

LaMachine: A meta-distribution for NLP software

Maarten van Gompel and Iris Hendrickx

Centre for Language and Speech Technology (CLST)

Radboud University, Nijmegen, the Netherlands

`proycon@anaproj.nl`, `i.hendrickx@let.ru.nl`

<https://proycon.github.io/LaMachine>

Abstract

We introduce LaMachine, a unified Natural Language Processing (NLP) open-source software distribution to facilitate the installation and deployment of a large amount of software projects that have been developed in the scope of the CLARIN-NL project and its current successor CLARIAH. Special attention is paid to encouragement of good software development practises and reuse of established infrastructure in the scientific and open-source software development community. We illustrate the usage of LaMachine in an exploratory text mining project at the Dutch Health Inspectorate where LaMachine was applied to create a research environment for automatic text analysis for health care quality monitoring.

1 Introduction

Software is a key deliverable and a vital component for research in projects such as those under the CLARIN umbrella. It is software that provides researchers the instruments to yield for their research; It is CLARIN's core mission to make digital language resources, including software, available to the wider research community.

We see that NLP software often takes on complex forms such as processing pipelines invoking various individual components, which in turn rely on various dependencies. Add dedicated web-interfaces on top of that and you obtain a suite of interconnected software that is often non-trivial to install, configure, and deploy. This is where LaMachine comes in.

LaMachine incorporates software providing different types of interfaces¹ that typically address different audiences. Whilst we attempt to accommodate both technical² and less-technical audiences³, there is a natural bias towards the former as lower-level interfaces are often a prerequisite to build higher-level interfaces on. Depending on the *flavour* of LaMachine chosen, it makes a good virtual research environment for a data scientist, whether on a personal computer or on a computing cluster, a good development environment for a developer or a good deployment method for production servers in for example CLARIN centres. We demonstrate how LaMachine can create a fully functioning and standalone research environment for text mining and NLP for Dutch texts in a use case project at the Healthcare Inspectorate.

2 Architecture

Being an open-source NLP software distribution, LaMachine is constrained to POSIX-compliant platforms; this primarily means Linux, but also BSD and, with some restrictions, macOS. This stands in contrast with mobile platforms (Android/iOS/etc), native Windows/mac desktop software, or certain interface types in general such as classical desktop GUI applications, all of which fall beyond the scope of LaMachine. Software that is incorporated must 1) bear some relevance to NLP, 2) be under a recognised open-source license, 3) be deposited in a public version controlled repository⁴ and 4) have a release protocol (with semantic versioning) using the proper technology-specific channels.

¹Command line interfaces, programming interfaces, web-user interfaces, webservices.

²Data scientists, DevOps, system administrators, developers.

³The wider researcher community, particularly the Humanities; also educational settings.

⁴e.g. Github, Gitlab, Bitbucket

LaMachine is a *meta distribution* as it can be installed in various contexts. At its core, LaMachine consists of a set of machine-parsable instructions on how to obtain, install and configure software. These are implemented using Ansible⁵. This is notably different from the more classical notion of Linux distributions, which generally provide their own repositories with (often binary) software packages. LaMachine builds on this already available infrastructure by taking these repositories as a given and only needs to know which repositories to use. Similarly, there are different programming-language-specific ecosystems providing their own repositories, such as the Python Package Index⁶ for Python, CRAN⁷ for R, CPAN⁸ for Perl, Maven Central⁹ for Java. LaMachine again relies on those to pull and install software from and never forks, archives, or modifies the software in any way. In doing so, we compel participating software projects to adhere to well-established distribution standards and ensure the software is more sustainable towards the future (?). Moreover, we ensure that LaMachine never becomes a prerequisite for the software but merely a courtesy or convenience.

LaMachine provides ample flexibility that allows it to be deployable in different contexts. First of all there is flexibility with regard to the target platform, where we support several major GNU/Linux distributions (Debian, Ubuntu, CentOS, RedHat Enterprise Linux, Fedora, Arch Linux), as well as macOS. Second, there is flexibility with regard to the form, where we support *containerisation* through Docker¹⁰, *virtualisation* through Vagrant¹¹ and VirtualBox¹², direct remote provisioning through Ansible (for production servers), or an installation that is either global to the machine or local in a custom directory for a specific user. Pre-built docker containers and virtual machines with a limited selection of participating software are regularly uploaded to the Docker Hub and Vagrant Cloud, respectively.

With LaMachine we also attempt to harmonise the metadata of all installed software, by converting metadata from upstream repositories to the CodeMeta standard¹³ (?; ?) where possible, or encouraging software developers to provide their own. This in turn enables other tools to do proper service discovery and provenance logging.

3 Software

As LaMachine was conceived initially¹⁴ as the primary means of distribution of the software stack developed at CLST, Radboud University Nijmegen, it includes a lot of our software. A full list goes beyond the scope of this overview; we will merely mention some CLARIN-NL/CLARIAH-funded tools: ucto (a tokeniser), Frog (an NLP suite for Dutch), FoLiA (Format for Linguistic Annotation, with assorted tools), FLAT (a web-based linguistic Annotation tool), PICCL (an OCR and post-OCR correction pipeline) and CLAM. We already include some relevant software by the University of Groningen and VU Amsterdam as well. Moreover, LaMachine incorporates a large number of renowned tools by external international parties, offering most notably a mature Python environment¹⁵, but also R and Java.

4 Case study

We participated the ZonMw project entitled “Tekstmining in het toezicht: een exploratieve studie naar de automatische verwerking van klachten ingediend bij het Landelijk Meldpunt Zorg”¹⁶ led by IQhealthcare¹⁷, the scientific center for healthcare quality of RadboudUMC. This project took place

⁵<https://www.ansible.com>

⁶<https://pypi.org>

⁷<https://cran.r-project.org/>

⁸<https://www.cpan.org>

⁹<https://search.maven.org>

¹⁰<https://www.docker.com>

¹¹<https://vagrant.org>

¹²<https://www.virtualbox.org>

¹³<https://codemeta.github.io/>, described in JSON-LD

¹⁴LaMachine exists since May 2015 and has been used extensively ever since by numerous users

¹⁵scipy, numpy, scikit-learn, matplotlib, nltk, spacy, pytorch, keras, gensim, etc...

¹⁶project website: <https://www.zonmw.nl/nl/onderzoek-resultaten/kwaliteit-van-zorg/programmas/project->

¹⁷<http://www.iqhealthcare.nl/nl/>

at the Dutch Health Inspectorate and aimed to apply text mining techniques to health care complaints that have been registered at the national contact point for health care (Landelijk Meldpunt Zorg <https://www.landelijkmeldpuntzorg.nl>) We investigated the usefulness of text mining to categorise and cluster complaints, to automatically determine the severity of incoming complaints, to extract patterns and to identify risk cases. This project turned out to be a good test case of the applicability and usefulness of the LaMachine software as a standalone research environment. As the complaint data is highly sensitive, this data could not leave the secure servers from the health inspectorate and was stored in an environment without internet access. We needed to bring the software to the data via a shared folder.

We created a pre-compiled virtual image of LaMachine¹⁸ and succeeded in running it as a 64-bits Arch Linux-based VM inside another VM with Windows 2012 in which we did have administrative rights but no internet access. In terms of hardware we ran on a machine with 8 cores and 32gb internal memory available. LaMachine provided a fully functional research environment and we ran all our experiments within LaMachine. We interacted with LaMachine both through the command line, which offers a standard shell and enables access to all lower-level tools and programming languages; or through the (offline) webbrowser to apply the Jupyter Notebook environment¹⁹, which offers a more graphical scripting environment for e.g. Python and R that has gained great popularity in the scientific community.

LaMachine comes with some simple data sharing facilities that allowed us to access the sensitive complaint data via a single shared dataspace between host and the VM. (Note that extensive data search and management functions are deliberately beyond the scope of LaMachine, and left to more high-level tooling.)

We used many of the available tools in LaMachine within this project: Frog for linguistic annotation of the textual content of the complaints, Timbl and the scikit-learn Python package (?) for classification, T-scan for feature creation and colibri-core for n-gram analysis.

5 Conclusion & Future work

The recent release of LaMachine v2, which constituted a full rewrite, has opened up LaMachine to outside contribution. Contributor documentation has been written, and at this stage, we greatly welcome external participants to join in. Use cases as the example in section 4 contributing to thorough testing and running of LaMachine in less ideal circumstances such as nested VM constructions and offline usage.

Aside from the incorporation of new relevant software, the main objectives for the future are to provide greater *interoperability* between the included tools through better *high-level interfaces* for the researcher. We see this as a bottom-up process and have now established a firm foundation to build upon. Note that such proposed interfaces, including the current portal application in LaMachine, are always considered separate independent software projects, which may be deployed by/in/for LaMachine, but also in other contexts. LaMachine remains ‘just’ a software distribution at heart.

Development of LaMachine presently takes place in the scope of the CLARIAH WP3 Virtual Research Environment (VRE) project²⁰, which has higher ambitions in accommodating the researcher and connectivity of data and services, and transcends also those of the CLARIN Language Resource Switchboard (?). An important part of our future focus will therefore be on interoperability with the higher-level tools emerging from the VRE efforts, but also with other parts of the CLARIN infrastructure; single-sign on authentication being a notable example here.

Acknowledgement

This research was funded by NWO CLARIN-NL, CLARIAH and the ZonMw project *Tekstmining in het toezicht: een exploratieve studie naar de automatische verwerking van klachten ingediend bij het Landelijk Meldpunt Zorg*, project number 516004614. We thank all project partners: the Dutch Health Inspectorate, IQhealthcare, and Tim Voets for their valuable contributions and help in the ZonMw project.

¹⁸ Installation of LaMachine begins with a single bootstrap command, see <https://proycon.github.io/LaMachine>

¹⁹ <https://jupyter.org/>

²⁰ <https://github.com/meertensinstituut/clariah-wp3-vre>