Leopoldo **Pla Sempere**

Software Architect · Part-time lecturer

b64encoded: bGVvcG9sZG9wbGFzZW1wZXJlQGdtYWlsLmNvbQ== | lpla.github.io | lpla | leopoldoplasempere

Experience

Part-time lecturer

San Vicente del Raspeig

University of Alicante

Oct. 2019 - present

- Teaching in **Mathematics and Computer Science** degrees
- Programming algorithmic fundamentals and advanced techniques for software specification
- Use real-world examples and challenges from Google, Amazon, Microsoft and Spotify.

Senior Data Scientist, Natural Language Processing and Machine Learning engineer

San Vicente del Raspeig

University of Alicante

Nov. 2017 - present

- Lead developer in **EU funded projects** in coordination with
  - University of Edinburgh, Prompsit, TAUS, Omniscien and John Hopkins University:
    - **ParaCrawl** (Provision of Web-Scale Parallel Corpora for Official European Languages)
    - **Paracrawl 2** (Broader Web-Scale Provision of Parallel Corpora for European Languages)
    - **Paracrawl 3** (Continued Provision of Web-Scale Parallel Corpora for European Languages) (paracrawl.eu)
  - University of Groningen, Jožef Stefan Institute, Prompsit:
    - **MaCoCu**: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages (macocu.eu)
- My work focus on **creating parallel corpora** for EU official languages **by a broad web crawling effort**.
- Coordinated the development of the **Open Source suite Bitextor** [bitextor/bitextor on Github] (bash, python, C++, snakemake, awk...), the production-ready tool that merges all this necessary technology chain for these CEF projects, run under **HPCC-Cloud scale** (Cambridge Service for Data-Driven Discovery, CSD3).
- 6 people team under my supervision.
- Cloud resources provided by **Amazon (AWS S3), Microsoft (Azure) and Docker**.
- OS administration of various Linux-based systems (Scientific Linux, Ubuntu, Fedora, Debian)

Data Scientist, Natural Language Processing and Machine Learning engineer

Elche

Prompsit Language Engineering

May 2015 - Oct. 2017

- Developing natural language related projects as Reverso Context, processing full range of corpora and language resources
  - developed for Reverso (Softissimo, approx. rank 300 in Global Alexa Ranking)
  - **several million page-views per day**, big data Cloud scale computing
  - grown in an agile development team
  - using technologies as SVN, pandas, jupyter, bash (UNIX tools), word2vec, keras, scipy and Apache Solr
  - Language resources I was developing were in all available languages in the app (Arabic, German, English, French, Hebrew, Italian, Japanese, Dutch, Polish, Portuguese, Romanian and Russian)
  - High responsibilities in product releases and Quality Assessment.

Full-stack developer intern

Elche

Prompsit Language Engineering

Feb. 2013 - May 2013

- Developing an internal news sentiment analysis suite
- Python and bash scripting for news crawling and plain text extraction from all downloaded HTML data
- AJAX, jQuery, CSSv3 and PHP for real-time interactive interface for dataset generation
- MySQL for dataset storage

Education

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Valencia

Polytecnhic University of Valencia

Sep. 2014 - Jun. 2015

- GPA: 8.81/10
- M.Sc. specialised in **natural language processing and deep learning**
- Coursed machine learning MOOCs of Andrew Ng (Stanford University) and Yaser Abu-Mostafa (Caltech)
- Master Thesis: Audio classical composer identification in MIREX 2015: submission based on Structural Analysis of Music

Computer Science Degree

San Vicente del Raspeig

University of Alicante

Sep. 2010 - Jun. 2014

- GPA: 8.53/10
- B.Sc. specialized in Informatics (robotics, machine learning, artificial vision, compilers)
- **Enroled in the group of high profile academics programme**
- Diploma Thesis: Dodecaphonic music composer assistant with OpenMusic

Publications & Awards

Building Domain-specific Corpora from the Web: the Case of European Digital Service Infrastructures — *Marseille, France*

LREC 2022 — *June, 2022*

van Noord et al.

MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages — *Ghent, Belgium*

EAMT 2022 — *June, 2022*

Bañón et al.

ParaCrawl: Web-Scale Acquisition of Parallel Corpora — *Online*

ACL 2020 — *6 July, 2020*

Bañón et al.

Computer Science Degree graduation and M.Sc scholarship — *San Vicente del Raspeig*

Best academic record — *2015*

University of Alicante, Generalitat Valenciana

B.Sc thesis defense — *San Vicente del Raspeig*

Highest mark (outstanding with honors) — *2014*

University of Alicante

Presentation & Associations

EAMT — *Remote*

Webmaster for the European Association of Machine Translation — *July 2021 - Present*

- Technical role managing services as eamt.org, lists.eamt.org, mt-archive.net and several EAMT congress websites

DockerCon 2022 — *Remote*

Presenter for <The Bitextor Pipeline> — *May. 2022*

- Introduced the Bitextor pipeline and explained its uses with Docker

Docker Community All-Hands #4 2021 — *Remote*

Presenter for <The Bitextor Pipeline> — *Dec. 2021*

- Introduced the Bitextor pipeline and explained its uses with Docker, including a live demo

Skills

Natural Language Processing tools	Moses, Marian, Bleualign, Bitextor, Bicleaner
Scripting and tools	git, CMake, tmux, Python, Lisp, Bash, AWK, Perl, systemd
Machine Learning Toolkits	TensorFlow, Keras, PyTorch, pandas, sklearn, numpy, scipy, matplotlib
Programming	C++, C, Java, Golang, JAVA, Rust, LaTeX, CMake, Snakemake
Languages	English (proficient), Spanish (native), Catalan (native)

Extracurricular Activity

Clarinet Professional degree — *Elche*

Professional Conservatory of Music of Elche — *Sep. 2004 - Jun. 2010*

- Specialized in jazz improvisation with Miguel García Ferrer
- MOOCs on jazz improvisation with Gary Burton (Berklee School of Music)
- MOOC on fundamentals of rehearsing music ensembles with Dr. Evan Feldman (University of North Carolina at Chapel Hill)
- After studies, gained experience in several youth orchestras, symphonic wind orchestras and rock, pop, swing and jazz bands, writing arrangements, improvising, while playing clarinet and piano as soloist. Non-professional experience with saxophone, guitar, ukulele, bass and drums.

Hardware-related projects

Self-taught

- Open source PCB layout design using EAGLE and KiCad, with real production in factories as OSHPark and JLCPCB, for 80's computers.
- Electronic repairs and programming on vintage video-game platforms. Winner of the "best sound" award at "Retroconsolas Alicante 2013" for the Amstrad CPC 464.
- Open source 3D design (Autodesk), modeling and printing. Volunteer with "3D Makers Elche" group during COVID-19 lockdown by 3D-printing more than two hundred face-masks for near hospitals.