

Übung 9

Bigramm-Sprachmodell

Themen: *open*, *defaultdict(lambda: defaultdict(int))*, *sum*, *Overfitting*

Schreiben Sie eine Klasse `Sprachmodell`, deren `init`-Methode einen Text als Argument erhält und

- ein Start- und ein Ende-Symbol hinzufügt
- die Häufigkeiten aller Zeichen-Bigramme berechnet
- und aus den Häufigkeiten Wahrscheinlichkeiten schätzt nach der Formel:

$$p(b|a) = \frac{f_{ab}}{\sum_b f_{ab'}}$$

Schreiben Sie zusätzlich eine Methode `perplexity`, welche einen Text als Argument erhält, ebenfalls über die Zeilen iteriert und die Wahrscheinlichkeit jeder Zeile wie folgt berechnet:

$$p(\mathbf{t}) = \prod_{i=1}^{n+1} p(t_i|t_{i-1})$$

Dabei ist t_0 das Startsymbol und t_{n+1} das Endesymbol. Diese Wahrscheinlichkeit wird dann noch in eine Perplexität umgerechnet:

$$perplexity(\mathbf{t}) = -\frac{\log p(\mathbf{t})}{|\mathbf{t}|}$$

Das Hauptprogramm liest eine Datei ein, deren Name auf der Kommandozeile übergeben wurde, und initialisiert damit das Sprachmodell.

Dann wird der Benutzer aufgefordert, eine Textzeile einzugeben. Zu dem eingegebenen Text werden ein Start- und ein Ende-Symbol hinzugefügt. Dann wird die Wahrscheinlichkeit des Textes mit folgender Formel berechnet und ausgegeben:

Das Programm endet, wenn der Benutzer eine leere Zeile eingibt.

Frage: Welches Problem taucht beim Experimentieren mit dem System auf?