Übung 4

DATENSET

Wir haben uns für das Datenset aus PCL 2 entschieden, welche weibliche und männliche Babynamen enthält. Zusätzlich das Babynamenset von http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/ . Diese beiden sets fügen wir zusammen und löschen Duplikate, um ein neues, grösseres Set zu erhalten. Da wir auf Zeichenebene trainieren, ist es nicht schlimm, dass es nicht ganz 1 MB gross ist, denn Buchstaben hat es genug.

PREPROCESSING

Die Files sind separiert nach männlichen Namen und weiblichen, was heisst, dass wir diese noch zusammenführen müssen. Hierfür müssen wir zuerst allen ein Namen ein Label geben (m/f) und anschliessend die beiden Files zusammenlegen und gründlich mischen. Nachdem die Files zusammengeführt sind, können wir sie aufteilen in ein Trainings-, Test- und Developmentset. Die Aufteilung der Menge sieht folgendermassen aus: Training 80%, Test 10% und Dev 10%. Vor jedem Trainingsdurchlauf teilen wir das Dev- und Trainset neu ein um ein Overfitting zu verhindern. Das Testset wird aber nicht verändert. Ausserdem müssen wir noch beachten, dass das Programm jeweils Name und Label separat einliest (z.B. ein .csv-File, da wird die Separierung einfacher).

TRAIINING

Wir trainieren auf Zeicheneben, da unser Modell sonst keine neuen Namen generieren könnte, was sehr schade wäre. Das heisst, wir müssen den Code anpassen, dass nicht nach Wörtern tokenisiert wird, sondern nach Zeichen (Ein Satz wird zu einem Wort und ein Wort wird zu einem Buchstaben).

- Hyperparamter:
  - epochs: 27 → so wird ein bisschen mehr trainiert, aber noch kein Overfitting
  - batch_size: 10 → das sollte laut stackexchange eine gute learningrate geben.
  - NUM_STEPS: 800 → gibt etwa 10 Namen pro step
  - HIDDEN_SIZE: 1000 →1500 scheint ein bisschen viel

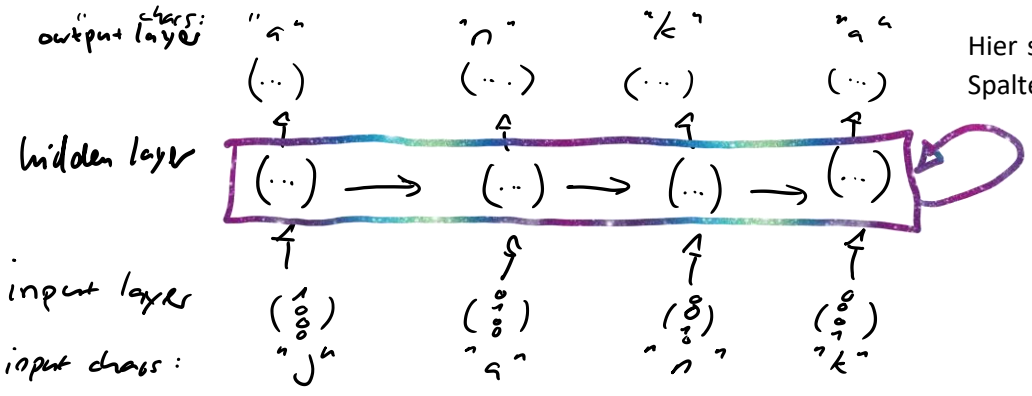
Probleme: Wir hatten keine Zeit um diese Übung komplett zu lösen ☹

CODE-VERÄNDERUNGEN

- Tokenisierung nach Buchstaben und nicht nach Wörtern
- Shuffle von Trainingsset und Devset → Overfitting vermeiden
- Hyperparameter anpassen

PERPLEXITÄT

Die Perplexität wird ziemlich sicher höher sein als beim Trainingset, da unser Modell kein wirklich gutes ist.



Hier sollte jeweils die richtige Spalte den höchsten Wert haben.