RNNs

Für diese Übung wurden aus persönlicher Motivation die auf Project Gutenberg frei verfügbaren Bücher von **Friedrich Wilhelm Nietzsche** in deutscher Sprache ausgewählt.

Im **ersten Trainingsschritt** wurde das Preprocessing halbautomatisch (ohne Skripte) ausgeführt. Das Datenset bestand aus 3 Büchern und war nach dem Preprocessing ca. 1,1 MB gross. Hiermit wurde für das erste Modell ein **Scoringwert** von **231,90** erzielt.

Für den **zweiten Trainingsschritt** wurde ausprobiert, wie sich der Scoringwert gegenüber dem ersten Modell verändert, wenn die **Grösse** und **Qualität** des **Datensets erhöht** werden.
Dafür wurden bei diesem Schritt alle verfügbaren Bücher (insgesamt 6 Stück) von Nietzsche verwendet (http://www.gutenberg.org/ebooks/7202, http://www.gutenberg.org/ebooks/7203, http://www.gutenberg.org/ebooks/7204, http://www.gutenberg.org/ebooks/7205, http://www.gutenberg.org/ebooks/7206, http://www.gutenberg.org/ebooks/7207).
Das Preprocessing wurde nun um ein Python-Skript ergänzt. Damit wurde das zeilenweise Abtrennen von Sätzen verbessert. Dieses Skript beachtet auch die in Nietzsches Büchern verwendeten Abkürzungen mit Punkt: «Dr.», «S.», «u.s.w.», «f.», «z.B.», «d.h.».

Das Preprocessing bestand des Weiteren aus halbautomatischen Schritten. So wurden mit dem Regex-Befehl

```
^.{0,3}((\r?\n)|$)
```

bzw.

```
^.{0,4}((\r?\n)|$)
```

diejenigen Zeilen entfernt, die nur aus Kapitelnummern bestanden (z.B. «6.» oder «112.»)

Durch den Regex-Befehl

```
^.{0,10}((\r?\n)|$)
```

konnte der gesamte Text stichprobenartig auf Fehler in Zeilenumbrüchen überprüft werden. Einige der oben genannten Abkürzungen konnten so entdeckt werden. Da aber auch einige Satzzeichen von Zeilenumbrüchen betroffen waren, wurde dies manuell überprüft und korrigiert.

Das Datenset hat somit eine Dateigrösse von ca. 2,2 MB erreicht und enthält 15084 Zeilen.

Das verbesserte Preprocessing und die vergrösserte Datenmenge haben den Scoringwert deutlich gesteigert. Als **«Perplexity»** wird nun für das **adaptierte Modell** ein Wert von **191,13** ausgegeben.

Da der zweite Trainingsschritt aus dem direkten Vergleich des Scoringwertes hinsichtlich der veränderten Datenmenge und der verbesserten Datenqualität bestand, wurden keine Hyperparameter verändert, sondern die Standardeinstellungen beibehalten (Standard-Vokabulargrösse von 10000). Somit kann für diesen Anwendungsfall gefolgert werden, dass ein grösseres und qualitativ hochwertigeres Datenset die Resultate merklich verbessern.

In weiteren Trainingsschritten – die aus Zeitgründen nicht weiter ausgeführt wurden – , wie z.B. der Anpassung der Vokabulargrösse, wird der Scoringwert sicherlich noch weiter optimiert werden können.