Offre de Stage
Traitement de textes scientifiques pour la catégorisation d'expressions d'incertitude (6 mois)

Plusieurs stages sont proposés par le laboratoire C.R.I.T de l'Université de Franche-Comté à Besançon, financés par les projet ANR InSciM, dans le domaine du Traitement Automatique des Langues (TAL).

Description du poste

L'objectif du stage est de traiter des corpus d'articles scientifiques en anglais provenant de différents sources, afin de pouvoir identifier et catégoriser les segments textules porteurs d'incertitude. Les traitements développés s'intégreront dans l'approche par règles linguistiques mise en place par l'équipe, et suivant un cadre d'annotation déjà établi.

Les objectifs du stage incluent :

- Moissonnage d'articles scientifiques dans différentes desciplines et formats (HTML, XML, LaTeX, ...).
- Extraction des contenus textuels, nettoyage, pré-traitements et importantion dans une base de données MySQL.
- Segmentation en phrases et catégorisation des segments textuels.
- Expérimentation avec des approches à base de règles et/ou approches par apprentissage automatique, et évaluations.

L'équipe du projet étaint internationale, le travail se fera principalement en anglais.

Profil recherché

- Vous êtes étudiant de niveau Master (de préférence Master 2, Bac +5) avec des compétences en Informatique et en Traitement Automatique des Langues (TAL).
- Vous avez de solides compétences en Python/R, MySQL et des outils en TAL.
- Vous avez des compétences en extraction de données et fouilles textuelles.
- Vous êtes organisé, rigoureux, autonome et doté d'une bonne communication.
- Vous avez un bon niveau en anglais (au moins B2).
- Les connaissances en HTML, XML et LaTeX sont un plus.
- Les connaissances en apprentissage automatique/profond sont un plus.

Détails

- **Durée** : 6 mois.
- **Date de début** : Début 2023, négociable.
- **Localisation :** Télétravail possible avec des déplacements ponctuels à prévoir au CRIT (rue Mégevand 25000 Besançon).

Organisation

Le laboratoire C.R.I.T

Le CRIT (Centre de Recherches Interdisciplinaires et Transculturelles) de l'Université de Franche-Comté intègre une équipe en Traitement Automatique des Langues (TAL) qui développe des méthodes linguistiques pour le traitement de textes écrits. Le stage sera encadré par Dr Iana ATANASSOVA, maître de conférences HDR au CRIT et responsable du projet ANR InSciM.

Projet ANR InSciM

Le projet ANR InSciM « Modélisation de l'incertitude en science » (https://anr.fr/Projet-ANR-21-CE38-0003, https://project-inscim.github.io/) a pour objectif d'étudier l'incertitude exprimée dans les articles scientifiques de différentes disciplines. Il s'agit d'utiliser les méthodes du TAL pour créer une ontologie disciplinaire de l'incertitude et annoter des corpus.

Comment candidater

Envoyez votre CV et lettre de motivation (avec noms et coordonnées de 2 référents) aux adresses suivantes :

- panggih_kusuma.ningrum@univ-fcomte.fr
- iana.atanassova@univ-fcomte.fr

Internship: Processing of Scientific Texts for the Categorisation of Uncertainty Expressions (6 months)

Several internships are proposed by the C.R.I.T laboratory at the University of Franche-Comté in Besançon, France, financed by the ANR InSciM project, in the field of Natural Language Processing (NLP).

Job Description

The objective of the internship is to process corpora of scientific articles in English from different sources, in order to identify and categorise the textual segments carrying uncertainty. The text processing algorithms that will be developed will be integrated at part of the linguistic approach implemented by the team, and will follow an established annotation framework.

The objectives of the internship include :

- Harvesting scientific articles in different disciplines and formats (HTML, XML, LaTeX, ...).
- Textual information extraction, cleaning, pre-processing and importing into a MySQL database.
- Sentence segmentation and categorisation of text segments.
- Experimentation with rule-based and/or machine learning approaches, and evaluation.

As the project team is international, the work will be done mainly in English.

Required Skills and Expertise

- You are a Master student (preferably Master 2, Bac +5) with skills in Computer Science and Natural Language Processing (NLP).
- You have knowledge in programming langages including Python/R, MySQL and NLP tools.
- You are familiar with data harvesting, data extraction and text mining.
- You are organised, autonomous and have good communication skills.
- You are fluent in English (at least B2 level).
- Knowledge of HTML, XML and LaTeX is a plus.
- Knowledge of machine learning/deep learning is a plus.

Details

- **Intership duration** : 6 months.

- **Start** : Beginning of 2023, negociable.
- **Location :** CRIT (rue Mégevand 25000 Besançon). Remote working is possible with several on-site meetings at CRIT.

Context

The C.R.I.T laboratory

The CRIT (Centre de Recherches Interdisciplinaires et Transculturelles) of the University of Franche-Comté includes a team in Natural Language Processing (NLP) which develops linguistic methods for the processing of written texts. The internship will be supervised by Dr Iana ATANASSOVA, assistant professor at the CRIT and PI of the ANR InSciM project.

ANR InSciM Project

The ANR InSciM project "Modelling Uncertainty in Science" (https://anr.fr/Projet-ANR-21-CE38-0003, https://project-inscim.github.io/) aims to study the uncertainty expressed in scientific articles from different disciplines. The main purpose is to use NLP methods to create a disciplinary ontology of uncertainty and annotate scientific corpora.

Comment candidater

You can apply by sending your CV and motivation letter (including names and coordinates of two referees) to:

- panggih_kusuma.ningrum@univ-fcomte.fr
- iana.atanassova@univ-fcomte.fr