PANACEA Project
Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
Grant Agreement no. 248064

Documentation of Nouns Classifier Web Service[1]

[dt_noun_classifier_{classname}]

Author: Muntsa Padró. Barcelona, 2012

Contact: muntsa.padro@upf.edu

1 Overview

Given a part of speech tagged text, this webservice performs the classification of nouns as belonging or not belonging to the given class. One webservice have been deployed for each available class. The classification is performed with a pre-trained Decision Tree. The ouptut is a LMF file with the classifier prediction for each noun. You can choose to have this prediction with a numerical score or with a ternary classification (yes/no/unknown). See *optional parameters* section for details.

2 Inputs, outputs and formats

2.1 Inputs

- *input:* PoS tagged corpus in tabular format (form, lemma and pos tag), with sentences marked with <s> and UTF8 encoding. Example:

```
<s>
El          el          DA0MS0
niño        niño        NCMS000
va          ir          VMIP3S0
a           a           SPS00
la          el          DA0FS0
escuela     escuela     NCFS000
.           .           Fp
</s>
```

  You can use FreeLing Morphosyntactic tagger Web Service v.3 to get this input using the option xmlcqp. If the sentences are not marked with <s> the classifier will also work, but it may have less accuracy.

- *label:* string that will identify the classification in the resulting LMF. This is useful if you plan to perform a merging with existing dictionaries or with other lexica automatically acquired. For example, if you are using different corpora from

different domains, this string should identify the domain. Thus, after the classification you will have two lexica, indicating the nominal classes for each domain, and you can merge them keeping this domain dependent information.

- *language:* language of the input. May be Spanish (es) or English (en)

Optional parameters:

- *inputIsURLlist:* whether the input is a list of urls containg PoS tagged corpora. If it is set to "no", the input should be a text with all annotated sentences. If it is set to "yes" the input should be a list of urls that contain different sentences.
- *lemmas*: list of lemmas to classify. If it is empty, all nouns in corpus will be classified (may take a long time)
- *minOccurrences*: minimum number of times a noun has to be seen in the corpus to be classified. If a list of lemmas is given, by default minOccurrences is set to 1. If no lemma list is given, the default value for this parameter is 5.
- *output_type*:

1. *scored*: each noun gets a score (between -1 and 1) indicating the confidence of the classifier. If the score is higher than 0, the noun is considered a member of the class, scores close to 1 indicate high confidence of the classifier. If the score is below zero, it is considered a non-member of the class (with more confidence as closer to -1 is the score).

   Example:

```
<Lexicon>
    <LexicalEntry id="le_1">
        <feat att="partOfSpeech" val="noun"/>
        <Lemma>
            <feat att="writtenForm" val="boy"/>
        </Lemma>
        <Sense>
            <feat att="corpusLabel" val="labour"/>
            <feat att="hum" val="0.85"/>
        </Sense>
    </LexicalEntry>
    <LexicalEntry id="le_2">
        <feat att="partOfSpeech" val="noun"/>
        <Lemma>
            <feat att="writtenForm" val="car"/>
        </Lemma>
        <Sense>
            <feat att="corpusLabel" val="labour"/>
```

```
            <feat att="hum" val="-0.90"/>
        </Sense>
    </LexicalEntry>
    <LexicalEntry id="le_3">
        <feat att="partOfSpeech" val="noun"/>
        <Lemma>
            <feat att="writtenForm" val="employment"/>
        </Lemma>
        <Sense>
            <feat att="corpusLabel" val="labour"/>
            <feat att="hum" val="0.4"/>
        </Sense>
    </LexicalEntry>
</Lexicon>
```

2. *filtered*: the nouns are filtered according to their score. If the score is positive and over a determined threshold the noun is considered to be a member of the class. If it is negative and under another threshold, it is considered to be a non-member of the class. The other cases are tagged as "unknown", since the classifier did not give enough confidence to their classification. The used thresholds are pre-set according to some experiments, if you want to use your own thresholds, you should get the *scored* output and use the Select Nouns from LMF lexicon Web Service to filter it with your thresholds.

   Example:

```
<Lexicon>
    <LexicalEntry id="le_1">
        <feat att="partOfSpeech" val="noun"/>
        <Lemma>
            <feat att="writtenForm" val="boy"/>
        </Lemma>
        <Sense>
            <feat att="corpusLabel" val="labour"/>
            <feat att="hum" val="yes"/>
        </Sense>
    </LexicalEntry>
    <LexicalEntry id="le_2">
        <feat att="partOfSpeech" val="noun"/>
        <Lemma>
            <feat att="writtenForm" val="car"/>
        </Lemma>
        <Sense>
            <feat att="corpusLabel" val="labour"/>
            <feat att="hum" val="no"/>
        </Sense>
    </LexicalEntry>
```

```
    <LexicalEntry id="le_3">
      <feat att="partOfSpeech" val="noun"/>
      <Lemma>
        <feat att="writtenForm" val="employment"/>
      </Lemma>
      <Sense>
        <feat att="corpusLabel" val="labour"/>
        <feat att="hum" val="unknown"/>
      </Sense>
    </LexicalEntry>
  </Lexicon>
```

2.2 Outputs

- *LMFoutput*: LMF file with the classification for each noun. This file can be merged with other files obtained in PANACEA platform using the LMF file merger Web Service.
- *weka*: weka file used to classify the nouns. Useful for debugging purposes
- *notFoundLemmas*: list of lemmas that did not appear in the corpus more than the minOccurrences

2.3 Related Web Services:

- Abstract nouns classifier Web Service: http://lod.iula.upf.edu/resources/264
- Artifact nouns classifier Web Service: http://lod.iula.upf.edu/resources/265
- Eventive nouns classifier Web Service: http://lod.iula.upf.edu/resources/227
- Human nouns classifier Web Service: http://lod.iula.upf.edu/resources/243
- Location nouns classifier Web Service: http://lod.iula.upf.edu/resources/244
- Matter nouns classifier Web Service: http://lod.iula.upf.edu/resources/266
- Process nouns classifier Web Service: http://lod.iula.upf.edu/resources/267
- Semiotic nouns classifier Web Service: http://lod.iula.upf.edu/resources/268
- Social nouns classifier Web Service: http://lod.iula.upf.edu/resources/269