**MT Übung 4**

**Thema: RNNs**

Wir haben uns für dieses Datenset (Shakespeare) entschieden, weil Shakespeares Werke einerseits „altes“ Englisch, sowie andererseits „gleichgebliebenes“ Englisch beinhalten. Ausserdem beinhalten Shakespeares Werke so viele Synonyme, dass manche sogar behaupten, dass diese Werke nicht nur von einer Person stammen konnten. Vokabulargrösse: 10000. Wir haben ein Datenset ausgesucht, welches mehr als die Mindestvokabulargrösse (über 1MB Text) beinhaltet, jedoch nicht allzu viel, damit der Server nicht allzu lange für seine Berechnungen benötigt.

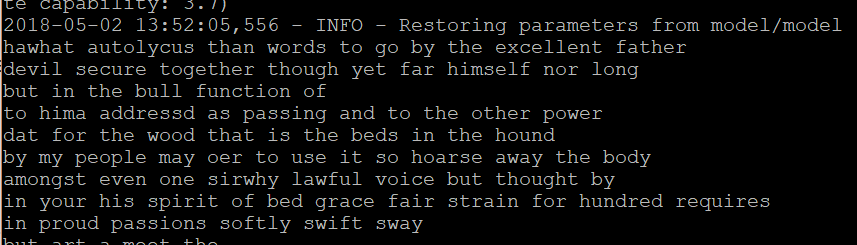
Anschliessend haben wir das Datenset durchmischt und im Anschluss in 1/10 als Dev-Set und der Rest als Trainings-Set unterteilt. Einige Datensets hätten uns auch sehr interessiert, wie beispielsweise Kundenmeinungen über Unfälle und medizinische Unterstützung, welche jedoch nicht zugänglich waren. Einige Links gingen nicht und einige Datensets brauchten viele Pre-Processing-Schritte. Da Romanesco keine Pre-Processing Schritte durchführt haben wir als Pre-Processing-Schritt auf alle Satzzeichen verzichtet, daher ist die Variante ohne Pre-Processing etwas eleganter, da sämtliche Satzzeichen noch da stehen.

Beim Code-Verändern war es schwierig herauszufinden, wo etwas fehlt, was besser, bzw. im Kontext passender sein würde. Wir hatten etwas Mühe gegeben, viele verschiedene Hyperparameter z.B –e, -b zu benuzen, um die tiefsten Perplexität zu erreichen. Das Training war zeitaufwendig. Aber man hatte immer einen guten Überblick mit den GPU’s, wo noch Platz auf dem Server ist.

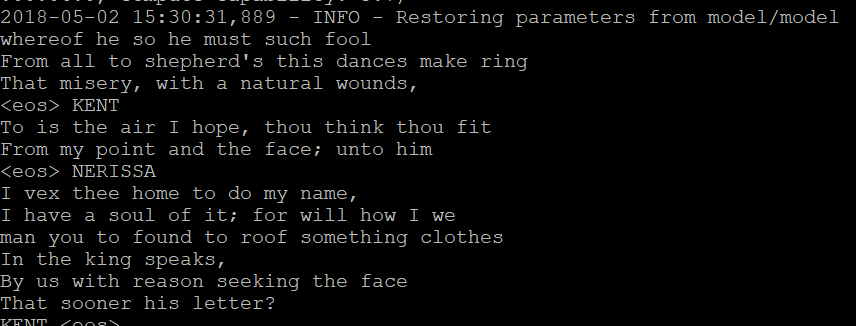
Als Perplexität, also wie überrascht unser System auf dem Dev-Set war, haben wir folgende Perplexity-Werte erhalten:

epoch(4):406 epoch(3)1:409 epoch(3)2:261 epoch(3)3:221

* without pre-processing:perplexity:136
* with pre-processing(batch size 40):perplexity:221



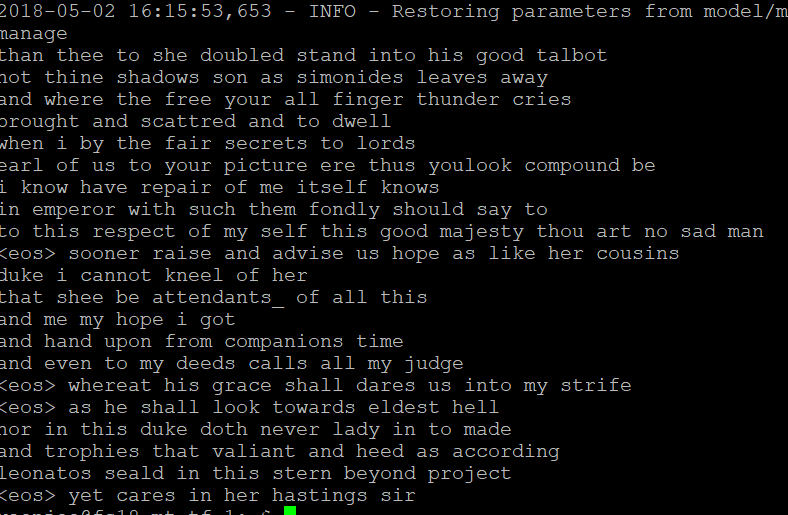
Mit Pre-Processing war die Perplexity höher, da wir die Satzzeichen weggelassen haben. Normalerweise wäre die Perplexity jedoch tiefer, wenn anschliessend noch ein Post-Processing stattfinden würde. Bei Shakespeare liegt die Perplexity ziemlich hoch, da es ein anspruchsvoller Text (und ausserdem nicht aus dem heutigen Jahrhundert stammt) ist.



Beim Sampling haben wir neuen Text generieren lassen. Die neuen Sätze die herausgekommen sind, sehen lustig aus. Es gibt Wörter die von der Reihenfolge nacheinander passen und andere wiederum die gar keinen Sinn machen. Uns sind die vielen „you“ also:

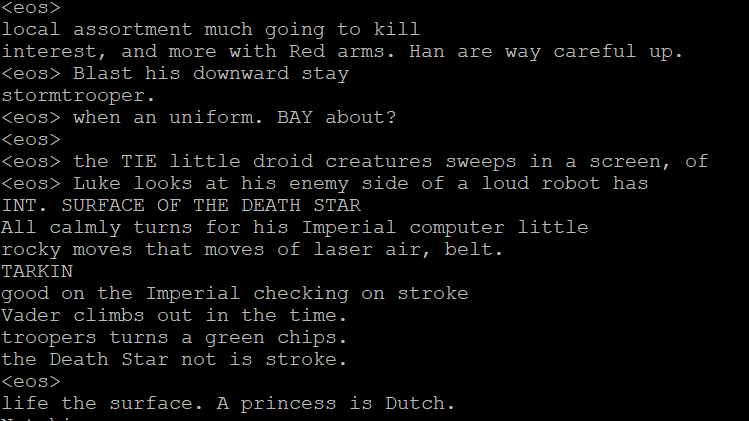
„thou“ oder „thee“ („you“) und „thy“ oder „thine“ („your“) und „ye“ (Plural „you“)

aufgefallen. Diese wurden beim Text generieren lassen nicht richtig verwendet: „…in your his spirit…“ oder „…thou think thou fit…“ sowie „…man you to found…“. Allgemein wurden die Personalpronomen nicht richtig angewendet: „…how I we man you to found…“ ist ein katastrophaler Satz. Beim Pre-Processing (im Gegensatz zu ohne Pre-Processing) wurden einige Wörter zusammengeschweisst : „hawhat“ statt „ha what“, “„hima“ statt „him a“, „sirwhy“ statt „sir why“. Dies liegt wahrscheinlich daran, dass die Satzzeichen weggelassen wurden und somit jeweils das letzte Wort und das erste Wort eines Satzes zusammengesetzt wurden. Da unser Text auf Englisch und nicht auf Deutsch geschrieben ist, ist es weniger schlimm (somit werden das „I“, Eigennamen und Nationalitäten ausnahmsweise klein geschrieben), dass alles klein geschrieben (lowercase) ist.



Zum Spaß haben wir uns für das zweite Datenset (Starwars) entschieden, um zu beobachten, ob ein neuer interessanter Text im Starwars-Stil nach dem Training generiert werden kann. Das Training war nicht so zeitaufwendig. Beim diesem Datenset funktionierte es ziemlich schnell, denn es liegt an der Datenmenge. Beim Training haben wir die Batchsize als 1 und Epoch als 10 fixiert. Als Perplexität, also wie überrascht unser System auf dem Dev-Set war, haben wir folgende Perplexity-Werte erhalten:

* without pre-processing:perplexity:27



Beim Sampling haben wir neuen Text generieren lassen. Die folgenden Wörter sind bei uns aufgefallen wie erwartet:

„uniform“, „princess“, „Imperial“, „Han“, „Vader“, „Death“, „robot“, „Luke“, „computer“

Kreative Sätze sind bei uns auch besonders aufgefallen. Es überrachste uns, dass nicht nur die Eigennamen richtig erkannt und korrekt generiert wurden, sondern auch manche Sätze in guter Struktur formuliert wurden wie die folgenden Sätze:

„A princess is Dutch“, „Luke looks at his enemy side of a loud robot“, „Vader climbs out in the time“