Služby spracovania textu vytvorené na FIIT STU

OK

PeWe@FIIT
personalized web group

11.11.2014

SK Lemmatizátor

- http://text.fiit.stuba.sk/lemmatizer/
- Lematizácia slovenského textu

- Rýchla
  - Plain-text
  - Najpravdepodobnejšie lemmy

- Úplna
  - XML alebo JSON
  - Všetky lemmy pre každý pojem na vstupe

URL: http://text.fiit.stuba.sk:8080/lematizer
/services/lemmatizer/lemmatize/fast
Method: POST
Headers: Content-Type: text/plain
Data: Už niekoľko rokov patrí Slovensko medzi krajiny s najvyšším počtom áut vyrobených na tisíc obyvateľov.

Response
Data: už niekoľko rok patriť slovensko medzi krajina s vysoký počet auto vyrobený na tisíc obyvateľ

URL: http://text.fiit.stuba.sk:8080/lematizer
/services/lemmatizer/lemmatize/full
Method: POST
Headers: Content-Type: text/plain
Data: Mám psa.

Response
Data:
```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<customLemmasHolders>
        <customLemmasHolder>
                <form>mám</form>
                <lemmas>
                        <lema>mama</lema><rank>0</rank>
                </lemmas>
                <lemmas>
                        <lema>mať</lema><rank>0</rank>
                </lemmas>
                <lemmas>
                        <lema>mámiť</lema><rank>0</rank>
                </lemmas>
        </customLemmasHolder>
        <customLemmasHolder>
                ...podobne pes ...
        </customLemmasHolder>
</customLemmasHolders>
```

Morpholyzer

- http://morpholyzer.fiit.stuba.sk:8080/PosTagger/
- Identifikácia slovných druhov v slovenskom texte
- Všetky, prípustné, najpravdepodobnejšia
- *Prvá ľudská sonda zacielená na kométu sa blíži k cieľu.*

Európa	Európa Podstatné meno substant. par. ženský r. singulár nominatív	Európa Podstatné meno substant. par. ženský r. singulár vokatív	
chce	chcieť Sloveso prézent nedokonavý v. singulár tretia afirmácia		
pristáť	pristáť Sloveso infinitív dokonavý v. afirmácia		
na	na Predložka vokalizovaná akuzatív	na Predložka vokalizovaná lokál	na Citoslovce

kométe	kométa Podstatné meno substant. par. ženský r. singulár datív	kométa Podstatné meno substant. par. ženský r. singulár lokál
Zobudia	zobudiť Sloveso prézent dokonavý v. plurál tretia afirmácia	
sondu	sonda Podstatné meno substant. par. ženský r. singulár akuzatív	

Metallurgy

- http://metallurgyapi.eu/
- Získavanie metadát k textom

- Obsah
  - Text zo zadanej web stránky

- Tokeny
  - Rozdelí text na slová a tokenizuje

- Kľúčove slová
  - top slová v dokumente (TF/IDF)
  - možnosť zadať väčší korpus dokumentov naraz

http://metallurgyapi.eu/content.json?url=http%3A%2F%2Fnitra.sme.sk%2Fc%2F6588534%2Fnajstarsiu-rotundu-v-strednej-europe-uvidi-verejnost.html


```
{
        success: true,
        content: " Pamiatka je najstaršou stojacou rotundou
                   nielen na Slovensku, ale zrejme aj ... "
}
```

http://metallurgyapi.eu/content.json?url=http%3A%2F%2Fnitra.sme.sk%2Fc%2F6588534%2Fnajstarsiu-rotundu-v-strednej-europe-uvidi-verejnost.html


```
{
        success: true,
        tokens: [
                "pamiatka",
                "je",
                "najstaršia",
                "stojac",
                "rotund",
                "nielen",
                "na",
                "slovensku",
                "al",
                "zrejm",
                "aj",
                "v",
                ...
        ]
}
```

http://metallurgyapi.eu/keywords.json?url=http%3A%2F%2Ftechcrunch.com%2F2012%2F11%2F09%2Fcan-social-media-influence-really-be-measured%2F


```
{
    success: true,
    language: "en",
    category: "Agriculture",
    keywords:  [
        {
            keyword: "algorithm",
            rating: 2.09635,
            idf: 4.79165
        },
        {
            keyword: "score",
            rating: 1.03667,
            idf: 1.71587
        },
    ]
}
```

CollEx

- http://metallurgyapi.eu:8080/collex/
- Extrakcia kolokácií v slovenskom jazyku

- Rozdelenie textu na vety
- Identifikácia kolokácií pre dvojice a trojice susediacich slov

- *„Zajtra ráno bude pekne.“*

```
<sentences>
  <sentence>
    <bigrams>
      <bigram words="zajtra ráno">
        <score name="PmiScore" value="4.114" collocation="true"/>
      </bigram>
      <bigram words="ráno bude">
        <score name="PmiScore" value="0.325" collocation="false"/>
      </bigram>
      <bigram words="bude pekne">
        <score name="PmiScore" value="0.728" collocation="false"/>
      </bigram>
    </bigrams>
    <trigrams>
      <trigram words="zajtra ráno bude">
        <score name="PmiScore" value="2.533" collocation="true"/>
      </trigram>
      <trigram words="ráno bude pekne">
        <score name="PmiScore" value="-0.080" collocation="false"/>
      </trigram>
    </trigrams>
  </sentence>
</sentences>
```

NER

- http://mus.fiit.stuba.sk/
- Extrakcia pomenovaných entít zo slovenských textov v prirodzenom jazyku
- Vstup
  - Text
  - url

- mus.fiit.stuba.sk/ner/?type=words&content=*Zdeno Chára oslávil 500. zápas v drese Bostonu v NHL asistenciou a víťazstvom 3:2 v Ottawe*

```
[
        {
                real_name: "Zdeno Chára",
                name: "Zdeno Chára",
                typ: "P",
                start_index: 2,
                end_index: 3
        },
        {
                real_name: "Boston",
                name: "Bostonu",
                typ: "L",
                start_index: 10,
                end_index: 10
        },
        {
                real_name: "NHL",
                name: "NHL",
                typ: "O",
                start_index: 12,
                end_index: 12
        },
        ...
]
```