

# Antoine SIMOULIN, PhD

📞 +1 312-447-4198 • ✉ antoine.simoulin@gmail.com • <https://simoulin.io>

## EXPERIENCES

---

### Expert Data Scientist, NLP

🏢 [Quantmetry](#) 📍 [Paris](#) 📅 [April 2017 – July 2022](#)

Known for its expertise in artificial intelligence, Quantmetry is a high-end consulting firm in Paris. Together with a data-science team of business and technical consultants, I participated in multiple projects to automate or optimize processes. I [led end-to-end projects for real-world problems](#). My role went [beyond implementation since I contributed from ideation, framing, deployment, to monitoring](#). Here are some projects and initiatives I contributed to:

- Putting neural models in [production](#) for classifying, summarizing, and automating email replies in one of the largest French insurance companies;
- [Deployment](#) for a solution of predictive maintenance with an international car manufacturer;
- [Statistical study](#) in breast cancer pathology. Design of a structuration method using natural language processing algorithms for medical unstructured records.

### Quantitative Analyst Intern

🏢 [Crédit Agricole Corporate and Investment Banking](#) 📍 [New York & Paris](#) 📅 [September 2015 – August 2016](#)

This one-year internship was completed as quantitative analyst in the securitization team from Paris and New York.

- Challenging work during the set-up of new worldwide operations. [Collaboration with clients, notation agencies and with intern teams](#) for analysis and interpretation of the data;
- I implemented and improved [Monte-Carlo's algorithms using CUDA](#) on graphic cards for the capital calculation of an internal insurance.

## EDUCATION

---

### PhD, Computer Science (NLP)

🏢 [University of Paris Cité](#) 📍 [Paris](#) 📅 [February 2019 - July 2022](#)

My PhD entitled, *Sentence embeddings and their relation with sentence structures*, focuses on [Natural language Processing](#) methods for building sentence embeddings. My work—advised by Professor Benoit Crabbé, member of LLF lab—studies how compositionality might be leveraged through neural network structures and linguistic biases. I design and [implement innovative neural networks](#) following tree or graph syntactic patterns inspired by linguistic insights. Along with linguistics, I scale these architectures and pre-train large language models such as a version of GPT-2 for French with over a [billion parameters](#).

### Dual Master Program (MSc), Data Science

🏢 [Ecole Polytechnique](#) 📍 [Paris](#) 📅 [2016 - 2017](#)

The [leading French research, academics, and innovation institution](#). Mathematical and numerical analysis. Statistical learning, machine learning. Very large-scale calculations and the control of mechanisms of distributed databases.

### Master of science (MSc), Simulation and Mathematical Engineering

🏢 [ENSTA Paris](#) 📍 [Paris](#) 📅 [2013 – 2017](#)

French engineering school accessible through selective classe préparatoire. Computer science & mathematics degree with [major in differential optimization](#): steepest descent methods, penalization, duality algorithm and simplex algorithm, high dimensional minimization, non-differential optimization and proximal methods, practical implementations in C++. Last year advised by Prof. Pierre Carpentier, director of UMA lab.

## SKILLS

---

### Languages

[English](#) Fluent (TOEIC 965/990) [German](#) Working knowledge

### Microsoft Office & Web

In-depth knowledge of Microsoft Office: [Excel](#), [VBA](#), [Word](#), [PowerPoint](#), [Access](#).  
Basic understanding of web design: [HTML](#), [CSS](#).

### Programming skills

In-depth knowledge of [Python](#), and familiar with [R](#), [Matlab](#), [C](#), [C++](#)  
In-depth knowledge of [SQL](#), and familiar with [Hadoop](#), [Spark](#)  
In-depth knowledge of [Linux/Unix/Shell environments](#)

### Miscellaneous

DIY, 3D printing, Badminton, Scuba diving.

# RESEARCH

---

## Publications

**Unifying Parsing and Tree-Structured Models for Generating Sentence Semantic Representations**, **NAACL 2022**, Student Research Workshop, Antoine Simoulin, Benoit Crabbé

**How Many Layers and Why? An Analysis of the Model Depth in Transformers**, **ACL 2021**, Student Research Workshop, Antoine Simoulin, Benoit Crabbé

**Contrasting Distinct Structured Views to Learn Sentence Embeddings**, **EACL 2021**, Student Research Workshop, Antoine Simoulin, Benoit Crabbé

**Generative Pre-trained Transformer in \_\_\_\_\_ (French)**, **TALN 2021**: Traitement Automatique des Langues Naturelles, Antoine Simoulin, Benoit Crabbé

**Deep Learning : des usages contrastés dans le monde socio-économique**, **Statistique et Société**, 8: 55-108, Rémi Adon, Abdellah Kaid Gherbi, Florian Arthur, Aurélia Nègre, Guillaume Basquias, Antoine Simoulin, Guillaume Hochard, Fouad Talaoui-Mockli, Nicolas Bousquet

**An innovative solution for breast cancer textual big data analysis**, **In submission**, Nicolas Thiebaut, Antoine Simoulin, Karl Neuberger, Issam Ibnouhsein, Nicolas Bousquet, Nathalie Reix, Sébastien Molière, Carole Mathelin

**Impact du dépistage : une expérience française**, **Mise à jour du Collège National des Gynécologues et Obstétriciens Français**, Carole Mathelin, Jules Colin, Sébastien Molière, Audrey Fleury, Christelle Linck, Marie Paté, Catherine Guldenfels, Antoine Simoulin, Karl Neuberger, Jeremie Jégu

## Teaching

I taught a [graduate level course in natural language processing \(NLP\)](#) at [Paris Cité University](#) between 2020 and 2022. The course includes 7 sessions and introduces statistical models (TF-IDF, Bag-of-Words, LDA, Embeddings, language models) for NLP. Around [25 students](#) from the mathematics department followed the course each year.

## Talks and Presentations

### **Pre-trained neural networks for text generation and their implications**

📅 [Machine Learning Meetup](#) 📍 [Epitech](#) engineering school, Nantes France 📅 [April 2021](#)

Around 30 students and professionals in the field of data science attended the talk. I presented my paper about the first large pre-trained generative model in French.

### **Implementing and deploying natural language processing projects**

📅 [AI Paris](#) 📍 [Paris](#) 📅 [December 2019](#)

Around 800 professionals in the field of data science attended the presentation. We presented the project of email classification at MAIF and the challenges to deploy a project in production.

### **Melusine open-source release**

📅 [BigData Paris](#) 📍 [Paris](#) 📅 [December 2019](#)

Open source release of Melusine, a library for email processing. Around 80 professionals in the field of data science attended the presentation.

### **Senometry project: analysis of textual medical records for structured data extraction**

📅 [NLP Meetup](#) 📍 [Paris](#) 📅 [May 2018](#)

Presentation to around 40 professionals in the field of data science. The research project consists in using NLP methods to automatically analyze data from medical records.

## Open Source contributions

- **GPT-fr** is a French large pre-trained language model for French. The base version, equivalent to OpenAI GPT in English, includes above one billion parameters.
- **PyTree** implements tree-structured neural networks in PyTorch. The package provides highly generic implementations as well as efficient batching methods. The project was listed among the winners of the **PyTorch Annual Hackathon 2021**.
- **Sentence embedding** pre-trained model trained on 1B sentence pairs. The project was listed among the winners during the **Hugging Face Community week using JAX/Flax for NLP & CV 2021**.
- **Melusine** is a high-level Python library for email processing developed by Quantmetry and MAIF.