The LLaMaPUn Library

September 28, 2014

1 Stop words

Stop words are very frequent words like *we*, *and*, *is*, which are very frequent, but don't carry much meaning. These words are excluded in some NLP applications. We tried to create a math specific list of stop words. Generally, it's not always easy to decide which words should be contained in the list. It seems to be a good idea to have different stop word lists for different applications. Note that a stop word list has to take the word tokenization into account, because words like *isn't* are tokenized in different ways.

2 Language detection

Our NLP tools are designed for English texts. To avoid noise from non-English documents, we created an interface for the libTextCat library, which supports n-gram-based language identification. An experiment showed that a subset of 10 960 documents of the arxiv corpus contained 31 non-English documents. Most of them are written in French.

3 Stemmer

Stemming words reduces the vocabulary size significantly. We've utilized the morpha stemmer, in a way that allows us to easily stem words. An interesting property of the stemmer is that it often assumes that words are the Latin plural form, so it stems *formula* to *formulum* etc. However, in most cases it isn't a problem, because it is consistent. Some words can be stemmed multiple times, e.g. *tilings* is stemmed to *tiling*, which in turn is stemmed to *tile*.

4 DNM library

For the natural language processing, we usually use XHTML documents. Their internal tree structure carries useful information about the structure of the document. However, conventional natural language processing tools are designed for plain text.

The goal of the DNM library is to find a way to switch between the DOM's tree structure and the plain text. At the moment, it writes plain text offsets into the DOM, allowing us to get the plaintext corresponding to a node, and, finding a node according to an offset by using a depth-first search.

Another challenge is the normalization of certain nodes such as math formulae. Apart from that, the library supports the normalization/skipping of tables, footnotes, cites, and equation groups. Normalization means in this case replacing them by tokens such as *MathFormula*, *TableStructure*, etc. These expressions are chosen in a way, that they are easy to interpret, and yet shouldn't occur (without white spaces) in a normal text. Note that these tokens are often changed in later normalization steps. For example, after moving all the characters to lower case and stemming the words, in our setting *MathFormula* currently becomes *mathformulum*.