NLP Assignment 2

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.

Entwickeln Sie 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 einen mehrstufigen (*stacked*) POS-Tagger für das Universal-Tagset mit dem Ziel, eine möglichst hohe *accuracy* zu erzielen. Denken Sie auch daran, dass vielleicht ein HMM-Tagging helfen kann, die *accuracy* zu steigern.
- 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.
- Stellen Sie die Qualität Ihres Taggers mit einer *confusion matrix* dar und geben Sie die erzielte *accuracy* an. Verwenden Sie dazu die Datei POS_German_minitest.txt, die aus Sätzen besteht, die in POS_German_train.txt nicht vorkommen. Die *confusion matrix* hilft Ihnen auch, den Tagger zu optimieren resp. weiter zu *stacken*.

Hinweise:

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

Abgabe : In Form eines `Jupyter Notebooks`
per Email an dominik.frefel@fhnw.ch bis 22. April 2018