

NLP Assignment 3

Dieses Assignment wird bewertet !

Die Zusammenarbeit unter Studierenden ist erwünscht, so lange sich diese auf die Diskussion von Konzepten oder Problemen mit python konzentrieren. Das Kopieren von Code ist nicht erlaubt, in diesem Falle werden alle Lösungen der beteiligten Parteien mit 0 Punkten bewertet; dazu werden alle Lösungen (manuell und automatisiert mit Plagiat-Checkern) auf Gruppenarbeit und Kopien untersucht.

Für dieses Assignment installieren Sie Keras: <https://keras.io/#installation>

Entwickeln Sie mit TensorFlow/Keras einen POS-Tagger für Deutsch mit dem Ziel, eine **möglichst grosse accuracy** zu erreichen:

- Reduzieren Sie in einem ersten Schritt das Stuttgart-Tübingen-Tagset (STTS) auf das Universal-Tagset. Dokumentieren Sie, wie Sie das detaillierte STTS auf die 12 Wortarten des Universal-Tagsets abbilden.
- Entwickeln Sie mit TensorFlow/Keras einen POS-Tagger für das Universal-Tagset mit dem Ziel, eine möglichst hohe *accuracy* zu erzielen..
- Verwenden Sie die Datei POS_German_train.txt um Ihren Tagger zu trainieren. Um eine gute Generalisierung zu erreichen, könnte es helfen, dieses Dataset während der Entwicklungsphase nochmals in ein Trainings- und ein Dev-Test Set aufzuteilen um optimale Hyperparameter zu finden.
- Geben Sie die erzielte *accuracy* an für das kleine Test-Set POS_German_minittest.txt; dieses besteht aus Sätzen, die in POS_German_train.txt nicht vorkommen.

Hinweise:

- Das Stuttgart-Tübingen-Tagset (STTS) ist im Dokument TIGER_scheme-syntax.pdf auf Seite 121 ff erläutert.
- Für die Notengebung wird nicht mit POS_German_minittest.txt sondern mit einer umfangreicheren Test-Datei getestet.

Abgabe : In Form eines Jupyter Notebooks
per Email an ivo.nussbaumer@fhnw.ch bis 12. Mai 2019