Naive Bayes Algorithmus

* **Multinomial Naive Bayes**

Input: eine Menge von Dokumente D, N die Gesamtzahl von Dokumenten.

Output: bedingte Wahrscheinlichkeit, dass ein Begriff in mehreren Dokumenten vorkommt

* + **Trainingsalgorithmus**

**Pseudocode:**

TrainMultinomialNB (C, D)

**Input:** Ein Array von Dokumenten D, eine Klasse C

**Output**: extrahierte Vokabular, Prior, die bedingte Wahrscheinlichkeit  
 := entspricht die Anzahl an Dokumente in der Klasse c

N: =die Gesamtzahl von Dokumenten  
 := Anzahl der Vorfälle von t Begriffe in Trainingsdokumenten aus Klasse c, sowie mehrfacher Vorfall eines Begriffes in einem Dokument

D: = Array von Dokumenten  
C: = Dokumentenklasse

t: = Begriff im Dokument

V: = ExtrahiereVokabular (D)

N: = zaehleDokumente(D) (Dokumenten des Arrays D werden gezählt)

String ExtrahiereVokabular (D): = Vokabular wird von Dokumenten des Arrays D extrahiert

Int zaehleDokumente(D): = Methode zum Zählen der Dokumente in Array D

Int zaehleDokumenteInKlasse (D, c): Methode zum Zählen der Dokumenten des Arrays D in der C

String VerketteTextallerDokumenteinKlasse (D, c): Methode zur Verkettung von Dokumenten des Arrays C in der Klasse c

bedingtewahrscheinlichkeit [t] [c] : =relative Häufigkeit von Begriffe t in Dokumente der Klasse c

int Prior[c] := Prior Wahrscheinlichkeit eines in der Klasse C auftretendes Dokument (Variable)

**for each** c ∈ C  
**Do**  =zaehleDokumenteInKlasse (D, c)  
 Prior [c]= /N  
 = VerketteTextallerDokumenteinKlasse (D, c)  
**for each** t ∈ V  
**do** bedingtewahrscheinlichkeit [t] [c]=

**return** V, Prior, bedingtewahrscheinlichkeit

* + **Testalgorithmus**

**Pseudocode:**

Input: Menge von Dokumenten D, eine Klasse C, bedingtewahrscheinlichkeit

Output: die beste Klasse (maximale a posteriori Klasse) in der Textklassifizierung

D: = Array von Dokumenten  
C: = Dokumentenklasse  
d: = Dokument in D  
V: = ExtrahiereVokabular (D) (Vokabular wird von Dokumenten des Arrays D extrahiert)  
String ExtrahiereMerkmalVomDokument (V, d): = Methode zur Extraktion bestimmter Merkmale im extrahierten Vokabular eines Dokumentes

int Prior[c] := Prior Wahrscheinlichkeit eines in der Klasse C auftretendes Dokument (Variable)

Int Punktzahl[c] := Variable zum Ausdruck der Punktzahl für die Klasse c

W= ExtrahiereMerkmalVomDokument (V, d)  
**for each** c ∈ C  
**do** Punktzahl[c]=log Prior[c]  
**for each** t ∈ W  
**do** Punktzahl[c]+=log bedingtewahrscheinlichkeit[t][c]

**Return** Punktzahl[c] (Punktzahl mit der maximalen Wert)

* **Bernouilli Naive Bayes**
  + **Trainingsalgorithmus**

TrainBernouilliNB (C, D)

**Input:** Menge von Dokumenten D, eine Klasse C

**Output**: extrahierte Vokabular, Prior, die bedingte Wahrscheinlichkeit

:= entspricht die Anzahl an Dokumente in der Klasse c

N: =die Gesamtzahl von Dokumenten

D: = Array von Dokumenten

C: = Dokumentenklasse

t: = Begriff im Dokument

V: = ExtrahiereVokabular (D) (Vokabular wird von Dokumenten des Arrays D extrahiert)

N: = zaehleDokumente(D) (Dokumenten des Arrays D werden gezählt)

Int zaehleDokumenteInKlasse (D, c): Methode zum Zählen der Dokumenten des Arrays D in der C

String VerketteTextallerDokumenteinKlasse (D, c): Methode zur Verkettung von Dokumenten des Arrays C in der Klasse c

int Prior[c] := Prior Wahrscheinlichkeit eines in der Klasse C auftretendes Dokument (Variable)

int zaehleMerkmalvomBegriff ():= Methode zum Zählen der Merkmale von Begriffe t in den verketteten Text aller Dokumente der Klasse c

**for each** c ∈ C

**do** =zaehleDokumenteInKlasse (D, c)

prior [c]= /N

= VerketteTextallerDokumenteinKlasse (D, c)

**for each** t ∈ V  
**do** int = zaehleMerkmalvomBegriff ()  
**for each** t ∈ V

**do** bedingtewahrscheinlichkeit [t] [c] =

**return** V, prior, bedingtewahrscheinlichkeit

* + **Testalgorithmus**

**Pseudocode:**

Input: Menge von Dokumenten D, eine Klasse C, bedingtewahrscheinlichkeit

Output:

D: = Array von Dokumenten

C: = Dokumentenklasse

d: = Dokument in D

V: = ExtrahiereVokabular (D) (Vokabular wird von Dokumenten des Arrays D extrahiert)

= ExtrahiereBegriffeVomDokument (V, d)

String ExtrahiereBegriffeVomDokument (V, d): = Methode zur Extraktion von Begriffe des Dokument d in das Array D von Dokumenten

String ExtrahiereVokabular (D)

int Prior[c]: = Prior Wahrscheinlichkeit eines in der Klasse C auftretendes Dokument (Variable)

Int Punktzahl[c]: = Variable zum Ausdruck der Punktzahl für die Klasse c

bedingtewahrscheinlichkeit [t] [c]: =relative Häufigkeit von Begriffe t in Dokumente der Klasse c

**for each** c ∈ C

**do** Punktzahl[c]=log prior[c]

**for each** t ∈ V

**do** **if** t ∈   
 **then** Punktzahl[c]+=log bedingtewahrscheinlichkeit[t][c]  
 **else** Punktzahl[c] += log (1 – bedingtewahrscheinlichkeit [t] [c])

**Return**  Punktzahl[c] (Punktzahl mit der maximalen Wert)