Woche 1: Grundlagen und Statistische Sprachmodelle

Muster in der Sprache entdecken
Vorbereitungsübung (Keine Programmierung erforderlich)

NLP-Kurs 2025 - Studentenversion

Zeit: 30-40 Minuten
Lernziel: Die Grundkonzepte statistischer Sprachmodelle durch praktische Entdeckung verstehen.

Teil 1: Einführung und Motivation (10 Minuten)

Wortvorhersage

Aufgabe 1: Vervollständigen Sie diese Sätze

a) Die Katze saß auf der _____

b) Er trinkt seinen Kaffee immer mit _____

c) Nach dem Regen scheint wieder die _____

d) Vielen Dank für Ihre _____

e) Es war einmal vor langer _____

Aufgabe 2: Reflexion
Warum konnten Sie diese Sätze so leicht vervollständigen? Notieren Sie drei Gründe:

1. _____
2. _____
3. _____

Entdeckungsmoment

Sie haben gerade das Grundprinzip von Sprachmodellen demonstriert: **Vorhersage basierend auf Mustern**. Ihr Gehirn hat tausende ähnliche Sätze gesehen und kann daher das wahrscheinlichste nächste Wort vorhersagen.

Alltägliche Anwendungen

Aufgabe 3: Wo begegnen Ihnen Sprachmodelle täglich?
Kreuzen Sie alle zutreffenden Anwendungen an und ergänzen Sie jeweils ein konkretes Beispiel:

- ☐ Smartphone-Tastatur: _____
- ☐ E-Mail-Programme: _____
- ☐ Suchmaschinen: _____
- ☐ Übersetzungsdienste: _____
- ☐ Sprachassistenten: _____
- ☐ Andere: _____

Schätzung: Wie viele Wörter sparen Sie täglich durch Autokorrektur/Vervollständigung? _____ Wörter

Teil 2: Das Zählprinzip (10 Minuten)

Muster zählen

Aufgabe 4: Analysieren Sie diesen kurzen Text

> Der Hund läuft. Der Hund bellt. Die Katze läuft. Die Katze schläft. Der Hund schläft.

Zählen Sie die Worthäufigkeiten und füllen Sie die Tabelle aus:

Wortpaar	**Häufigkeit**	**Wahrscheinlichkeit**
"Der Hund"	_____	_____
"Die Katze"	_____	_____
"Hund läuft"	_____	_____
"Hund bellt"	_____	_____
"Hund schläft"	_____	_____
"Katze läuft"	_____	_____
"Katze schläft"	_____	_____

Aufgabe 5: Berechnen Sie bedingte Wahrscheinlichkeiten
Basierend auf Ihren Zählungen:

1. $P(\text{läuft} \mid \text{Hund}) =$ _____
2. $P(\text{schläft} \mid \text{Katze}) =$ _____
3. $P(\text{Hund} \mid \text{Der}) =$ _____

Hinweis

Bedingte Wahrscheinlichkeit $P(B \mid A)$ bedeutet: "Wahrscheinlichkeit von B, gegeben dass A bereits eingetreten ist".
Formel: $P(B \mid A) = \frac{\text{Anzahl(A und B zusammen)}}{\text{Anzahl(A insgesamt)}}$

Teil 3: N-Gramm-Modelle (15 Minuten)

Von Uni- zu Trigrammen

Aufgabe 6: Verstehen Sie den Unterschied
Gegeben sei der Satz: "Der alte Mann ging langsam nach Hause"
Zerlegen Sie den Satz in:
Unigramme (einzelne Wörter):

_____

Bigramme (Wortpaare):

_____

_____

Trigramme (Drei-Wort-Sequenzen):

_____

_____

Aufgabe 7: Welches N ist besser?
Vervollständigen Sie die Sätze basierend auf unterschiedlichen Kontextlängen:

1. **Unigramm** (häufigstes Wort überhaupt): "____"
2. **Bigramm** (nach "Guten"): "Guten ____"
3. **Trigramm** (nach "Ich hätte gerne"): "Ich hätte gerne ____"

Welches Modell liefert die sinnvollste Vorhersage? Warum?

Zum Nachdenken

Kompromiss zwischen Kontext und Datenmenge:

- Mehr Kontext (höheres N) = bessere Vorhersagen
- Mehr Kontext = weniger Trainingsdaten pro Muster
- Typischer Kompromiss: Trigramme oder 4-Gramme

Wahrscheinlichkeiten berechnen

Aufgabe 8: Satzwahrscheinlichkeit
Gegeben seien diese Bigramm-Wahrscheinlichkeiten:

- $P(\text{Katze} \mid \text{Die}) = 0.3$
- $P(\text{schläft} \mid \text{Katze}) = 0.4$
- $P(\text{gerne} \mid \text{schläft}) = 0.6$

Berechnen Sie die Wahrscheinlichkeit des Satzes "Die Katze schläft gerne":
$P(\text{Satz}) = P(\text{Die}) \times P(\text{Katze} \mid \text{Die}) \times P(\text{schläft} \mid \text{Katze}) \times P(\text{gerne} \mid \text{schläft})$
Angenommen $P(\text{Die}) = 0.1$:
$P(\text{Satz}) =$ _____

Teil 4: Das Problem unbekannter Wörter (5 Minuten)

Zero-Probability-Problem

Aufgabe 9: Was passiert mit neuen Wörtern?
Ihr Modell wurde mit diesen Sätzen trainiert:

- "Die Katze schläft"
- "Der Hund bellt"
- "Die Katze läuft"

Nun erscheint der Satz: "Der **Vogel** singt"

1. Was ist $P(\text{Vogel} \mid \text{Der})$ ohne Smoothing? _____
2. Warum ist das problematisch?

3. **Add-One-Smoothing:** Wir tun so, als hätten wir jedes Wort einmal mehr gesehen.
   Original: "Der" → Hund (1×), Vogel (0×)
   Mit Smoothing: "Der" → Hund (2×), Vogel (1×)
   Neue Wahrscheinlichkeit $P(\text{Vogel} \mid \text{Der})$ = _____

Teil 5: Reflexion und Anwendungen (5 Minuten)

Grenzen von N-Gramm-Modellen

Aufgabe 10: Wo versagen N-Gramme?
Betrachten Sie diesen Satz: "Der Schlüssel zu dem Schrank, der in der Küche steht, _____"

1. Ein Trigramm-Modell sieht nur: "Küche steht ___". Was würde es vorhersagen?
   _____
2. Was sollte tatsächlich folgen (denken Sie an "Der Schlüssel")?
   _____
3. Welches Problem zeigt dies auf?

Zum Nachdenken

Zusammenfassung - Was Sie gelernt haben:

- Sprachmodelle sagen das nächste Wort basierend auf Kontext vorher
- N-Gramm-Modelle zählen Wortmuster und berechnen Wahrscheinlichkeiten
- Höheres N = mehr Kontext, aber weniger Trainingsdaten
- Smoothing löst das Problem unbekannter Wörter
- N-Gramme können keine langen Abhängigkeiten modellieren

Ausblick auf Woche 2: Neuronale Sprachmodelle können diese Limitierungen überwinden!

Bonusaufgabe für Zuhause

Sammeln Sie 10 WhatsApp/SMS-Nachrichten und analysieren Sie:

1. Welche Bigramme kommen am häufigsten vor?
2. Wie oft nutzen Sie die Autokorrektur-Vorschläge?
3. Finden Sie Beispiele, wo die Vorhersage falsch war?

Bringen Sie Ihre Beobachtungen zur nächsten Vorlesung mit!