/nametag/

Instalace nástroje

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Instalace

Instalace probíhá stažením a rozbalením do adresáře.

Vytvoříme si nový adresář např. `nametag`:

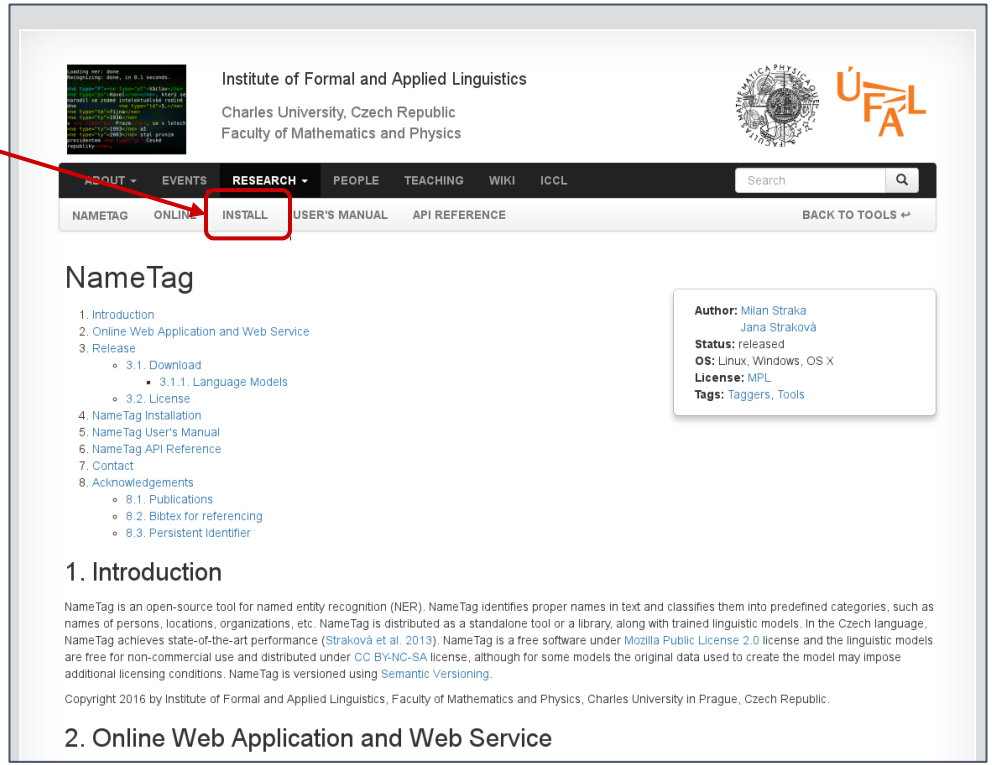
- z příkazové řádky Unix: `mkdir nametag`
- případně jak jsme zvyklí ve správci

Stačí umístit do osobních složek uživatele, např. `/home/<uživatel>/nametag`.

Budeme potřebovat:

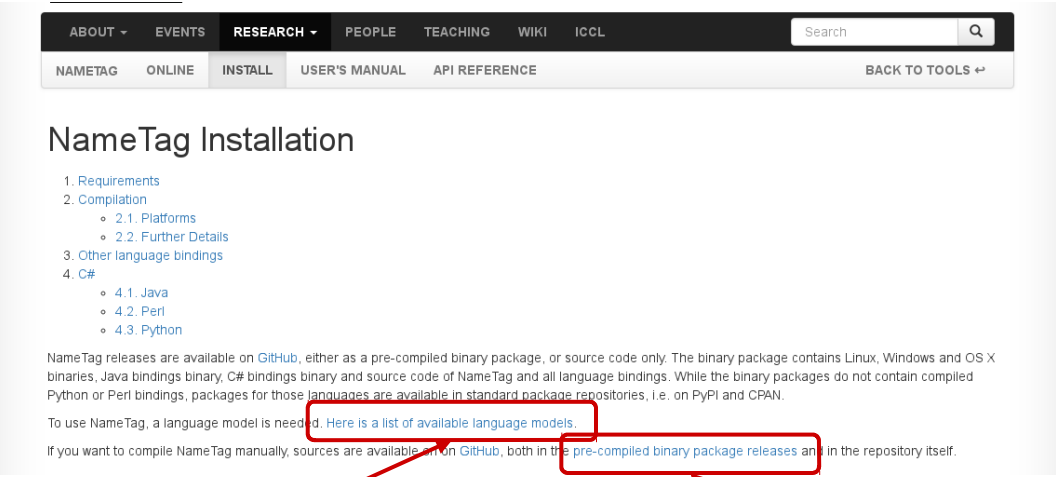
1. program NameTag
2. předtrénovaný model k programu

Instalace



http://ufal.mff.cuni.cz/nametag

Instalace

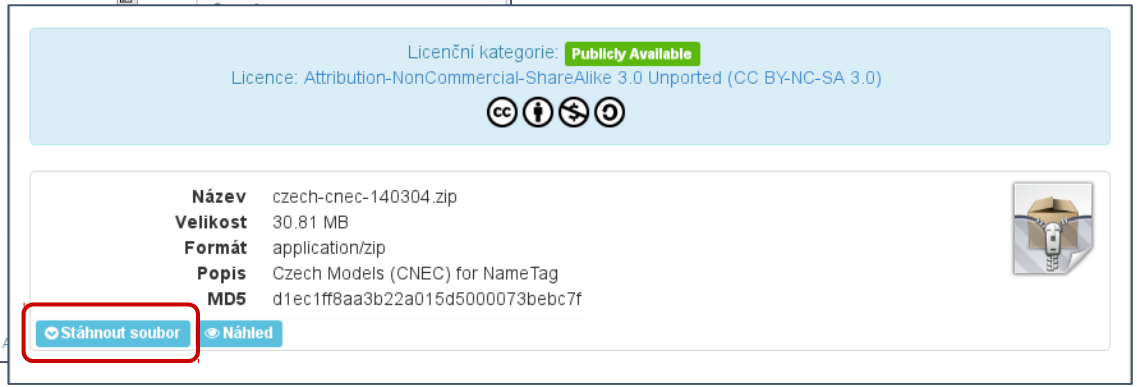


1. odkaz na předtrénovaný model

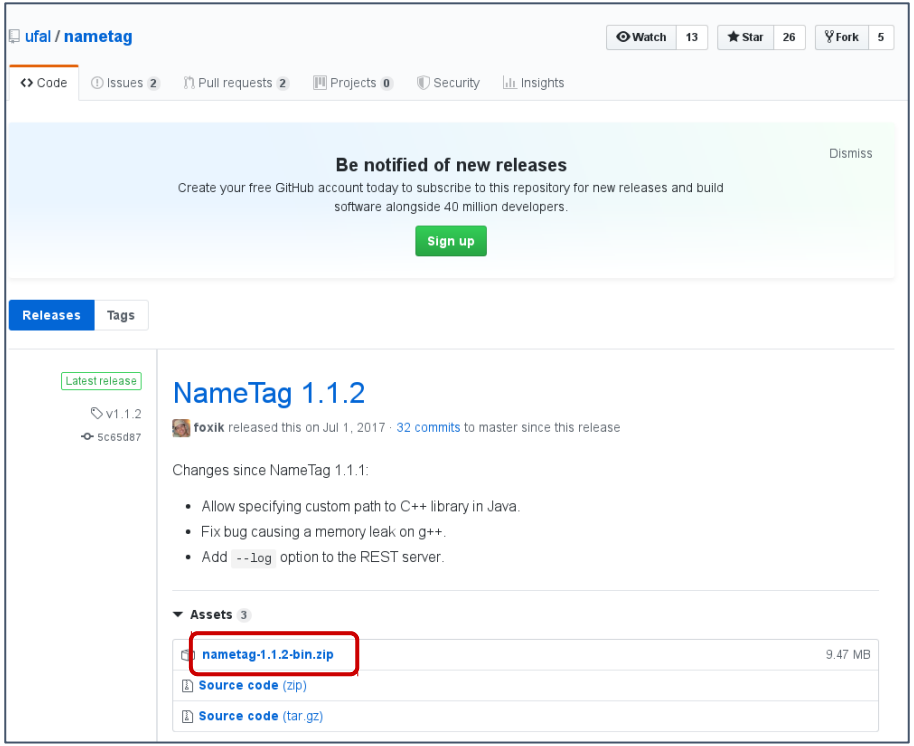
2. odkaz na program v repozitáři GitHub

http://ufal.mff.cuni.cz/nametag/install

Instalace - stažení modelu



Instalace - stažení programu



Instalace - rozbalení

Program i model rozbalíme a máme nainstalováno.

- z příkazové řádky Unix:
    - `unzip czech-cnec-140304.zip`
    - `unzip nametag-1.1.2-bin.zip`
- případně jak jsme zvyklí rozbalovat `zip` soubory ve správci

První spuštění

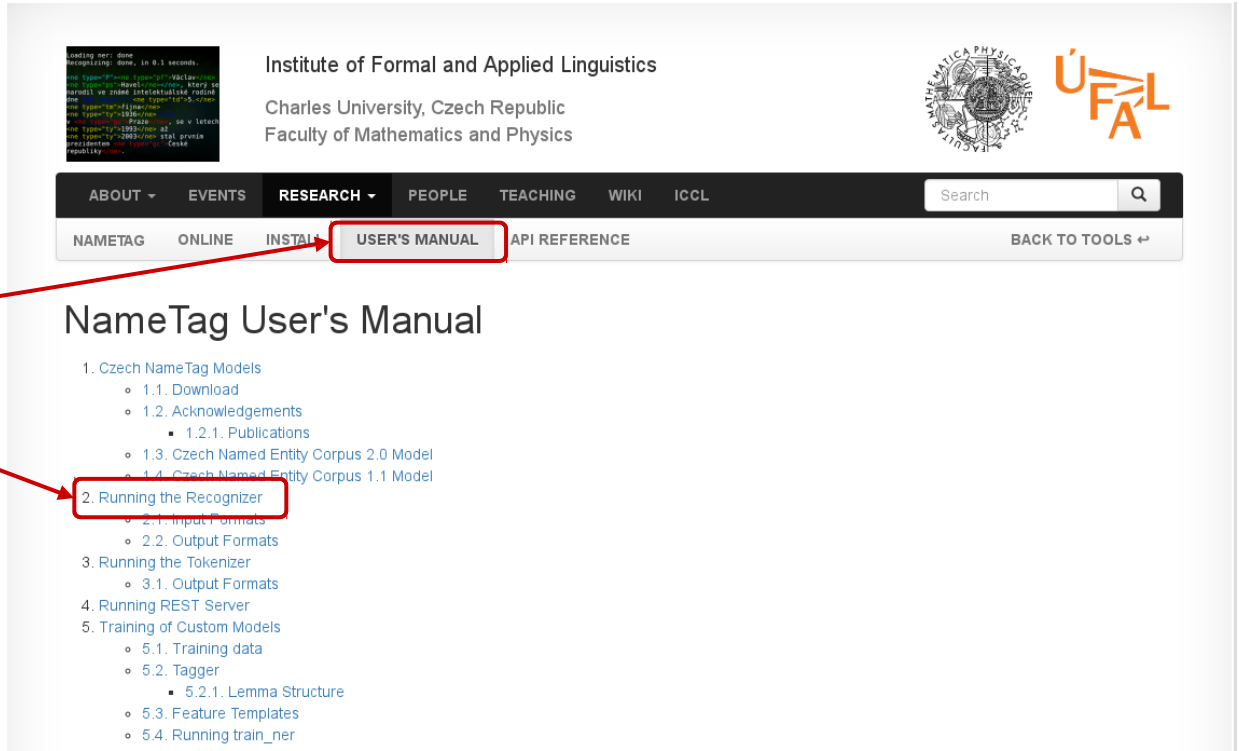
Najdeme odpovídající binárku programu pro náš počítač v adresáři `nametag-1.1.2-bin`:

- 32-bit linux: `bin-linux32`, 64-bit linux: `bin-linux64`
- 32-bit Windows: `bin-win32`, 64-bit Windows: `bin-win64`

Spuštění programu: `nametag-1.1.2-bin/bin-linux64/run_ner`

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - Uživatelský manuál



http://ufal.mff.cuni.cz/nametag/users/manual

Spuštění rozpoznávače - Uživatelský manuál

2. Running the Recognizer

The NameTag Recognizer can be executed using the following command:

```
run_ner recognizer_model
```

The input is assumed to be in UTF-8 encoding and can be either already tokenized and segmented, or it can be a plain text which is tokenized and segmented automatically.

Any number of files can be specified after the `recognizer_model`. If an argument `input_file:output_file` is used, the given `input_file` is processed and the result is saved to `output_file`. If only `input_file` is used, the result is saved to standard output. If no argument is given, input is read from standard input and written to standard output.

The full command syntax of `run_ner` is

```
Usage: run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
```

2.1. Input Formats

The input format is specified using the `--input` option. Currently supported input formats are:

- `untokenized` (default): the input is tokenized and segmented using a tokenizer defined by the model,
- `vertical`: the input is in vertical format, every line is considered a word, with empty line denoting end of sentence.

2.2. Output Formats

The output format is specified using the `--output` option. Currently supported output formats are:

- `xml` (default): Simple XML format without a root element, using `<sentence>` element to mark sentences and `<token>` element to mark tokens. The recognized named entities are encoded using `<ne type="...">` element.

  Example input:

  ```
  Václav Havel byl český dramatik, esejista, kritik komunistického režimu a později politik.
  ```

  A NameTag identifies a first name (`pf`), a surname (`ps`) and a person name container (`P`) in the input (line breaks added):

  ```
  <sentence><ne type="P"><ne type="pf"><token>Václav</token></ne> <ne type="ps"><token>Havel</token></ne></ne>
  <token>byl</token> <token>český</token> <token>dramatik</token><token>,</token> <token>esejista</token><token>,</token>
  <token>kritik</token> <token>komunistického</token> <token>režimu</token> <token>a</token> <token>později</token>
  <token>politik</token><token>.</token></sentence>
  ```

- `vertical`: Every found named entity is on a separate line. Each line contains three tab-separated fields: *entity_range*, *entity_type* and *entity_text*. The *entity_range* is composed of token identifiers (counting from 1 and including end-of-sentence; if the input is also `vertical`, token identifiers correspond exactly to line numbers) of tokens forming the named entity and *entity_type* represents its type. The *entity_text* is not strictly necessary and contains space separated words of this named entity.

  Example input:

  ```
  Václav Havel byl český dramatik, esejista, kritik komunistického režimu a později politik.
  ```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Uložme si příkladovou větu do souboru:

```
echo "Václav Havel byl prvním prezidentem České republiky." > veta.txt
```

Spustíme rozpoznávač:

```
nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
Loading ner: done
<sentence><token>“</token><ne type="P"><ne type="pf"><token>Václav</token></ne> <ne type="ps"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním</token> <token>prezidentem</token> <ne type="gc"><token>České</token> <token>republiky</token></ne><token>.</token><token>”</token></sentence>
Recognizing done, in 0.000 seconds.
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - jedna věta

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spuštění rozpoznávače - parametr

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spustíme rozpoznávač s parametrem:

```
nametag-1.1.2-bin/bin-linux64/run_ner --output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

Spuštění rozpoznávače - parametr

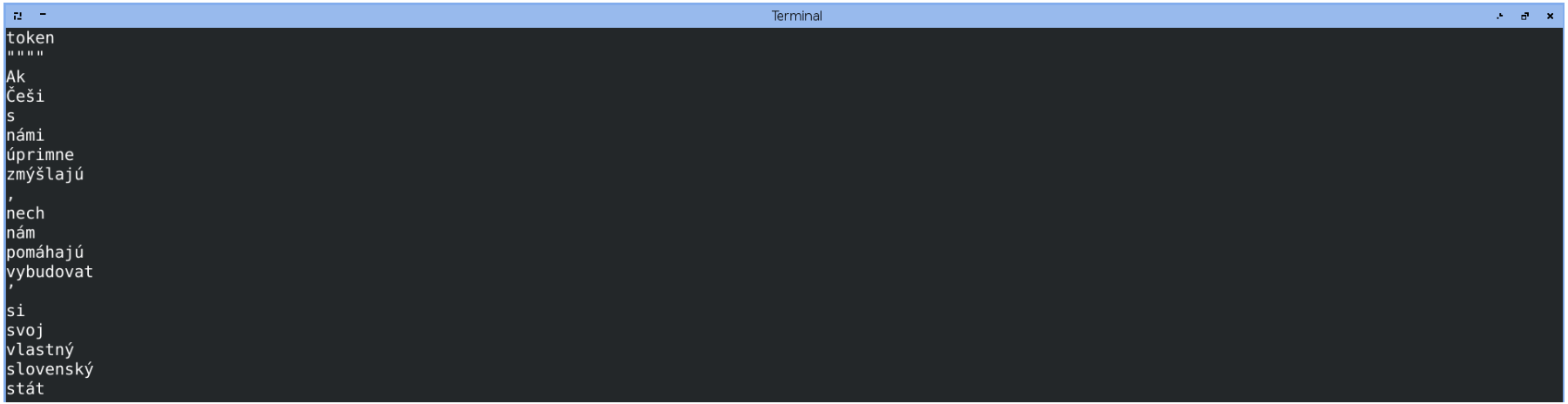
```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

Spustíme rozpoznávač s parametrem:

```
nametag-1.1.2-bin/bin-linux64/run_ner --output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
```

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner --output=vertical czech-cnec-140304/czech-cnec2.0-140304.ner veta.txt
Loading ner: done
2,3     P       Václav Havel
2       pf      Václav
3       ps      Havel
7,8     gc      České republiky
Recognizing done, in 0.000 seconds.
```

Spuštění rozpoznávače na našich datech



Hladik, Radim, 2018, *Czech Sociological Review 1993-2016*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-2703

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Spuštění rozpoznávače na našich datech

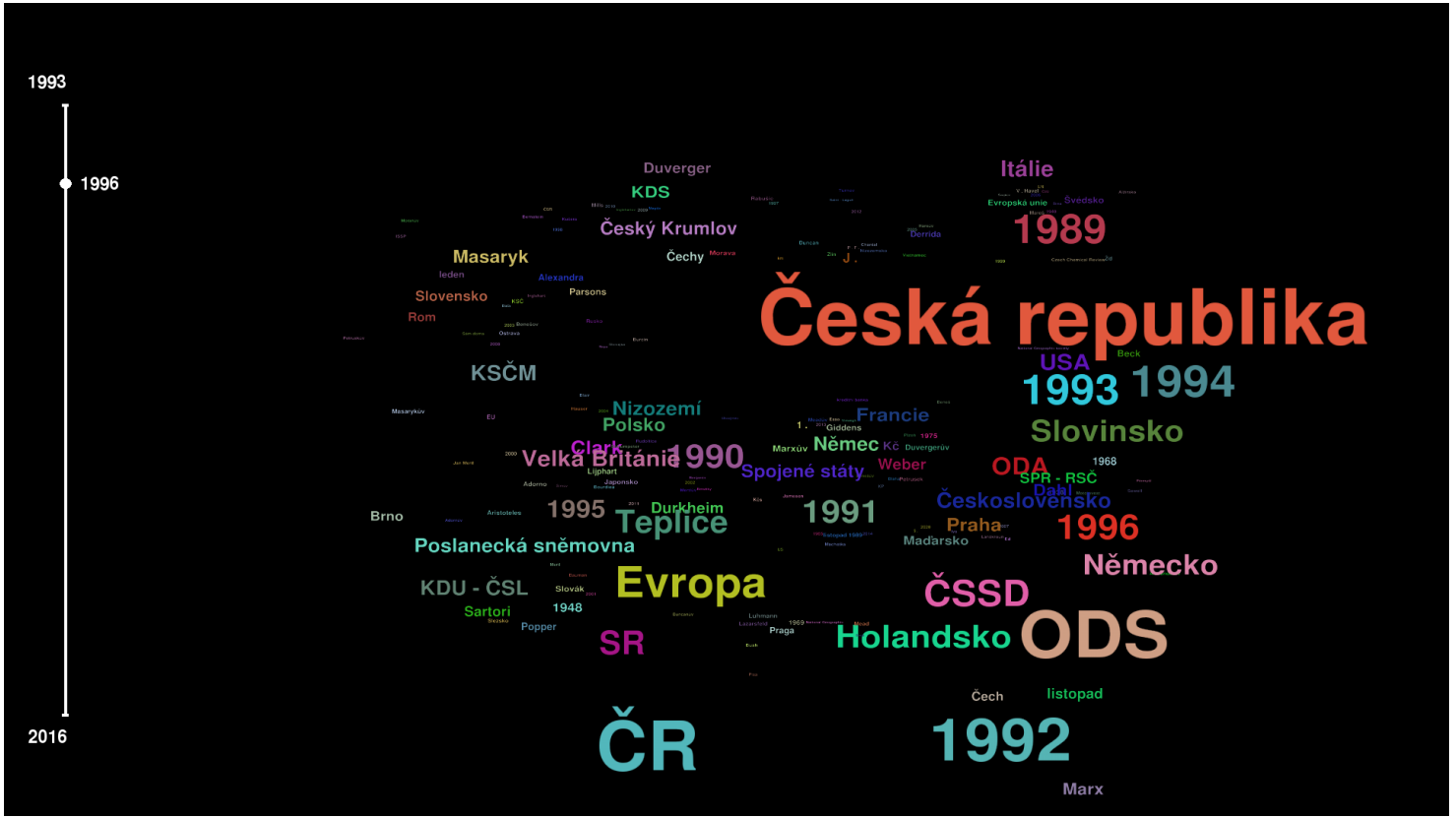
```
gris@ametyst:~/nametag_tutorial$ nametag-1.1.2-bin/bin-linux64/run_ner
Usage: nametag-1.1.2-bin/bin-linux64/run_ner [options] recognizer_model [file[:output_file]]...
Options: --input=untokenized|vertical
         --output=vertical|xml
         --version
         --help
gris@ametyst:~/nametag_tutorial$
```

```
nametag-1.1.2-bin/bin-linux64/run_ner --input=vertical -output=vertical
czech-cnec-140304/czech-cnec2.0-140304.ner slova.txt:jmena.txt
```

Výstup rozpoznávače

```
3         gc      Ak
4         pc      Češi
42,43     gc      slovenskej republiky
122       pc      Čechů
124       pc      Slovákům
197       gc      Slovenska
271       gc      ČSR
338       ps      Henlein
345       gc      Slovensku
469,470   gr      Podkarpatské Rusi
520       gc      Slovenska
537       gr      Slezsku
539       ps      Masaryk
618       ty      1918
642       gc      Rakouska
```

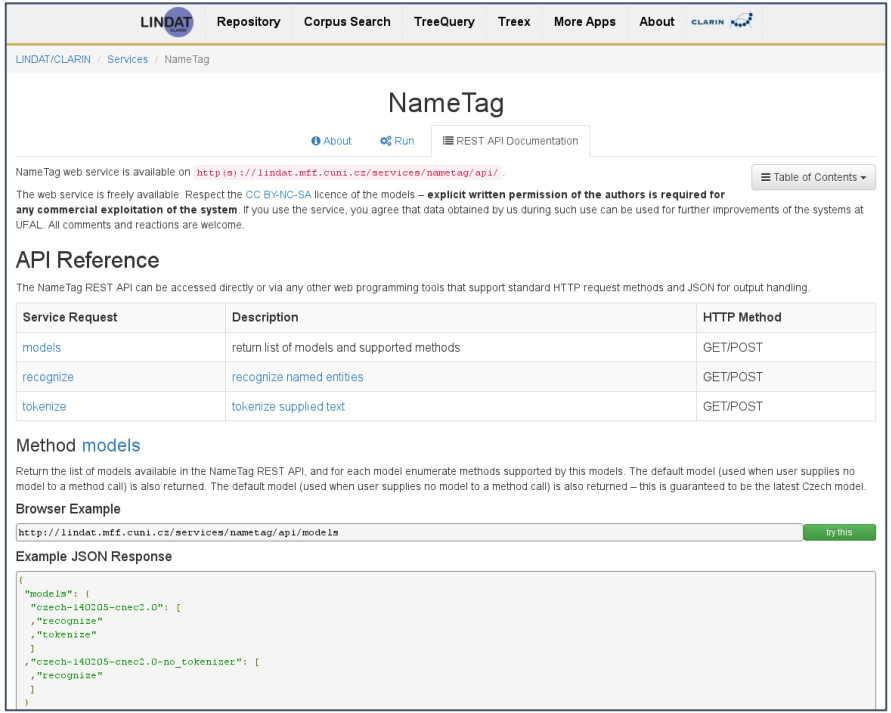
Vizualizace nejčastějších vlastních jmen



Webová služba

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

REST API



http://lindat.mff.cuni.cz/services/nametag/api-reference.php

REST API v prohlížeči

Do adresy v prohlížeči zadáme

```
http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.
```

REST API v prohlížeči

Do adresy v prohlížeči zadáme

```
http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.
```

Vrátí se nám

REST API v prohlížeči

Do adresy v prohlížeči zadáme

```
http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.
```

Vrátí se nám

REST API v prohlížeči

Do adresy v prohlížeči zadáme

```
http://lindat.mff.cuni.cz/services/nametag/api/recognize?data=Václav Havel byl prvním prezidentem České republiky.
```

Vrátí se nám

formát JSON

API pomocí CURL (příkazová řádka)

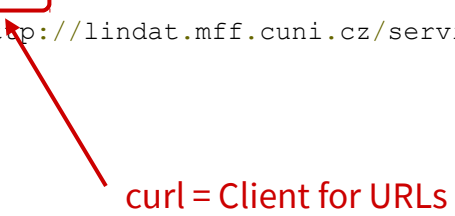
Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

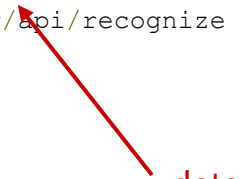
```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```



API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```



API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

API pomocí CURL (příkazová řádka)

Na příkazové řádce zadáme

```
curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.'
http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním
</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
gris@ametyst:~$
```

CURL - přidání parametru

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním
</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
gris@ametyst:~$
```

CURL - přidání parametru

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním
</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
gris@ametyst:~$
```

CURL s parametrem:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

CURL - přidání parametru

```
gris@ametyst:~$ curl --data-urlencode 'data=Václav Havel byl prvním prezidentem České republiky.' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "<sentence><ne type=\"P\"><ne type=\"pf\"><token>Václav</token></ne> <ne type=\"ps\"><token>Havel</token></ne></ne> <token>byl</token> <token>prvním
</token> <token>prezidentem</token> <ne type=\"gc\"><token>České</token> <token>republiky</token></ne><token>.</token></sentence>"
}
gris@ametyst:~$
```

CURL s parametrem:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

```
gris@ametyst:~$ curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "1,2\tP\tVáclav Havel\n1\tpf\tVáclav\n2\tps\tHavel\n6,7\tgc\tČeské republiky\n"
}
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo “Václav Havel byl prvním prezidentem České republiky.” > veta.txt
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo “Václav Havel byl prvním prezidentem České republiky.” > veta.txt
```

CURL se vstupem ze souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo “Václav Havel byl prvním prezidentem České republiky.” > veta.txt
```

CURL se vstupem ze souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

Přesměrování výstupu do souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
> result_json.txt
```

CURL - vstup ze souboru, výstup do souboru

Větu uložíme do souboru:

```
echo “Václav Havel byl prvním prezidentem České republiky.” > veta.txt
```

CURL se vstupem ze souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
```

Přesměrování výstupu do souboru:

```
curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize
> result_json.txt
```

```
gris@ametyst:~$ curl -F 'data=@veta.txt' -F 'output=vertical' http://lindat.mff.cuni.cz/services/nametag/api/recognize > result_json.txt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   709    0   363  100   346    602    574 --:--:-- --:--:-- --:--:--   601
gris@ametyst:~$ cat result_json.txt
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "1,2\tP\tVáclav Havel\n1\tpf\tVáclav\n2\tps\tHavel\n6,7\tgc\tČeské republiky\n"
}
gris@ametyst:~$
```

CURL - konverze JSON výstupu do textu


```
{
 "model": "czech-cnec2.0-140304"
,"acknowledgements": [
  "http://ufal.mff.cuni.cz/nametag#nametag_acknowledgements"
 ,"http://ufal.mff.cuni.cz/nametag/users-manual#czech-cnec_acknowledgements"
 ,"http://ufal.mff.cuni.cz/morphodita#morphodita_acknowledgements"
 ]
,"result": "1,2\tP\tVáclav Havel\n1\tpf\tVáclav\n2\tps\tHavel\n6,7\tgc\tČeské republiky\n"
}
```


```
cat result_json.txt | PYTHONIOENCODING=utf-8 python3 -c "import json,sys; print(json.load(sys.stdin)['result'], end='')" > result.txt
```

```
gris@ametyst:~$ cat result_json.txt | PYTHONIOENCODING=utf-8 python3 -c "import json,sys; print(json.load(sys.stdin)['result'], end='')" > result.txt
gris@ametyst:~$ cat result.txt
1,2     P       Václav Havel
1       pf      Václav
2       ps      Havel
6,7     gc      České republiky
gris@ametyst:~$
```

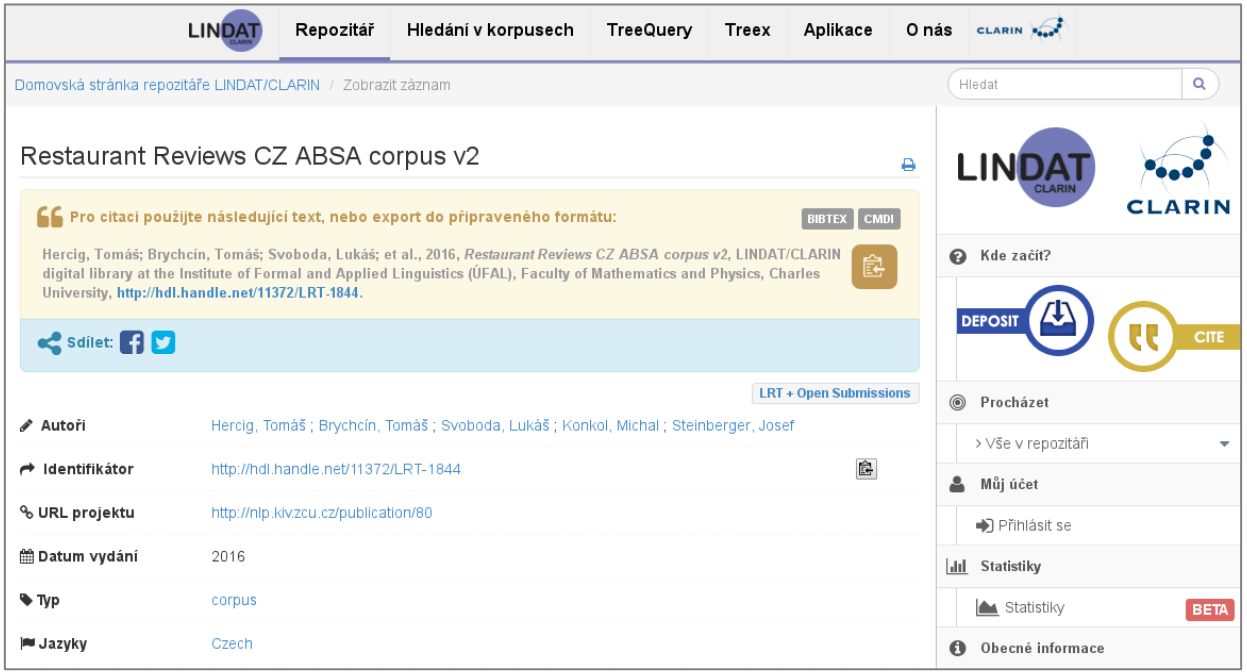
Další příklad: Korektor

Vybrané nástroje na příkazové řádce
Jana Straková, ÚFAL MFF UK

Korpus: Restaurant Reviews CZ ABSA corpus v2

Restaurant Reviews CZ ABSA corpus v2

Hercig, Tomáš; Brychcín, Tomáš; Svoboda, Lukáš; et al., 2016, *Restaurant Reviews CZ ABSA corpus v2*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-1844



Prohlížení korpusu

```
<?xml version="1.0" encoding="UTF-8"?>
<Reviews>
    <Review rid="ALL">
        <sentences>
            <sentence id="1411">
                <text>Jídlo taky nic moc</text>
                <Opinions>
                    <Opinion category="null" from="0" polarity="negative" target="Jídlo" to="5"/>
                    <Opinion category="food" from="0" polarity="negative" target="NULL" to="0"/>
                </Opinions>
            </sentence>
            <sentence id="1410">
                <text>Jídlo od Nás dostává 5 z 5 a pan Doksanský (spolumajitel) i jeho personál jsou profesi
pití ke konkrétním jídlům</text>
                <Opinions>
                    <Opinion category="null" from="44" polarity="positive" target="spolumajitel" to="56"/>
                    <Opinion category="null" from="65" polarity="positive" target="personál" to="73"/>
                    <Opinion category="null" from="137" polarity="positive" target="pití" to="141"/>
                    <Opinion category="null" from="156" polarity="positive" target="jídlům" to="162"/>
                    <Opinion category="null" from="122" polarity="positive" target="víno" to="126"/>
                    <Opinion category="food" from="0" polarity="positive" target="NULL" to="0"/>
                    <Opinion category="service" from="0" polarity="positive" target="NULL" to="0"/>
                </Opinions>
            </sentence>
            <sentence id="941">
                <text>Vloni trochu na kvalitě polevili,ale nadále bezva</text>
```

Motivace: doplnění chybějící diakritiky

```
<sentence id="942">
    <text>Vsichni vime, ze Bio maso je podstatne drazsi nez maso klasicke, ale ceny zde jsou opravdu na ceny z Bio chovu, neni neco spatne?</text>
    <Opinions>
        <Opinion category="null" from="50" polarity="neutral" target="maso" to="54"/>
        <Opinion category="null" from="69" polarity="negative" target="ceny" to="73"/>
        <Opinion category="null" from="17" polarity="neutral" target="Bio maso" to="25"/>
        <Opinion category="food" from="0" polarity="neutral" target="NULL" to="0"/>
        <Opinion category="price" from="0" polarity="negative" target="NULL" to="0"/>
    </Opinions>
</sentence>
```

```
<sentence id="10091">
    <text>V hamburgeru byl jakysi malinky karbanatek a nejvyraznejsi bylo celkove rajce.</text>
    <Opinions>
        <Opinion category="null" from="72" polarity="positive" target="rajce" to="77"/>
        <Opinion category="null" from="32" polarity="negative" target="karbanatek" to="42"/>
        <Opinion category="null" from="2" polarity="negative" target="hamburgeru" to="12"/>
        <Opinion category="food" from="0" polarity="negative" target="NULL" to="0"/>
    </Opinions>
</sentence>
```

Motivace: doplnění chybějící diakritiky

```
<sentence id="942">
    <text>Vsichni vime, ze Bio maso je podstatne drazsi nez maso klasicke, ale ceny zde jsou opravdu na ceny z Bio chovu, neni neco spatne?</text>
    <Opinions>
        <Opinion category="null" from="50" polarity="neutral" target="maso" to="54"/>
        <Opinion category="null" from="69" polarity="negative" target="ceny" to="73"/>
        <Opinion category="null" from="17" polarity="neutral" target="Bio maso" to="25"/>
        <Opinion category="food" from="0" polarity="neutral" target="NULL" to="0"/>
        <Opinion category="price" from="0" polarity="negative" target="NULL" to="0"/>
    </Opinions>
</sentence>
```

```
<sentence id="10091">
    <text>V hamburgeru byl jakysi malinky karbanatek a nejvyraznejsi bylo celkove rajce.</text>
    <Opinions>
        <Opinion category="null" from="72" polarity="positive" target="rajce" to="77"/>
        <Opinion category="null" from="32" polarity="negative" target="karbanatek" to="42"/>
        <Opinion category="null" from="2" polarity="negative" target="hamburgeru" to="12"/>
        <Opinion category="food" from="0" polarity="negative" target="NULL" to="0"/>
    </Opinions>
</sentence>
```

Korektor: spellchecker a oprava gramatiky

http://ufal.mff.cuni.cz/korektor

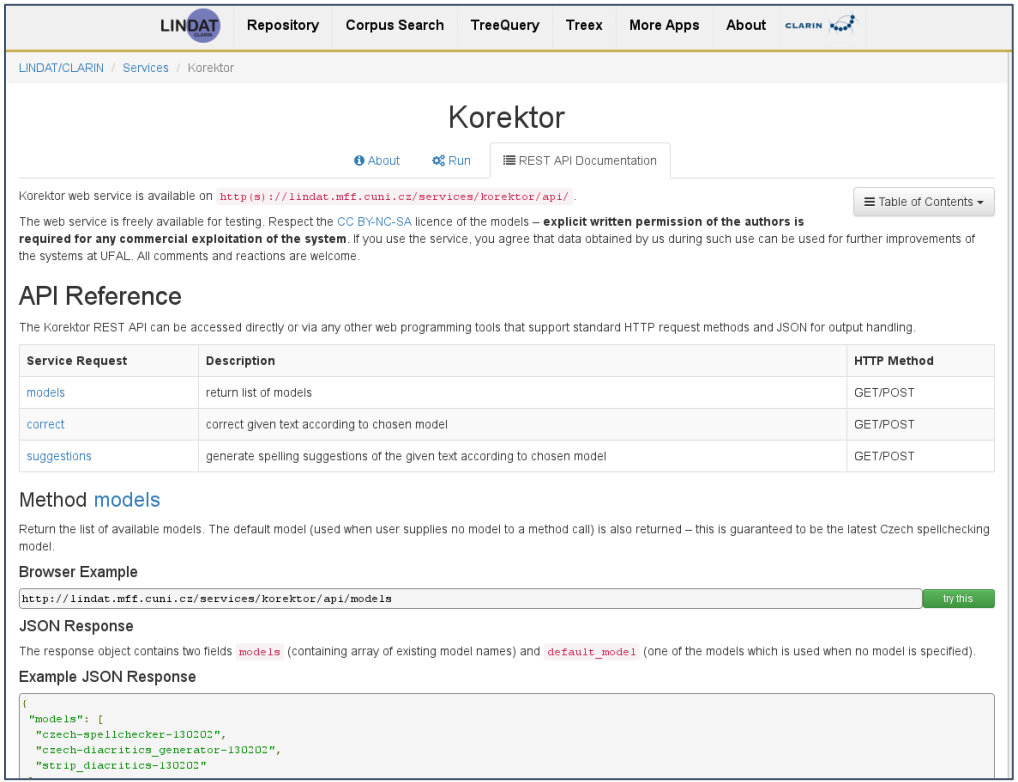
Získání dat

Získáme všechny věty (řádky s obsahem `<text>`), značku `<text>` odstraníme:

```
grep "<text>" CzechABSA-v2_format_SemEval2016.xml | sed "s/<text>//;s/<\/text>//" > texty.txt
```

```
                Jídlo taky nic moc
                Jídlo od Nás dostává 5 z 5 a pan Doksanský (spolumajitel) i jeho personál jsou profesionálové kteří Vám doporučí kvalitní víno nebo jiné pití k
e konkrétním jídlům
                Vloni trochu na kvalitě polevili,ale nadále bezva
                Kdo ví, třeba bych tam ještě po pozitivní konstruktivní kritice dostal přes ústa
                Kdo Vám v dnešní době nabídne výběr z 5 sklutečně fresh jídel, která mají i super chuť
                Vsichni vime, ze Bio maso je podstatne drazsi nez maso klasicke, ale ceny zde jsou opravdu na ceny z Bio chovu, neni neco spatne?
                Jistě se k vám brzy vrátím a všem vřele tuto restauraci doporučuji
                Mají tu fantastické jídlo a příjemné prostředí!
                Všichni byli sice milí, ale v pořadu vypadala hospůdka U Štěpána rozhodně lákavěji
                Ku příkladu dnes 29.10.2013 nabízejí LOSOSO S GRILOVANOU ZELENINOU, RISSOTO a nesmím zapomenout na FRANFURSKOU POLÉVKU
                Všechno bylo naprosto bez obtíží, obsluha super, jídlo bez vady
                Když budu chtít zas někdy někoho pozvat,rozhodně ne sem
                všechno je super ,tak se nenechte odradit
                Vynikající obsluha pozorná vítající hned při příchodu
                Rozsah jídelního lístku velký ,jídlo nám vždy chutnalo
                RESTAURACI VŠEM DOPOPRUČUJI
                Pro mě prostě vynikající steaky v příjemném prostředí, příště kupuji zase a moc se těším!
                Pro změnu byl steak krvavý a nechutný
                SKVĚLÁ KUCHTIČKA,JÍDLO CHUTNÉ A PIVO???DOBŘE CHLAZENÉ
                S obsluhou jsme byli spokojeni
                Poté přišel nejspíš vedoucí, jídlo odnesl a donesl nové
                Prostředí nic moc
```

Korektor jako webová služba



https://lindat.mff.cuni.cz/services/korektor/api-reference.php

Zjištění dostupných modelů

```
curl http://lindat.mff.cuni.cz/services/korektor/api/models
```

Zjištění dostupných modelů

```
curl http://lindat.mff.cuni.cz/services/korektor/api/models
```

```
gris@ametyst:~/ufal/work/prednasky_a_slajdy/2019_lindat_tutorial$ curl http://lindat.mff.cuni.cz/services/korektor/api/models
{
 "models": [
  "czech-spellchecker-130202",
  "czech-diacritics_generator-130202",
  "strip_diacritics-130202"
 ],
 "default_model": "czech-spellchecker-130202"
}
```

Zjištění dostupných modelů Korektoru

```
curl http://lindat.mff.cuni.cz/services/korektor/api/models
```

```
gris@ametyst:~/ufal/work/prednasky_a_slajdy/2019_lindat_tutorial$ curl http://lindat.mff.cuni.cz/services/korektor/api/models
{
 "models": [
  "czech-spellchecker-130202",
  "czech-diacritics_generator-130202",
  "strip_diacritics-130202"
 ],
 "default_model": "czech-spellchecker-130202"
}
```

Doplnění diakritiky pomocí Korektoru

```
curl -F 'data=@texty.txt' -F 'model=czech-diacritics_generator-130202'
http://lindat.mff.cuni.cz/services/korektor/api/correct | PYTHONIOENCODING=utf-8 python -c
"import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])" > opravene_texty.txt
```

Doplnění diakritiky pomocí Korektoru

```
curl -F 'data=@texty.txt' -F 'model=czech-diacritics_generator-130202'
http://lindat.mff.cuni.cz/services/korektor/api/correct | PYTHONIOENCODING=utf-8 python -c
"import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])" > opravene_texty.txt
```

Doplnění diakritiky pomocí Korektoru

```
curl -F 'data=@texty.txt' -F 'model=czech-diacritics_generator-130202'
http://lindat.mff.cuni.cz/services/korektor/api/correct | PYTHONIOENCODING=utf-8 python -c
"import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])" > opravene_texty.txt
```

```
 Jídlo od Nás dostává 5 z 5 a pan Doksanský (spolumajitel) i jeho personál jsou profesionálové kteří Vám doporučí kvalitní víno neb
dlům
 Vloni trochu na kvalitě polevili,ale nadále bezva
 Kdo ví, třeba bych tam ještě po pozitivní konstruktivní kritice dostal přes ústa
 Kdo Vám v dnešní době nabídne výběr z 5 skutečně fresh jídel, která mají i super chuť
 Všichni víme, že Bio maso je podstatně dražší než maso klasické, ale ceny zde jsou opravdu na ceny z Bio chovu, není něco špatně?
 Jistě se k vám brzy vrátím a všem vřele tuto restauraci doporučují
 Mají tu fantastické jídlo a příjemné prostředí!
 Všichni byli sice milí, ale v pořadu vypadala hospůdka U Štěpána rozhodně lákavěji
 Ku příkladu dnes 29.10.2013 nabízejí LOSOSO S GRILOVANOU ZELENINOU, RISSOTO a nesmím zapomenout na FRANFURSKOU POLÉVKU
 Všechno bylo naprosto bez obtíží, obsluha super, jídlo bez vady
 Když budu chtít zas někdy někoho pozvat,rozhodně ne sem
 všechno je super ,tak se nenechte odradit
```