Diskrete Stochastik
Klasse 3Ia/5Ibb/5iCbb

Fachhochschule Nordwestschweiz
Hochschule für Technik

Programmieraufgabe: Bayes-Spam-Filter

Ziel der Aufgabe ist es, einen Bayes-Spam-Filter zu programmieren.

Folgende Schritte werden erwartet:

1. Sie verschaffen sich einen Überblick über die Funktionsweise, etwa via
   (a) `http://www.math.kit.edu/ianm4/~ritterbusch/seite/spam/de`
   (b) `https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering`.
2. Sie erstellen ein Java-Programm, was folgendes leistet:
   (a) Die Emails aus ham-anlern.zip bzw. spam-anlern.zip werden nacheinander eingelesen und als Ham bzw. Spam markiert. Dabei wird für jedes Wort in einer Ham- bzw. Spam-Mail gezählt, in wie vielen Ham- bzw. Spam-Mails das Wort vorkommt. Ein Wort muss dabei kein sinnvolles Wort sein. Sie können also jeweils die gesamte Mail, inklusive Header, einlesen und ein Leerzeichen als Worttrennungssymbol verwenden.
   (b) Sie implementieren eine Funktion, die für eine gegebene Mail die Spamwahrscheinlichkeit gemäss der in der obigen Quelle hergeleiteten Formel berechnet. (Dabei können Sie alle Wörter (und nicht nur die signifikantesten Wörter) zur Berechnung hinzuziehen.)

   Falls ein Wort, etwa `money` in der Anlernphase nur in Ham-Mails vorkam, dann hat eine Mail, die das Wort `money` enthält, eine Spamwahrscheinlichkeit von 0 (**Begründen Sie dies!**), selbst dann, wenn die Mail auch `Viagra, enlargement, Kenia, win,...` enthält. Dies ist natürlich unerwünscht. Fügen Sie deshalb in der Anlernphase jedes Wort, was in einer der Ham-Mails aber in keiner Spam-Mail vorkommt, mit einer "Anzahl" α (sehr klein (< 1)) in den Spam-Korpus ein und umgekehrt für jedes Wort, welches in einer Spam- aber keine Ham-Mail vorkommt.
   (c) Bestimmen Sie geeignete Werte für den Schwellenwert, wann eine Mail als Spam klassifiziert werden soll und für obiges α, so dass ihr Spamfilter gut arbeitet. Nutzen Sie dazu die Mails in ham-kalibrierung.zip und spam-kalibrierung.zip.
   (d) Geben Sie an, wie viel Prozent der Mails in ham-test.zip bzw. spam-test.zip korrekt klassifiziert wurden.
3. Erwartet wird eine Abgabe, bei der nach dem Ausführen des Programms eine Zusammenfassung aller Werte (Schwellenwert, α, Erkennungsraten) angezeigt wird.

Allgemeine Hinweise:

1. Sie können in Gruppen bis zu drei Personen arbeiten.
2. Bei vollständiger Lösung wird auf die Note des kommenden Tests 0.3 drauf addiert. (Aus systemtechnischen Gründen liegt die Erfahrungsnote zwischen 1.0 und 6.0.)
3. Es ist nicht nötig, das Programm hinsichtlich Effizienz zu optimieren.
4. Das Programm sollte verständlich kommentiert sein.
5. Eigentlich gehe ich davon aus, dass Sie aus Fairnessgründen nicht versuchen, zu betrügen. Dennoch werde ich dies (auch mit Hilfe von Tools) kontrollieren. Falls dabei ein Täuschungsversuch festgestellt wird (also: (verschleierte) Kopien von Teilen existierender Programme (Internet oder Kollegen)), wird die Note des nächsten Tests auf 1.0 gesetzt.

Abgabe: 25./27.10.2021