MT FS19
Iris Suarez Malo – 15-713-258

Übung 4 - RNNs

Bericht zur Übung

Datenset:

Als Datenset habe ich mich für eines aus der, zur Verfügung gestellten, Excel Datei entschieden. Dabei handelt es sich um das «German Political Speeches Corpus and Visualization», welches unter folgendem Link verfügbar ist:
http://adrien.barbaresi.eu/corpora/speeches/ .
Für dieses Datenset habe ich mich entschieden, da es sich dabei um ein reines Text Corpus handelt. Es ist eine Sammlung von politischen Reden auf Deutsch. Zudem enthält es ungefähr 122000 Sätze, was ich für diese Übung als eine gute Grösse erachte.

Preprocessing:

Da das Datenset als xml-Datei vorliegt, habe ich die Daten in eine Textdatei umgewandelt und in Sätze und und Worte gesplitet und dann pro Zeile einen Satz abgespeichert.

Hyperparameter:

Als erstes habe ich die Vokabulargrösse von 10000 auf 15000 vergrössert. Diesen Entscheid traff ich, da mehr als 10000 verschiedene Wörter im Datenset vorkommen.
Als zweites habe ich die Batchgrösse auf 32 verringert da die Resultate mit 64 nicht gut waren.

Bereits durch diese zwei kleinen Änderungen hat sich das Resultat sehr verbessert.
Dabei sind bei mir keine Probleme aufgetretten.

Resultate:

Unverändert:

Verändert: