Natural language processing: One Python NER evaluation library to rule them all

Hintergrund:

Zur zuverlässigen, reproduzierbaren und vergleichbaren Evaluierung von Named-Entity-Recognition(NER)-Modellen ist es notwendig, gewöhnliche Scores wie Precision / Recall und F1 auf gleiche Art und Weise zu bestimmen. Bisher gibt es einige, unterschiedliche Verfahren, deren Precision/Recall/F1-Werte sich dementsprechend unterscheiden. Die Auswertung auf allen Verfahren ist recht aufwendig, da jedes Verfahren die Daten nur jeweils in einem bestimmten Eingangsformat verarbeiten kann und dieses Format zwischen den Ansätzen unterschiedlich ist, sowie sich die Programmiersprachen der Tools jeweils unterscheiden.

Aufgabenstellung:

Ziel der Arbeit ist die Implementierung einer Python-Wrappers, der mehrere dieser Evaluierungs-Verfahren integriert und dabei transparent die Eingabedaten vor der Verarbeitung umkonvertiert, mittels des gewählten Verfahrens auswertet, und die Ergebnisse in einer homogenen Darstellung umwandelt und dem Aufrufer übergibt. Der zweite Schritt besteht darin, im Falle von ein oder mehrere dieser Verfahren mittels test-driven-development Test-Cases (Eingabedaten und erwartete Rückgabedaten) zu definieren und ein oder mehrere Verfahren direkt in Python neu zu implementieren.

Anforderungen:

- Verständnis für fremden Code in Python und ggf. Java
- Gute, eigene Python-Kentnisse inkl. NumPy und ggf. Scikit-learn (optional)

Betreuung und Dauer:

Die Arbeit wird auf eine Dauer von 3 (Bachelorarbeit) Monaten ausgelegt. (Auch ggf. Forschungsmodul möglich) Während der gesamten Arbeit steht Ihnen ein Betreuer zur Seite.

Falls Sie Interesse oder Fragen an dieser Abschlussarbeit haben, freuen wir uns über Ihre Bewerbung. johann.frei@informatik.uni-augsburg.de

Weiterführende Links:

- Über Token Classification / Sequence Tagging / Named-Entity-Recognition: https://huggingface.co/docs/transformers/en/tasks/token_classification bzw. https://en.wikipedia.org/wiki/Named-entity_recognition

- BratEval: https://github.com/READ-BioMed/brateval
- ConllEval: https://github.com/sighsmile/conlleval
- SeqEval: https://github.com/chakki-works/seqeval