TreeCloud
Unitex: an increased synergy
Claude Martineau

To cite this version:
Claude Martineau. TreeCloud
hal-01702091

HAL Id: hal-01702091
https://hal-upec-upem.archives-ouvertes.fr/hal-01702091
Submitted on 19 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
TreeCloud & Unitex: an increased synergy
Claude Martineau
Projet PEPS CNRS/UPE Eclavt

Unitex/GramLab is a corpus analyser and annotation tool
- Based on Automata and RTNs with outputs
- Multilingual: Up to 22 languages (French, English, ..., Greek, ..., Korean, Thai)
- Unicode 3.0 (UTF-8, UTF-16LE, UTF-16BE)
- Cross-platform: Linux, macOS, Windows
- Open source: https://github.com/UnitexGramLab
- Website and binary installers: https://unitex-monitor.snlab pau.fr

Under development since 2001 by a group of passionate volunteers

How and Why to plug Unitex into TreeCloud?
1) Unitex transforms the input text into a new text with all the forbidden/stopping words replaced by the XXX word
2) The new text is sent to TreeCloud with XXX as the unique forbidden word (the unique word in the stop list)

Get a larger and more accurate coverage of forbidden words

Under development since 2001 by a group of passionate volunteers

Some examples of trees in different languages

TreeCloud is a tree cloud visualization of a text

TreeCloud builds a tree visualization of a text, which looks like a tag cloud where the tags are displayed around a tree to reflect the co-occurrence distance between the words in the text.

The adjacent grammar contains three pathways:
- The grey box represents a call to a subgraph that recognizes auxiliary and modal verbs.
- The second box uses some lexical masks (MODV, VCONJ, ...) to match grammatical phrases.
- The ANTIPOS-XXX lexical mask in the last box matches a list of stop words given by the user.

Get a larger and more accurate coverage of forbidden words

Under development since 2001 by a group of passionate volunteers

Some examples of trees in different languages

Tree of the 50 most frequent words in a Arabic article

Tree of the 45 most frequent compound and person nouns in a Jules Verne’s novel

2017 on-line version of TreeCloud: an improved implementation

Introduction of the concept of file processing

In the online version of 2014, there is only a single Unitex processing for each language. The necessary resources were hard-coded into the program.

In the new version, there can be several processing files for each language. Furthermore, the Spanish language (Latin and Cyrillic) has been added. In order to manage these many languages (language, processing) a concept of processing file has been set up.

For example, in the processing file below, the first line indicates the path for French resources, these pairs (language, processing) a concept of processing file has been set up.

Furthermore, the Serbian language (Latin and Cyrillic) has been added. In order to manage these many languages (language, processing) a concept of processing file has been set up.

In order to display verbs into the tree, it could be useful to get three UDDMA (since they have many influenced forms). The same principle is used for nouns with nouns or adjectives with cases (e.g., Serbian or Greek).

Some examples of trees in different languages

Tree of the 30 most frequent words in a Serbian article

Conclusion

Plug-in of Unitex into TreeCloud provides:
- A more accurate representation of forbidden words
- All kinds of multimedia to be recognized in the text and presented in the tree
- A visual representation of some grammatical or semantic categories of the words.
- A faster construction of the tree (via a careful use of the Unitex API)


Several ways to use Unitex/GramLab

- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.
- Several visualizations are possible for the annotated text that includes all the tagged patterns.

Conclusion

Plug-in of Unitex into TreeCloud provides:
- A more accurate representation of forbidden words
- All kinds of multimedia to be recognized in the text and presented in the tree
- A visual representation of some grammatical or semantic categories of the words.
- A faster construction of the tree (via a careful use of the Unitex API)