

## Documentation of P clue/ lexical class from Weka computer Web Service<sup>1</sup>

Author: Muntsa Padró. Barcelona, 2012

Contact: [muntsa.padro@upf.edu](mailto:muntsa.padro@upf.edu)

### 1 Overview

Given a training set encoded as vectors of cue (or feature) occurrences in weka format, this web service computes  $P(\text{cue}/\text{class})$ : the probability of seeing each cue as a member or non-member of the class using MLE approach (counts frequencies of appearance of each cue in each class).

### 2 Inputs, outputs and formats

#### Inputs:

- **weka\_signatures**: classified instances encoded as cue vectors. Each slot of the vector contains the number of times each feature has been observed for that instance. Also, we add special slots: total number of occurrences in the first slot and correct class (1 or 0) and lemma (or any identifier of the instance) in the two last slots.

**Format:** The vectors should be encoded in a weka file, in UTF.8. For example, for the class of eventive nouns in English we would have some examples of eventive nouns and some of non-eventive nouns:

```
@relation eventive.arff
@attribute "[total_occurences]" numeric
@attribute "[cue_1]" numeric
@attribute "[cue_2]" numeric
...
@attribute "[cue_n]" numeric
@attribute "[eventive]" {0,1}
@attribute "[lemma]" string

@data
5,0,0,...,3,0,visa
386,0,1,...,162,0,characteristic
23,1,0,...,0,1,ceremony
270,0,2,...,0,1,assembly
```

**Important:** the cue counts must be encoded as integers, this is, no relative frequency needs to be given but the number of times each cue has been seen.

#### Outputs:

The output is a comma separated file with the frequencies each cue has been observed with members and non-members of the class. Also, information about the number of tokens in each class is given. Example:

```
#cue;data size class;data p class;data size no class;data p no class;
```

---

<sup>1</sup> This documented is licensed under a Creative Commons Attribution 3.0 Spain License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/es/>.



**PANACEA Project**  
**Platform for Automatic, Normalized Annotation and**  
**Cost-Effective Acquisition**  
*Grant Agreement no. 248064*



lc\_1;1301732;0.000883438372876;1516522;0.00137419701132;  
lc\_2;1301732;0.000520076329075;1516522;0.000600716639785;  
lc\_3;1301732;0.0243222107162;1516522;0.0177992801951;  
lc\_4;1301732;0.0;1516522;0.0;  
lc\_5;1301732;0.0511119032182;1516522;0.017487382313;

### **Related Web Service:**

P clue/ lexical class from Weka computer Web Service: <http://lod.iula.upf.edu/resources/225>