PANACEA Project

Platform for **A**utomatic, **N**ormalized **A**nnotation and

Cost-**E**ffective **A**cquisition
Grant Agreement no. 248064

Documentation of Weka noun signatures creator Web Service[1]

Author: Muntsa Padró. Barcelona, 2012

Contact: muntsa.padro@upf.edu

1 Overview

This web service creates a weka file containing context information of a list of nouns in a given corpus. The context information for each noun is extracted using a set of Regular Expressions and it is encoded in one vector (one line per noun in the weka file). Each slot in the vector represents the number of times the regular expression in this position has been observed with the given noun.

2 Inputs, outputs and formats

2.1 Inputs

- *corpusId:* Already indexed CQP corpus ID from which to extract the signatures. You can index your PoS tagged corpus using the CQP indexer Web Service.
- *regularExpressions:* List of Regular Expressions to be applied separated by line breaks. The order of the REs in this file will be the order in the weka vectors.

Optional parameters:

- *className:* Name of the class to be included in the weka file.
- *indicators:* Indicators file informing about the belonging of different nouns to the studied class. Format: one word per line with binary values of belonging/not belonging to the class separated by tab. In UTF-8. Example:

```
temperature	0
tea	1
```

- *lemmas*: If the information about belonging, not belonging to the class (*indicators*) is not available, you may want to include a list of nouns to be processed. The format is a list of lemmata separated by line breaks, in UTF-8. If this and indicators fields are empty, all nouns in corpus will be processed (may take a long time).

- *minOccurrences*: minimum number of times a noun has to be seen in the corpus to be included in the output file. If a list of lemmas is given, by default minOccurrences is set to 1.
- *vector_type*: type of vector desired at the output. There are three possible values:
  - frequency: number of times each cue has been seen over total number of occurrences (relative frequency).
  - absolute: number of times each cue has been seen in absolute counts.
  - binary: one if the cue has been seen once or more, zero otherwise.

3 Outputs

- *weka*: weka file with noun vectors found in the given corpus.
- *notFoundLemmas:* list of lemmas that did not appear in the corpus more than the minOccurrences threshold.
- *concordances*: sentences in the corpus in which the selected nouns appear and informationa bout which Regular Expressions matched in each sentence. Useful for developing and testing the Res.

4 Related Web Service:

Weka noun signatures creator Web Service: http://lod.iula.upf.edu/resources/226