Bachelorarbeit

Textklassifizierung mittles NLP (in Python)

FRAGESTELLUNG:

Zum Zweck der Politikberatung, Technologiebewertung und Weiterem bietet eine sogenannte PEST(EL oder STEEPLE)-Analyse einen effektiven Ansatz, um förderliche und hinderliche Einflussfaktoren in Bezug auf relevante Kategorien herauszuarbeiten. Das Akronym PEST leitet sich hierbei von den englischsprachigen Kategorien ab, d. h. Political, Economic, Social, Technological.
Da die manuelle Informationsbeschaffung und gleichzeitige Textannotation durch Experten zeitaufwendig und angesichts der wachsenden Zahl relevanter Referenzen fast immer unvollständig ist, verspricht die Anwendung von Methoden der Computerlinguistik Abhilfe.
Das übergeordnete Ziel ist es, eine künstliche Intelligenz (KI) zu schaffen, welche neue Texte automatisch nach einer der Kategorien zuordnet und feststellt, ob sie in diesbezüglich förderlich oder hinderlich ist, so dass Forschende schneller Textquellen für ihre Forschung sichten können. Als erster Schritt auf dem Weg zu einer KI wird in dieser Arbeit untersucht, ob eine Klassifizierung nach Kategorien möglich ist und, falls ja, soll ebenfalls eine Liste von charakteristischen Schlüsselwörtern erstellt werden.

IHRE TÄTIGKEITSSCHWERPUNKTE:

- Vertraut machen mit den Ausgangstexten und der Datenvorverarbeitung in Bezug auf Normal-, Token-, Lemmatisierung, Segmentierung, etc.
- Entwicklung und Evaluierung eines Klassifikators für die STEEPLE/X-Kategorie, wobei X keine Relevanz in irgendeiner STEEPLE Kategorie anzeigt
- Kritische Bewertung (i) potenzieller Modellverzerrungen (z. B. aufgrund der Datenvorauswahl) und (ii) der Leistungsfähigkeit des Modells hinsichtlich der Klassifizierungsqualität bei neuen Textquellen

WIR ERWARTEN:

- Freude und Neugierde an der Entwicklung von Modellen der Computerlinguistik sowie eine selbstständige, gewissenhafte und termingerechte Arbeitsweise
- Bestandene Grundlagenmodule in Informatik und Digital Humanities
- Gute Kommunikationsfähigkeiten in Deutsch und Englisch, insb. da dies Projekt ein interdisziplinäres Gemeinschaftsprojekt zweier Arbeitsgruppen des DBFZ's ist

WIR BIETEN:

- Einen guten fachlichen Einstieg in die Thematik sowie eine kompetente und motivierte Unterstützung bei der Bearbeitung der Aufgabenstellung
- Ein familienbewusstes, modernes Arbeitsumfeld in einem kollegialen Arbeitsklima
- Gute Anbindung an öffentliche Verkehrsmittel

MÖGLICHER BEGINN:

2023-03-01
(bzw. nach Vereinbarung)

DAUER:

23 Wochen
(Verlängerung DBFZ-seitig möglich)

BEARBEITUNGSORT:

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

ANSPRECHPARTNER:

Dr. rer. nat. Marco Selig
Arbeitsgruppenleiter DataLab
Telefon: +49-341-2434-854

BEWERBUNGSUNTERLAGEN:

Bitte bewerben Sie sich mit Ihrer aussagefähigen Bewerbung inkl. Motivationsschreiben und aktueller Immatrikulationsbescheinigung (nur 1 Anhang möglich, vorzugsweise als PDF, max. 5 MB).
E-Mail: bewerbung@dbfz.de

Für eine verschlüsselte Übermittlung Ihrer Bewerbung können Sie das Uploadformular Cryptshare nutzen.
www.dbfz.de/stellen