Dévoiler les sources du théâtre classique français à travers l'alignement et la comparaison automatiques de textes

Un stage en traitement automatique des langues, algorithmique et fouille de données au LIGM, Laboratoire d'Informatique Gaspard Monge (Université Gustave Eiffel & CNRS), Champs-sur-Marne (France), dans l'équipe Modèles et algorithmes, encadré par Philippe Gambette (http://igm.univ-mlv.fr/~gambette/, philippe.gambette@univ-eiffel.fr)

Directeur du laboratoire : Stéphane Vialette (stephane.vialette@univ-eiffel.fr)

Contexte

Pendant son doctorat à Sorbonne Université, Céline Fournial a analysé les sources du théâtre classique français, créant une base de données de centaines de pièces en français publiées ou jouées de 1550 à 1650, inspirées de sources antiques, religieuses, espagnoles, italiennes ou françaises (https://celinefournial.github.io/hyperpieces/). Le texte intégral est disponible pour plus de 200 de ces pièces, ce qui permet des recherches automatiques de similarités entre textes, à l'aide d'algorithmes et d'approches inspirées de la bioinformatique.

Objectifs

L'objectif de ce stage est de développer des outils automatiques d'aide à l'étude des sources des pièces de théâtre.

La première étape consistera en une analyse détaillée des pièces et de leurs sources dans le corpus de Céline Fournial (en collaboration avec elle), grâce au développement d'outils d'alignement des pièces et de leurs sources à différents niveaux, en fonction de leur degré de similarité. Ces similarités pourront concerner non seulement des séquences de mots consécutifs (n-grammes) ou des distributions de mots, mais aussi des similarités entre structures de pièces (distribution des actes et scènes, relations entre personnages).

La seconde étape consistera à développer un algorithme pour proposer automatiquement des sources possibles pour une pièce de théâtre, avec un score de confiance pour chacune.

Ce stage conduira à développer des algorithmes rapides et efficaces, utilisant des outils ou techniques issues du traitement automatique des langues, de l'algorithmique du texte, de la théorie des graphes ou de l'apprentissage automatique.

Références

- Alfie Abdul-Rahman, Glenn Roe, Mark Olsen, Clovis Gladstone, Richard Whaling, Nicholas Cronk, Robert Morrissey and Min Chen (2017), « Constructive Visual Analytics for Text Similarity Detection », *Computer Graphics Forum* 36(1), p. 237-248, https://doi.org/10.1111/cgf.12798.
- Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman (1990), « Basic local alignment search tool », *Journal of Molecular Biology* 215(3), 1990, p. 403-410, http://doi.org/10.1016/S0022-2836(05)80360-2.
- Mohamed Amine Boukhaled, Zied Sellami et Jean-Gabriel Ganascia (2015), « Phœbus : un logiciel d'extraction de réutilisations dans des textes littéraires », Sessions démonstrations de *TALN 2015*, https://www.aclweb.org/anthology/2015.jeptalnrecital-demonstration.2.pdf.
- Céline Fournial (2019), *Imitation et création dans le « théâtre moderne » (1550-1650) : la question des cycles d'inspiration*, thèse de doctorat en littérature et civilisation française à Sorbonne Université.
- Russell Horton, Mark Olsen, Glenn Roe (2010), « Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections ». *Digital Studies/le Champ Numérique* 2(1), http://doi.org/10.16995/dscn.258.
- Zied Sellami, Jean-Gabriel Ganascia et Mohamed Amine Boukhaled (2015) « MEDITE : logiciel d'alignement de textes pour l'étude de la génétique textuelle », Sessions démonstrations de *TALN 2015*, https://www.aclweb.org/anthology/2015.jeptalnrecital-demonstration.1.pdf.

Compétences attendues

Compétences algorithmiques, programmation en Python, curiosité pour la recherche interdisciplinaire

Contact

Merci de contacter Philippe Gambette (philippe.gambette@univ-eiffel.fr) pour toute question relative à ce stage, en joignant un CV s'il vous intéresse.

Uncovering the sources of French classical theater through automatic text alignment and comparison

An internship in natural language processing, algorithmics and data mining, at LIGM, Laboratoire d'Informatique Gaspard Monge (Université Gustave Eiffel & CNRS), Champs-sur-Marne (France), in the team Models and algorithms, supervised by Philippe Gambette (http://igm.univ-mlv.fr/~gambette/, philippe.gambette@univ-eiffel.fr)

Head of the lab: Stéphane Vialette (stephane.vialette@univ-eiffel.fr)

Context

During her doctorate at Sorbonne Université, Céline Fournial uncovered the sources of classical French theater and created a database of hundreds of French plays published or played from 1550 to 1650 with antique, religious, Spanish, Italian or French sources (https://celinefournial.github.io/hyperpieces/). The full text is available for more than 200 of those plays: this allows to automatically find similarities between texts, with algorithms and approaches inspired from bioinformatics.

Goals

The goal of this internship is to build automatic tools to help studying the sources of theater plays.

It will include as a first step a detailed analysis of the plays and their sources in Celine Fournial's corpus (in collaboration with her), by programming tools aligning the plays and their sources at different levels, depending on their similarities. These similarities may cover not only word sequences (n-grams) or word distributions, but also the play structures (distributions of acts and scenes, relationships between characters).

The second step will consist in building an algorithm to automatically provide candidate sources for a theater play, with a confidence score for each.

This internship will lead to develop fast and efficient algorithms, using tools or techniques from natural language processing, stringology, graph theory or machine learning.

References

- Alfie Abdul-Rahman, Glenn Roe, Mark Olsen, Clovis Gladstone, Richard Whaling, Nicholas Cronk, Robert Morrissey and Min Chen (2017), "Constructive Visual Analytics for Text Similarity Detection", *Computer Graphics Forum* 36(1), p. 237-248, https://doi.org/10.1111/cgf.12798.
- Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman (1990), "Basic local alignment search tool", *Journal of Molecular Biology* 215(3), 1990, p. 403-410, http://doi.org/10.1016/S0022-2836(05)80360-2.
- Mohamed Amine Boukhaled, Zied Sellami et Jean-Gabriel Ganascia (2015), "Phœbus : un logiciel d'extraction de réutilisations dans des textes littéraires", Sessions Démonstrations *TALN 2015*, https://www.aclweb.org/anthology/2015.jeptalnrecital-demonstration.2.pdf.
- Céline Fournial (2019), *Imitation et création dans le "théâtre moderne" (1550-1650) : la question des cycles d'inspiration*, PhD thesis at Sorbonne Université.
- Russell Horton, Mark Olsen, Glenn Roe (2010), "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections". *Digital Studies/le Champ Numérique* 2(1), http://doi.org/10.16995/dscn.258.
- Zied Sellami, Jean-Gabriel Ganascia et Mohamed Amine Boukhaled (2015) "MEDITE : logiciel d'alignement de textes pour l'étude de la génétique df.textuelle", Sessions Démonstrations *TALN 2015*, https://www.aclweb.org/anthology/2015.jeptalnrecital-demonstration.1.p

Expected skills of the student

Algorithmic skills, programming skills in Python, curiosity for interdisciplinary research.

Contact

Please contact Philippe Gambette (philippe.gambette@univ-eiffel.fr) for any question about this insternship, or with a CV to apply.