1. (4 Punkte) In einer Halle kommt es an 1 % der Tage zu einem Brand. Der Rauchmelder schlägt bei einem Brand mit 99 % Wahrscheinlichkeit Alarm, an Tagen ohne Brand gibt es an 2 % der Tage einen Fehlalarm. Heute schlägt der Melder Alarm.
   (a) Benenne Hypothese, Beobachtung, Prior und Likelihood.
   (b) Wie wahrscheinlich brennt es?

Lösung:

(a) Hypothese H: Es brennt. Beobachtung E: Alarm. Prior $P(H) = 0{,}01$. Likelihood $P(E \mid H) = 0{,}99$, außerdem $P(E \mid \neg H) = 0{,}02$.

(b) $$P(H \mid E) = \frac{0{,}01 \cdot 0{,}99}{0{,}01 \cdot 0{,}99 + 0{,}99 \cdot 0{,}02} = \frac{0{,}0099}{0{,}0099 + 0{,}0198} = \frac{1}{3}$$
Nur bei einem Drittel der Alarme brennt es wirklich: Fehlalarme sind doppelt so häufig wie echte, weil Brände so selten sind.

2. (5 Punkte) Eine Krankheit tritt bei 2 % der Bevölkerung auf. Ein Test ist bei 95 % der Kranken positiv und bei 3 % der Gesunden (falsch) positiv.
   (a) Wie viele von 10 000 Personen sind krank und positiv getestet, wie viele gesund und positiv getestet?
   (b) Wie wahrscheinlich ist eine positiv getestete Person krank?
   (c) Die Person macht einen zweiten, unabhängigen Test derselben Art, der wieder positiv ist. Wie wahrscheinlich ist sie jetzt krank? (Tipp: Der Posterior aus b) ist der neue Prior.)

Lösung:

(a) Krank: 200, davon positiv $0{,}95 \cdot 200 = 190$. Gesund: 9 800, davon positiv $0{,}03 \cdot 9\,800 = 294$.

(b) $\frac{190}{190 + 294} = \frac{190}{484} \approx 0{,}3926$, also knapp 40 %.

(c) $\frac{0{,}3926 \cdot 0{,}95}{0{,}3926 \cdot 0{,}95 + 0{,}6074 \cdot 0{,}03} \approx 0{,}9534$, also gut 95 %. Der zweite Test macht die Diagnose sehr viel sicherer.

3. (6 Punkte) Von 100 E-Mails sind 40 Spam. Gezählt wurde, in wie vielen E-Mails die Wörter „gratis“ und „Gewinn“ vorkommen:

	Spam	kein Spam
Anzahl E-Mails	40	60
enthält gratis	24	6
enthält Gewinn	20	3

(a) Eine neue E-Mail enthält beide Wörter. Berechne mit der naiven Annahme, wie wahrscheinlich sie Spam ist.

(b) Eine E-Mail enthält „gratis“, aber nicht „Gewinn“. Wie wahrscheinlich ist sie Spam?

(c) Was ist an der Annahme naiv?

Lösung:

(a) Erwartete Anzahlen, wenn die Wörter unabhängig vorkämen: Spam $40 \cdot \frac{24}{40} \cdot \frac{20}{40} = 12$, kein Spam $60 \cdot \frac{6}{60} \cdot \frac{3}{60} = 0{,}3$. $P(\text{Spam}) = \frac{12}{12+0{,}3} \approx 0{,}976$

(b) „nicht Gewinn“: bei Spam $\frac{20}{40}$, sonst $\frac{57}{60}$. Spam $40 \cdot \frac{24}{40} \cdot \frac{20}{40} = 12$, kein Spam $60 \cdot \frac{6}{60} \cdot \frac{57}{60} = 5{,}7$. $P(\text{Spam}) = \frac{12}{12+5{,}7} \approx 0{,}678$

(c) Es wird angenommen, dass die Wörter unabhängig voneinander vorkommen. In Wirklichkeit stehen „gratis“ und „Gewinn“ oft zusammen. Man nimmt die Annahme trotzdem, weil man sonst für jede Kombination von Wörtern eigene Zählungen bräuchte.

4. (6 Punkte) Ein naiver Bayes-Klassifikator mit Laplace-Glättung soll Texte den Klassen Sport und Kochen zuordnen. Trainingsdaten:

Text	Klasse
tor tor heute	Sport
sieg heute	Sport
nudeln salz	Kochen
nudeln wasser	Kochen
heute nudeln	Kochen

(a) Gib die Priors an und die Wortanzahlen je Klasse. Wie groß ist das Vokabular?

(b) Berechne $P(\text{heute} \mid \text{Sport})$, $P(\text{tor} \mid \text{Sport})$, $P(\text{heute} \mid \text{Kochen})$ und $P(\text{tor} \mid \text{Kochen})$ mit Laplace-Glättung.

(c) Zu welcher Klasse gehört der Text „heute tor“? Mit welcher Wahrscheinlichkeit?

(d) Was ändert sich beim Text „heute tor mittag“?

Lösung:

(a) $P(\text{Sport}) = \frac{2}{5}$, $P(\text{Kochen}) = \frac{3}{5}$. Sport: tor 2, heute 2, sieg 1, zusammen 5 Wörter. Kochen: nudeln 3, salz 1, wasser 1, heute 1, zusammen 6 Wörter. Vokabular: 6 Wörter.

(b) $P(\text{heute} \mid \text{Sport}) = \frac{2+1}{5+6} = \frac{3}{11}$, $P(\text{tor} \mid \text{Sport}) = \frac{3}{11}$, $P(\text{heute} \mid \text{Kochen}) = \frac{1+1}{6+6} = \frac{2}{12}$, $P(\text{tor} \mid \text{Kochen}) = \frac{0+1}{12} = \frac{1}{12}$.

(c) Sport: $\frac{2}{5} \cdot \frac{3}{11} \cdot \frac{3}{11} \approx 0{,}0298$. Kochen: $\frac{3}{5} \cdot \frac{2}{12} \cdot \frac{1}{12} \approx 0{,}0083$. Sport mit $\frac{0{,}0298}{0{,}0298+0{,}0083} \approx 0{,}78$.

(d) Nichts: „mittag“ kommt in den Trainingsdaten nicht vor, ist nicht im Vokabular und wird übergangen.

5. (4 Punkte) (a) Warum rechnet man beim naiven Bayes-Klassifikator mit den Logarithmen der Wahrscheinlichkeiten? Warum ändert sich dadurch nicht, welche Klasse gewinnt?

(b) Wozu dient die Laplace-Glättung? Was würde ohne sie mit einem Text passieren, der ein Wort enthält, das in den Trainingstexten einer Klasse nie vorkommt?

Lösung:

(a) Bei langen Texten werden hunderte Wahrscheinlichkeiten multipliziert, das Produkt ist so klein, dass der Computer es als 0 speichert. Mit $\log(a \cdot b) = \log a + \log b$ addiert man stattdessen. Der Logarithmus wächst monoton, die Klasse mit dem größten Produkt hat auch die größte Summe.

(b) Ohne Glättung wäre die Wahrscheinlichkeit dieses Wortes für die Klasse 0 und damit das ganze Produkt 0, egal wie gut der Rest des Textes passt. Die Glättung zählt zu jeder Anzahl 1 dazu (und im Nenner die Größe des Vokabulars), so ist keine Wahrscheinlichkeit 0.

6. (4 Punkte) Gegeben sind die drei Dokumente „der hund bellt“, „der hund schläft“ und „die katze schläft“. Es gilt $\text{IDF} = \ln \frac{1+n}{1+\text{df}} + 1$ (n Anzahl der Dokumente, df Anzahl der Dokumente mit dem Wort).
   (a) Berechne die IDF-Werte für „der“, „hund“ und „bellt“.
   (b) Gib den normierten TF-IDF-Vektor des ersten Dokuments an (Einträge für der, hund, bellt).
   (c) Welches Wort hat im ersten Dokument das größte Gewicht? Warum ist das sinnvoll?

Lösung:

(a) der und hund kommen in 2 Dokumenten vor: $\ln \frac{4}{3} + 1 \approx 1{,}2877$. bellt nur in einem: $\ln \frac{4}{2} + 1 \approx 1{,}6931$.

(b) Jedes Wort kommt einmal vor (TF = 1). Länge $\sqrt{1{,}288^2 + 1{,}288^2 + 1{,}693^2} \approx 2{,}487$,
normiert: (0,5179; 0,5179; 0,6809).

(c) „bellt“: Es kommt nur in diesem Dokument vor und unterscheidet es deshalb am besten von den anderen. „der“ und „hund“ stehen auch im zweiten Dokument.

7. (6 Punkte) Schreibe zwei Funktionen für einen naiven Bayes-Klassifikator mit Laplace-Glättung, ohne scikit-learn: `trainiere(texte, klassen)` zählt die Wörter je Klasse (Wörter durch Leerzeichen getrennt), bestimmt das Vokabular und die Priors. `klassifiziere(text, ...)` gibt die wahrscheinlichste Klasse zurück; unbekannte Wörter werden übergangen. Rechne mit Logarithmen. Teste mit den Trainingsdaten „tor tor heute“ (Sport), „sieg heute“ (Sport), „nudeln salz“ (Kochen), „nudeln wasser“ (Kochen), „heute nudeln“ (Kochen) und den Texten „heute tor“, „salz heute“ und „sieg nudeln“.

Lösung:

```
from collections import Counter
import math

def trainiere(texte, klassen):
    anzahl = {k: Counter() for k in klassen}
    for text, k in zip(texte, klassen):
        anzahl[k].update(text.split())
    vokabular = {w for text in texte for w in text.split()}
    prior = {k: klassen.count(k) / len(klassen) for k in anzahl}
    return anzahl, vokabular, prior

def klassifiziere(text, anzahl, vokabular, prior):
    bester, bester_wert = None, None
    for k in anzahl:
        gesamt = sum(anzahl[k].values())
        wert = math.log(prior[k])
        for w in text.split():
            if w in vokabular:
                wert += math.log((anzahl[k][w] + 1) / (gesamt + len(vokabular)))
        if bester is None or wert > bester_wert:
            bester, bester_wert = k, wert
    return bester

texte = ['tor tor heute', 'sieg heute', 'nudeln salz', 'nudeln wasser', 'heute nudeln']
klassen = ['Sport', 'Sport', 'Kochen', 'Kochen', 'Kochen']
modell = trainiere(texte, klassen)
for t in ['heute tor', 'salz heute', 'sieg nudeln']:
    print(t, klassifiziere(t, *modell))
```

Ausgabe: heute tor Sport, salz heute Kochen, sieg nudeln Kochen.