Bachelor Thesis

Text classification through NLP (in Python)

BACKGROUND:

For the purpose of political counselling, technological assessment, and more, the framework of PEST(EL or STEEPLE) analysis provides an effective approach to work out promotive and prohibitive influencing factors with respect to relevant categories; e.g. Political, Economic, Social, Technological – hence the name PEST.
As manual information retrieval and simultaneous text annotation by experts is time consuming and with respect to the growing number of relevant references almost always incomplete, applied natural language processing (NLP) promises a remedy.
The overall goal is to create an artificial intelligence (AI) that automatically annotates new texts according to one of the category classes and whether it is promotive or prohibitive in that respect such that scientists can quickly harvest textual references for their research. As a first step towards an AI, this thesis project studies whether category classification is feasible and, if so, should compose a list of distinctive keywords.

YOUR TASKS:

- Get familiar with the source texts and data preprocessing in terms of lemmatisation, normalisation, segmentation, tokenisation, etc.
- Develop and evaluate a classifier for the STEEPLE/X categories, where X stands for relevance in neither STEEPLE category
- Critically assess (i) potential model biases (e.g., due to data selection) and (ii) model performance in terms of quality on new text sources together with experts from the field of bioeconomy

YOU HAVE:

- Joy and curiosity in developing NLP models for a noble cause as well as a self-reliant, conscientious, and timely way of working
- Passed basics lectures on computer science and digital humanities
- Strong communication skills in German and English, especially since this project is a joint venture of the work groups Resource Mobilisation and DBFZ's DataLab involving scientists from different disciplines

WE OFFER:

- A good introduction to the topic as well as competent and motivated support in the processing of the tasks
- A family-friendly, modern working environment in a collegial working atmosphere
- Good public transport connections

BEGINNING:

2023-03-01 (or by arrangement)

DURATION:

23 weeks
(prolongation DBFZ-wise possible)

PROCESSING LOCATION:

Deutsches Biomasseforschungszentrum gemeinnützige GmbH
Torgauer Straße 116, D-04347 Leipzig

CONTACT:

Dr. rer. nat. Marco Selig
DataLab work group leader
Phone: +49-341-2434-854

APPLICATION DOCUMENTS:

Please submit your compelling application (only in a single attachment, preferably as PDF, max. 5 MB)
e-Mail: bewerbung@dbfz.de

For an encrypted transmission of your application you can use the upload form Cryptshare.
www.dbfz.de/stellen