Fast automatic intertextuality detection in a large textual corpus

Practical information

**Supervisor:** Gregory Kucherov (senior CNRS researcher)

**Co-supervisor:** Philippe Gambette (associate professor at UPEM)

**Contact information:** gregory.kucherov@u-upem.fr, philippe.gambette@u-pem.fr

**Location:**
LIGM (CNRS, UPEM, ESIEE, Ecole des Ponts ParisTech)
5 boulevard Descartes
77420 Champs-sur-Marne (RER A Noisy-Champs)

Context

Starting from the beginning of the 2000s, the process of massive digitization of printed documents has been underway, resulting in a multitude of digital corpora comprising a total of millions of documents. This evolution requires the development of new **big-data-oriented** algorithms for searching large corpora of human language texts, and more specifically for detecting **intertextuality** -- various similarity relationships between pairs of documents. Besides the size of the data, this task is further complicated by different types of intertextuality one may want to detect and, on the other hand, by OCR errors induced by the digitization process.

Objectives

In this internship, we will work with a collection of about 60 GB (200 000 files) extracted from Gallica. Gallica (http://gallica.bnf.fr/) is an open digitized collection of Bibliothèque Nationale de France containing over 3 million documents including many OCRed text documents (books, periodicals, press articles, ...). Our goal will be to develop and implement new algorithms for intertextuality detection that will be tested on this collection. These algorithms will be inspired by methods developed in bioinformatics for searching large databases of DNA sequences. In particular, they will follow the *filtering* approach and will rely on efficient data structures for representing the corpus. Advanced algorithmic techniques such as *spaced seeds* and *Bloom filters* will be used.

This internship may be followed by a PhD.

References

- R. Horton, M. Olsen, and G. Roe (2010), *Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections*, *Digital Studies / Le champ*
numérique 2(1)


