AlvisNLP/ML

Robert Bossy

Mathématique Informatique et Génome – Bibliome
Institut National de la Recherche Agronomique

26 January 2012 / Bibliome meeting

What is AlvisNLP/ML?

AlvisNLP/ML is an automatic text corpus processing pipeline.

Main features

- Modular
- Configurable
- Unified data model

Usage

`alvisnlp [OPTIONS] plan.xml`

`plan.xml`

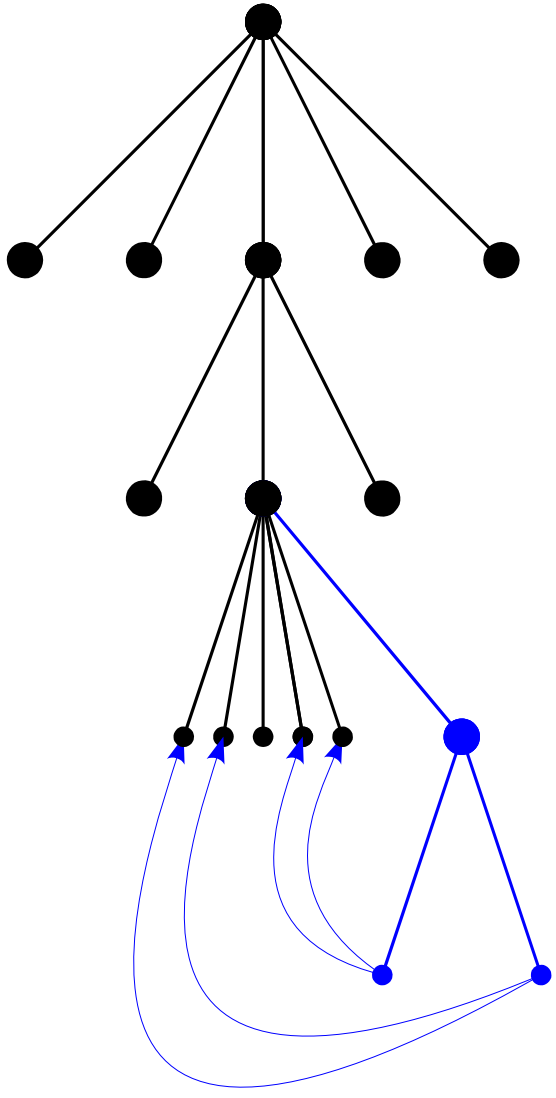
- XML file that contains the sequence of modules and their parameters.

→ reproducibility of results

Options

`-help`	display all options
`-log logfile`	save the log in the file *logfile*
`-moduleDoc module`	display documentation for *module*

Data model: limited depth hierarchy



Corpus

Documents

identifier

Sections

name, contents

Annotations

position

Relations

name

Tuples

roles

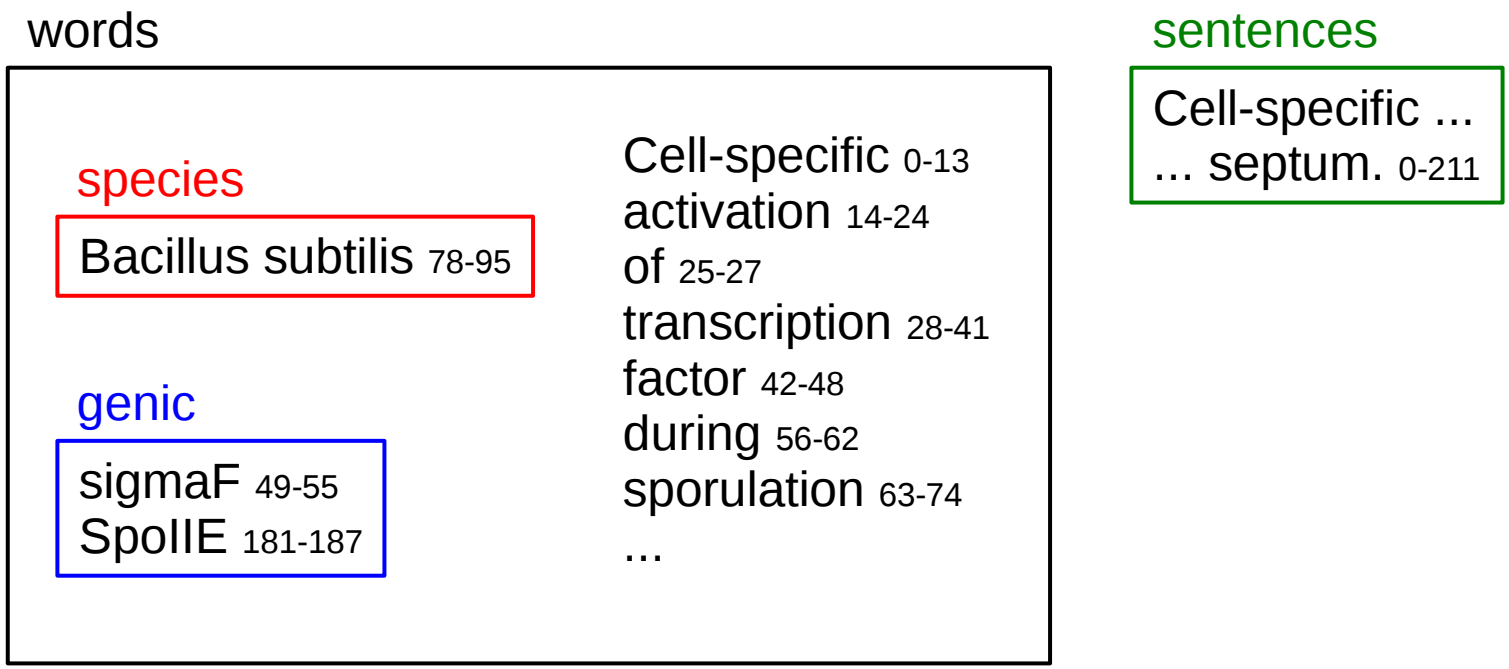
Data model: layers

Document id: 10323866

Section name: sentence1

Section contents:

Cell-specific activation of transcription factor sigmaF during sporulation in Bacillus subtilis requires the formation of the polar septum and the activity of a serine phosphatase (SpoIIE) located in the septum.



Data model: relations and tuples

genic-interaction

agent	target
SpoIIE$_{181-187}$	sigmaF$_{49-55}$

- Table = Relation
- Row header = Role
- Line = Tuple

dependencies

sentence		head	dependent
Cell-specific...	septum.$_{0-211}$	activation$_{14-24}$	Cell-specific$_{0-13}$
Cell-specific...	septum.$_{0-211}$	activation$_{14-24}$	sigmaF$_{49-55}$
Cell-specific...	septum.$_{0-211}$	sigmaF$_{49-55}$	factor$_{42-48}$
Cell-specific...	septum.$_{0-211}$	factor$_{42-48}$	transcription$_{28-41}$
		...	

Data model: features

- Features are key-value pairs.
- Any element can have features.

Documents

10323866-1 pmid=*10323866*, sentence=*1*, set=*train*

Annotations

sigmaF$_{49-55}$ pos=*NP*, canonical-form=*sigF*

Tuples

head	**dependent**	
activation$_{14-24}$	Cell-specific$_{0-13}$	label=*mod_att:N-ADJ*

Expression language

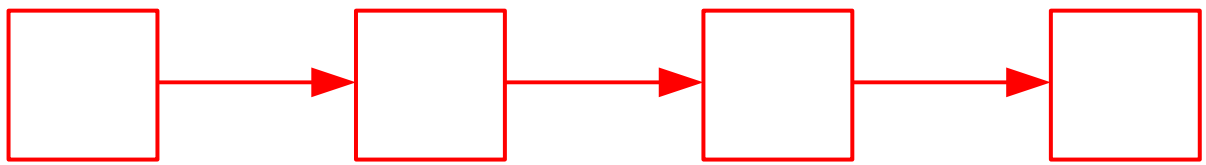
The expression language allows to navigate across the data structure. Expressions are evaluated from a *context* element.

Examples

- `documents`
  all documents of a corpus.
- `documents.sections.layer:words`
  all annotations in layer *words* in all sections of all documents of the corpus.
- `after:words(:2)`
  up to 2 annotations in layer *words* after the current annotation.
- `relations:dependencies.tuples.args:head`
  the argument with role *head* in tuples of relation *dependencies* in current section.

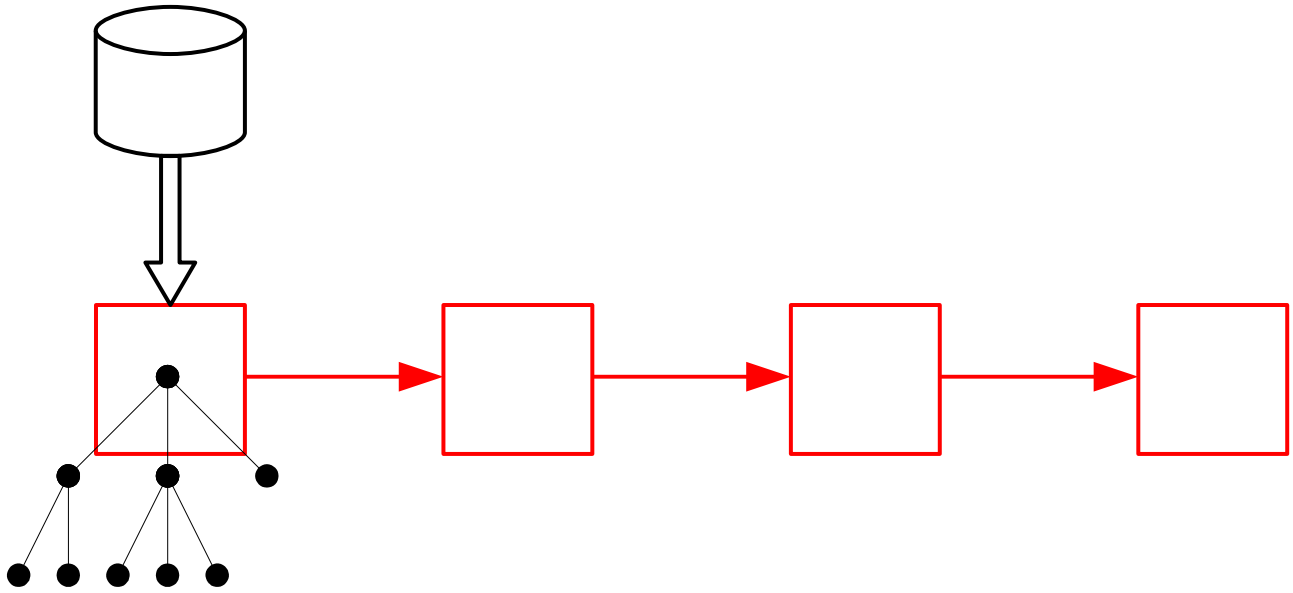
Execution model

The corpus starts empty.

- 

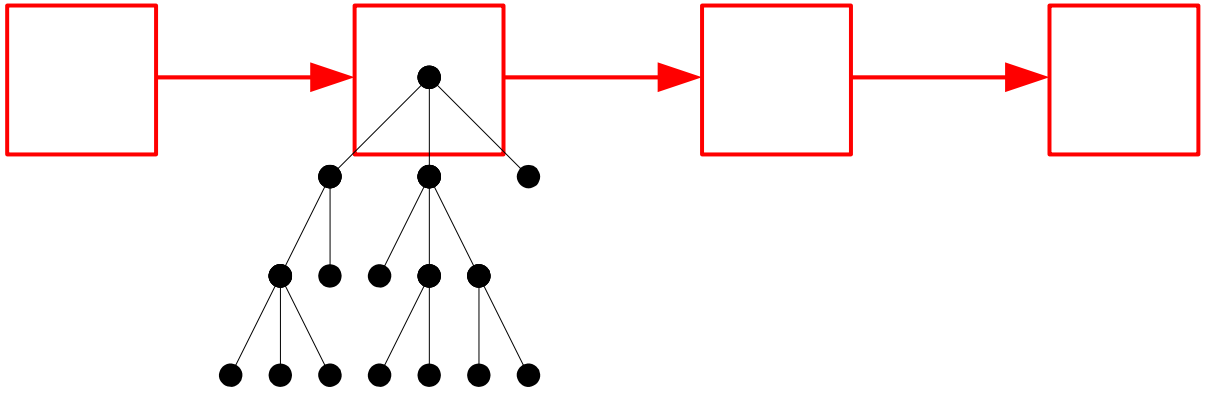
Execution model

The first module reads files, then creates documents and sections.



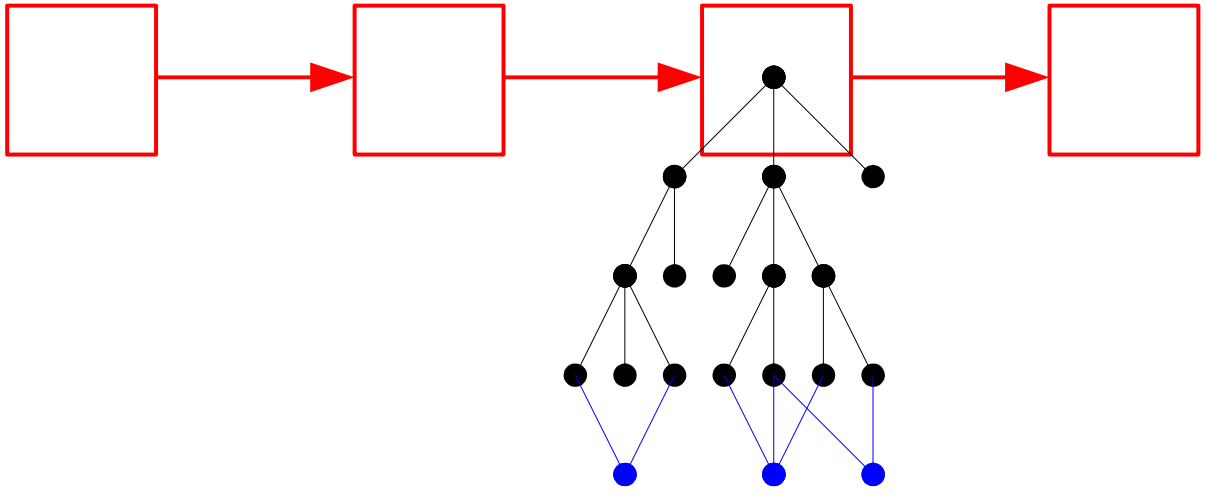
Execution model

This module creates annotations.



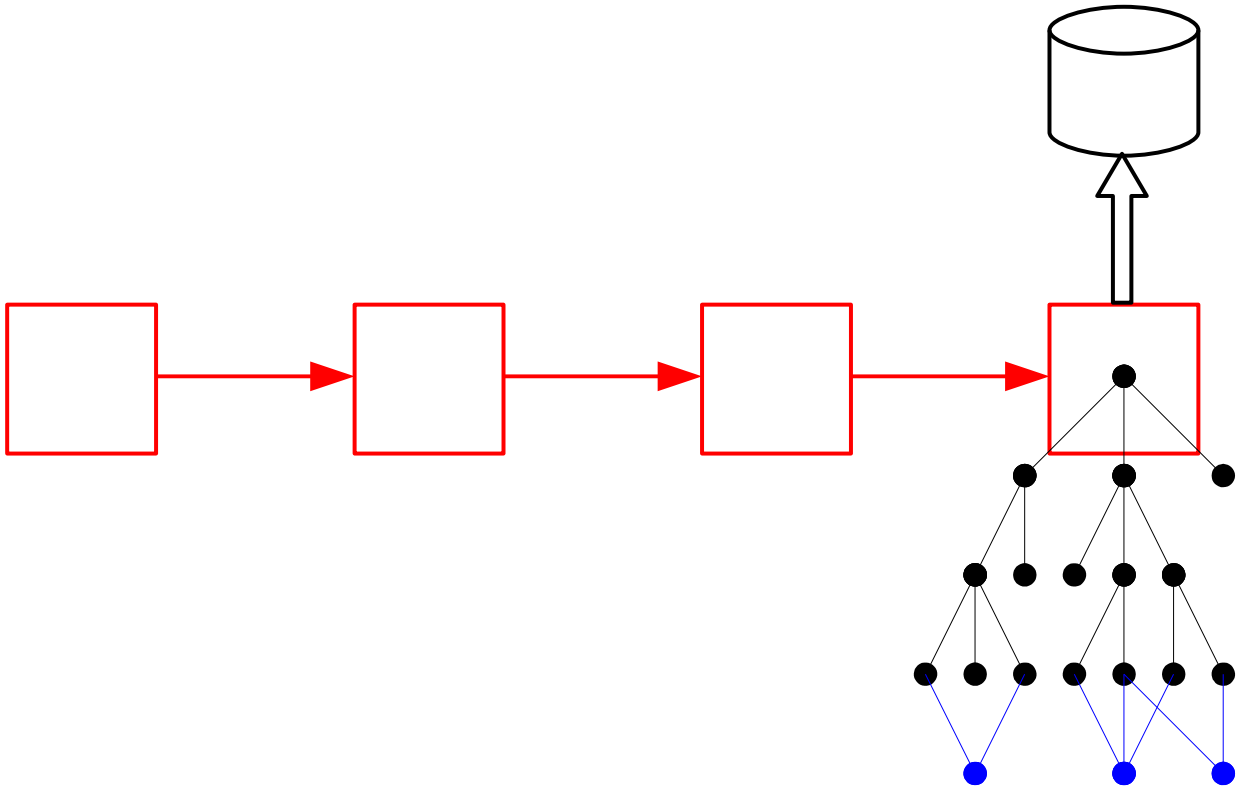
Execution model

That module creates tuples.



Execution model

The last module exports the corpus into files.



Release policy

Subversion repository

- Very latest changes.
- May NOT work properly, or even compile.
- Daily/Weekly commits.

devel

- Passes regression tests.
- Weekly/Monthly update.

Numbered version

- Compiles and passes regression tests.
- No hole in documentation.
- No experimental or obsolete modules.
- Latest: 0.4 (July 2010).

How can I contribute?

As an end user

- Criticize the documentation.
- Report bugs and feature requests.

As an experimented user

- Share your (useful) plans.
- Write some documentation.
- Design more regression tests.

As a developer

- Develop new modules.
- Develop general features.

Support

Redmine

`migale.jouy.inra.fr/redmine/projects/alvisnlp`

- Bug and feature trackers.
- Some help.
- Installation guide.
- Module documentations.

Troubleshooting

Please:

1. Give me the log (and the plan if I cannot access it).
2. Take a look at the documentation before asking me.

Tutorial

Where?

On *bibdev*:

```
$ tar -x -z -f alvisnlp_stfilter.tar.gz
$ cd stfilter
```

Environment

`JAVA_HOME`	path to Java
`CLASSPATH`	paths to external libraries (Weka)
`ALVISNLP_HOME`	path to AlvisNLP/ML

```
$ . environ.sh
```

Launch

```
$ alvisnlp common.plan
```