Übung 4 – Bericht – Martina Forster – mafors

Als Datenset habe ich Movie-Reviews von https://www.kaggle.com/nltkdata/movie-review/version/3 genommen, weil die Textdaten schon ‚lowercase' und ‚white-space tokenisiert' sind und es eine genug grosse Datenmenge war. Mein Preprocessing-Skript liest einfach alle Sätze aus der betreffenden Text-Kolonne im .csv (es hatte sowieso nur 1 Satz pro Zeile), macht Shuffling und generiert schliesslich ‚train.txt' und ‚dev.txt' daraus. Beim ersten Training habe ich alle Default-Parameter von Romanesco verwendet (Epochenzahl war 10, Batchsize 20 und Vocabsize 10'000), beim zweiten Training habe ich die (Default-) Epochenzahl auf 30 erhöht, um zu sehen, ob eine so simple Veränderung einen grossen Unterschied macht. Ich habe extra nur diese eine Veränderung gemacht, weil ich den reinen Effekt der Epochenzahl sehen wollte. Die anderen Parameter habe ich gleich gelassen wie beim ersten Training. Die Idee war, herauszufinden, ob ein dreimal so langes Training einen grossen Unterschied in der Perplexität macht. Probleme dabei: Ich hatte zuerst etwas Mühe, herauszufinden, welches eigentlich das „main"-Skript ist, bzw. welches danach die anderen Skripts aufruft. Es hat dann aber ganz gut geklappt, ich habe mir einfach alle Files genau angeschaut und mir einen Überblick verschafft. Die Perplexität beim ersten Training (unverändert) war 104.11, während ich beim zweiten Training (adaptiert) eine Perplexität von 110.81 bekommen habe.