Übersicht

Ich habe zwei verschiedene Datensets ausgewählt, um zwei verschiedene Arten von Texten zu testen. Das erste Datenset ist eine Sammlung von knapp 58'000 Familiennamen aus der nlp Bibliothek und das zweite die gesammelten Werke von Edgar Allan Poe (Gedichte und Kurzgeschichten) aus dem Gutenberg Projekt. Ich habe erwartet, dass das Programm die Namen wesentlich besser verarbeiten kann als die literarischen Texte und so war es auch, wie sich anhand der Perplexität und des generierten Outputs feststellen lässt. Dazu später mehr.

Familiennamen-Datenset

Das Datenset ist etwa einen halben Megabyte gross. Ich habe es ausgewählt, da ich erwartete damit sehr gute Ergebnisse erzielen zu können. Namen sind relativ kurz und haben immer wiederkehrende Muster, wie etwa ähnliche Endungen. Es ist sehr simpel aufgebaut, auf jeder Zeile befindet sich ein Name. Damit erfüllt es die Bedingungen für romanesco, zusätzliche pre-processing war nicht nötig. Ohne etwas am Code zu ändern erreichte es eine **Perplexität von 1.01**, also nahezu perfekt. Das wirkte zunächst etwas zu gut um wahr zu sein, aber der Output besteht tatsächlich aus echten Namen (wie z.B Schmitz) und fast-echten Namen (wie z.B. Macdougal, korrekt wäre: McDougal). Meine Vermutung stellte sich also als richtig heraus.
Output befindet sich in names_sample.txt.

Edgar Allan Poe Datenset

Dieses Datenset ist knapp ein Megabyte gross. Ich habe es ausgewählt, da ich testen wollte wie das System mit hochkomplexen Texten umgeht, im Gegensatz zu den Namen, ich erwartete, dass es schlecht abschneiden würde. Dieses Datenset war etwas komplexer aufgebaut. Zwar befand es sich bereits als plain-file, aber befand sich noch in einem auf menschliches Lesen zugeschnittenes Format. Die Texte waren also nicht satzsegmentiert. Mit Hilfe von des Pre-processing (Script befindet sich im repository) habe ich den Text für romanesco aufbereitet. Wie erwartet war die erzielte **Perplexität von 151.82** wesentlich schlechter als bei den Namen. Der Output bildet häufig grammatikalisch nicht korrekte Sätze, immerhin ist die Sprache aber äusserst düster wie beim Autor und der Stil sogar noch mysteriöser und verworrener.
Output befindet sich in poe_sample.txt.