

Thema: Übungsblatt 5: Editierdistanz und datenbasiertes Training
Von: Timo Baumann, timo.baumann@oth-regensburg.de
Datum: 09.05.2022
Abgabe bis: 14.05.2022 24:00

Ü 7 abzugebende bewertete Übungen

bevor String lesen N-1 Startz

Bitte geben Sie die Lösungen zu den Übungsaufgaben über das Moodle ab. **Für dieses Aufgabenblatt geben Sie bitte in Gruppen von 2-3 Studierenden ab.**

Ü 7.1 Zeilenweises Vorgehen für N-Gramme

getline()
in /Wn

In der vergangenen Übung haben Sie das Trainingsmaterial als einen langen Text behandelt. In der Anwendung nutzen wir das Material zeilenweise und gehen davon aus, dass jede Zeile einen separaten Text enthält. (Dadurch erhalten wir viele unterschiedliche Möglichkeiten für Textanfänge und Textenden.)

Formen Sie Ihr bisheriges Programm (oder die Musterlösung) so um, dass

- jede Zeile separat gelesen, in N-Gramme zerlegt und dem Modell hinzugefügt wird
- stellen Sie den Symbolen jeder Zeile $N - 1$ spezielle Symbole (z.B. "<s>") voran und ein spezielles Symbol (z.B. "</s>") hinterher.
- Generieren Sie jetzt neue Zeilen, indem Sie als initiale Präfix $N - 1$ mal das Startsymbol angeben und beenden Sie die Generierung sobald das Endesymbol generiert wurde.

Ü 7.2 Evaluation von N-Grammen

Die Evaluation eines Sequenzmodells erfolgt wie beschrieben über die Berechnung der Entropie (oder alternativ: Perplexität) auf gegebenen Daten.¹

1. Erweitern Sie Ihr bisheriges Programm so, dass es für eine gegebene Historie und ein Folgesymbol die Wahrscheinlichkeit laut Modell ausgibt. (Sie können dies weiter direkt aus dem Text berechnen oder aber vorher für alle N-Gramme auszählen wie in der Musterlösung).

¹Jenseits der Ausführungen bei Jurafsky und Martin (Kapitel 3) ist möglicherweise auch <https://towardsdatascience.com/perplexity-in-language-models-87a196019a94> hilfreich.

2. Ihr Programm soll einen weiteren Text (bzw. Textzeile) entgegennehmen und jeweils den Logarithmus der Wahrscheinlichkeit jedes Symbols aufaddieren; am Ende teilen Sie durch die Länge des evaluierten Textes.

Experimentieren Sie: wie ist die Kreuzentropie Ihres Modells für die Trainingsdaten? Wie hoch ist sie, wenn Sie eine Art Daten (z.B. Merkel) in 90% Training und 10% Test aufteilen, auf Training das Modell trainieren und die Testdaten zur Evaluation nutzen?

Wie ändert sich die Entropie mit N?

Ü 7.3 Discounting für N-Gramme

Implementieren Sie Laplace-Discouting wie besprochen. Verbessert sich dadurch das Modell?

$\langle \text{SS} \rangle$ $\langle \text{S} \rangle$

Vorgehen:

Funktion die w_x berechnet

```
getline()
string texti
while (getline(input, texti)) {
    }
}
```

$$5 = \frac{2 \cdot 2^{x-1} \cdot n-1}{3 \cdot 3-1} = \frac{2 \cdot 2^{n-1}}{2} = 2^{n-1}$$

$$5 = \frac{3 \cdot 3^{x-1} \cdot n-1}{3 \cdot 3-1} = \frac{3 \cdot 3^{n-1}}{2} = 3^{n-1}$$

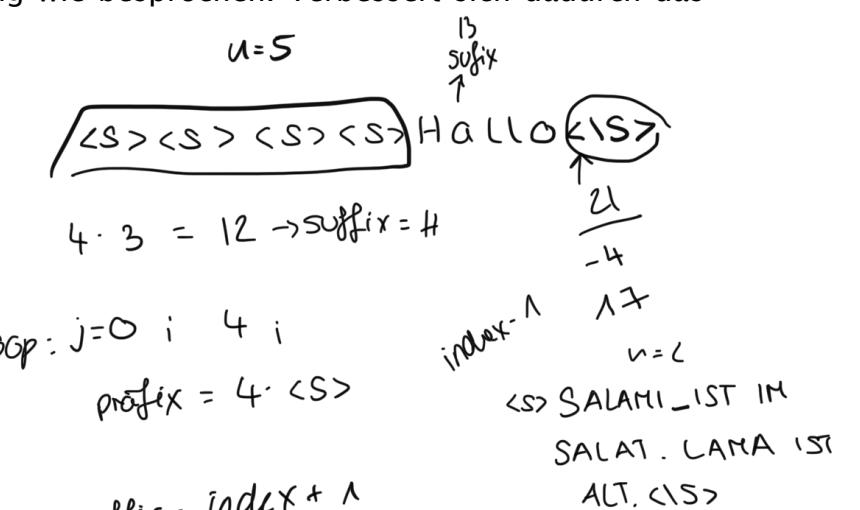
$$5 = \frac{3 \cdot 3^{x-1} \cdot n-1}{3 \cdot 3-1} = \frac{3 \cdot 3^{n-1}}{2} = 3^{n-1}$$

x=1

loop: $j=0 \quad i=4 \quad i$

prefix = $4 \cdot \langle \text{S} \rangle$

$$4 \cdot 3 = 12 \rightarrow \text{suffix} = 4$$



x=2:

$$\text{suffix} = \text{index} + 1$$

$$\text{prefix} = 3 \cdot \langle \text{S} \rangle + \text{index}$$

$$(3^1, 3^2, 3^3) = 2 \cdot 3^2 \cdot 3^1$$

3. <S>

le

$\langle \text{S} \rangle \langle \text{S} \rangle$