MT ÜBUNG 4

PETRA WITTWER - 11-914-488

Als Fragestellung für diese Aufgabe habe ich mir überlegt, ob ein RNN in der Lage ist aus Kochrezepten zu lernen und somit im Sampling sinnvolle neue Rezepte generieren kann. Bei der Suche nach Daten hatte ich sehr viel Glück und stiess auf ein Datenset auf Kaggle, dass Rezepte als JSON zusammen mit einer Kategorisierung enthält[1]. Das Zipfile mit den Daten befindet sich im Repositoty. Um diese JSON Daten in Romanesco zu verwenden, habe ich diese mit einem Python-Skript vorverarbeitet. Im aus full_format_recipes.json erzeugten File recipes.txt, entspricht ein Rezept einer Zeile. In einer Zeile befindet sich jeweils der Titel des Rezepts, dann die Zutaten, Anweisungen und schliesslich die Kategorien. Dabei haben Anweisungen und Zutaten jeweils die Überschrift «Directions» bzw. «Ingredients» Das so erzeugte File enthält ca. 20'000 Rezepte und ist ca. 32MB gross. Meine Hoffnung war es, dass das RNN diese Strukturierung aus Titel -> Zutaten -> Anweisungen -> Kategorien lernt. Ich habe mir überlegt, dass das Hinzufügen von mehr Layern dazu führen könnte, dass die einzelnen Layer auf entweder Struktur oder zugehörige Wörter fokussieren könnten. Dies hat allerdings im Sampling nie so richtig funktioniert. Was eingermassen gut funktioniert, ist das Erzeugen der Überschriften, gefolgt von (syntaktisch) sinnvollem Inhalt. Auf “Ingredients” folgen meistens tatsächlich Zutaten zusammen mit Mengenangaben. Diese Mengenangaben haben überraschend gut funktioniert: Butter kommt in Samples meist mit der Mengenangabe «Stick» vor, Salz beispielsweise aber meistens mit «Teaspoons».

Bei den Hyperparametern habe ich mich für 3 Epochen und eine Vokabulargrösse von 6000 entschieden. In meinen früheren Versuchen habe ich festgestellt, dass der Loss kaum abnimmt nach den ersten 3 Epochen. Die Vokabulargrösse habe ich recht stark reduziert da in der Domäne insgesamt recht wenige Wörter vorkommen. Da ich ja unter anderem auch darauf fokussiert war möglichst gute Samples zu bekommen, habe ich überlegt, dass ein kleineres Vokabular auch zu sinnvolleren Kombinationen führen sollte.

Beim Scoring auf dem Dev-Set habe ich eine Perplexität von 19.66 erhalten.

[1] https://www.kaggle.com/hugodarwood/epirecipes/version/2