Übung 5: Naive-Bayes-Modelle

Rechenübung

Man kann das Naive-Bayes-Modell für die Klassifikation von Texten einsetzen. Dann sind ganze Texte zu desambiguieren und alle Wörter des zu klassifizierenden Textes bilden die "Kontextwörter". Die Trainingsdaten bestehen hier aus einzelnen Texten und ihrer Klasse.

Gegeben sind die folgenden Trainingstexte für ein Text-Klassifikationsmodell mit den Klassen Sport und Politik:

Sport: *die Bayern holen die Meisterschaft*

Politik: *die Wahl in Bayern wird wiederholt*

- Extrahieren Sie die Häufigkeiten der Wörter und Textklassen.
- Schätzen Sie die Parameter eines Naive-Bayes-Modelles. Machen Sie eine interpolierte Backoff-Glättung zuerst mit den Apriori-Wahrscheinlichkeiten der Wörter und dann mit einer uniformen Verteilung über ein Vokabular von 1000 Wörtern.
- Klassifizieren Sie den Text: *Bayern wiederholt die Prüfung*

Implementierungsaufgaben

Teil 1: Training eines Wortbedeutungsdesambiguierers

Laden Sie das Korpus mit der Adresse `www.cis.uni-muenchen.de/~schmid/lehre/StatNLP/data/zeit-10M-tagged.gz` herunter und dekomprimieren Sie es.

Extrahieren Sie alle Vorkommen der Wörter "Staat" und "Kind" und die benachbarten Inhaltswörter, die maximal 50 Positionen entfernt sind. Inhaltswörter sind Wörter mit den Tags "ADJA", "ADJD", "ADV", "NE", "NN", "VVFIN", "VVINF", "VVPP", "VVIZU". Das Ergebnis dieses Schrittes ist eine Tabelle mit Häufigkeiten von Wortpaaren (s, w), wobei s entweder "Staat" oder "Kind" ist und w ein Kontextwort ist. Ignorieren Sie hier die Wortflexion und arbeiten Sie nur mit den Lemmata in der Datei.

Schätzen Sie die geglätteten bedingten Wahrscheinlichkeiten der Kontextwörter w gegeben die "Bedeutung" s nach der Formel:

$$p(w|s) = \frac{f(s,w) + \alpha(s)\, p(w)}{f_1(s) + \alpha(s)}$$

Hier ist s entweder "Kind" oder "Staat" und $\alpha(s)$ ist die Zahl der *unterschiedlichen* Inhaltswörter, die in der Nachbarschaft von s aufgetaucht sind. $p(w)$ wird wie folgt geschätzt:

$$p(w) = \frac{f_2(w)}{N}$$

wobei die allgemeine Worthäufigkeit $f_2(w)$ wie folgt definiert ist:

$$f_2(w) = f(Staat, w) + f(Kind, w)$$

und die Gesamthäufigkeit N sich so ergibt:

$$N = f_1(Staat) + f_1(Kind)$$

mit

$$f_1(s) = \sum_w f(s, w)$$

Diese Glättungsmethode heißt Witten-Bell-Glättung. Während bei der addiere-λ-Methode bei allen Wortpaaren (s,w) derselbe Wert λ addiert wird, wird hier ein Wert addiert, der proportional zur allgemeinen Wortwahrscheinlichkeit p(w) ist.

Am besten iterieren Sie zunächst über alle extrahierten Wortpaare und berechnen $f_1(s)$ und $f_2(w)$ durch Aufsummieren der Wortpaarhäufigkeiten. Dann iterieren Sie noch einmal über alle Wortpaare (s,w) und berechnen $p(w|s)$.

Für ein Wort w, das nur im Kontext von *Staat* aufgetaucht ist, muss auch die Wahrscheinlichkeit $p(w|Kind)$ berechnet werden und umgekehrt.

Da die beiden Wörter *Kind* und *Staat* annähernd gleich häufig sind, ist es nicht notwendig, die Apriori-Wahrscheinlichkeiten der beiden Klassen zu berücksichtigen, da sie annähernd gleich und damit irrelevant sind.

Speichern Sie die Parameter in einer Datei.

Teil 2

Schreiben Sie ein Programm, welches Testdaten mit Hilfe des gelernten Naive-Bayes-Modelles klassifiziert und testen Sie es mit einigen Beispielen.

Das Programm bekommt zwei Dateinamen als Argumente. In der ersten Datei sind die Modellparameter gespeichert, die in Aufgabenteil 1 berechnet wurden. In der zweiten Datei steht ein zu desambiguierender Text. Das Format der Textdatei entspricht dem Format der Trainingsdaten, abgesehen davon dass die Wörter "Kind" und "Staat" durch das Pseudowort "Kind:Staat" ersetzt wurden. Sie können für diesen Zweck auch einen Teil der Trainingsdaten beiseite legen (und dann nicht im Training verwenden). Für jedes Vorkommen des Pseudowortes "Staat:Kind" im Text sollen $\prod_{w \in C} p(w|Staat)$ und $\prod_{w \in C} p(w|Kind)$ ausgegeben werden, wobei C die Liste der Kontextwörter ist.