LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

NACHHOLKLAUSUR ZUR VORLESUNG BACHELORMODUL ,,SYMBOLISCHE PROGRAMMIERSPRACHE“ WS 2020/2021

FLORIAN FINK

NACHHOLKLAUSUR ZUR VORLESUNG AM 13.04.2021

VORNAME:	
NACHNAME:	
MATRIKELNUMMER:	
STUDIENGANG:	B.Sc. Computerlinguistik, B.Sc. Informatik, Magister

Die Klausur zur Vorlesung besteht aus **7 Aufgaben** und umfasst **12 Seiten**. Die Punktzahl ist bei jeder Aufgabe angegeben. Die Bearbeitungsdauer beträgt **45 Minuten**. Bitte überprüfen Sie, ob Sie ein vollständiges Exemplar erhalten haben.

Aufgabe	mögliche Punkte	erreichte Punkte
1. Precision, Recall und Accuracy	6	
2. Quiz	5	
3. Naive-Bayes	8	
4. Objektorientierung	4	
5. Information Retrieval	6	
6. NLTK und lexikalische Information	7	
7. POS Tagging	4	
Summe	40	
Note		

Einwilligungserklärung
Hiermit stimme ich einer Veröffentlichung meines Klausurergebnisses unter Verwendung meiner Matrikelnummer im Internet zu.

Datum: _____ Unterschrift: _____

NAME:

Aufgabe 1 *Precision, Recall und Accuracy* [6 Punkte]
Ein Klassifizierer teilt 100 Texte in die 4 Klassen A, B, C und D ein. Bei der Auswertung ergibt sich folgende Wahrheitsmatrix (confusion matrix):

```
  |  A  B  C  D |
--+-------------+
A |<30> 7  13 3 |
B |  8 <1> 4  3 |
C |  7  5 <3> 3 |
D |  7  1  2 <3>|
--+-------------+
(row = reference; col = test)
```

Berechnen Sie

(a) die Accuracy des Klassifizierers,

(b) die Precision des Klassifizierers für Klasse C,

(c) den Recall des Klassifizierers für Klasse C,

Ihr Rechenweg muss nachvollziehbar sein.

(2+2+2 = 6 Punkte)

NAME:

Aufgabe 2 *Quiz* [5 Punkte]
Bearbeiten Sie folgende Aufgaben.

(a) Die beiden Ereignisse A und B seien *statistisch unabhängig*. Wie kann die Wahrscheinlichkeit $P(A$ und $B)$ berechnet werden (in Abhängigkeit von $P(A)$ und $P(B)$)?

(b) (Anmerkung: Das Vorkommen von Worten ist im Allgemeinen **nicht** *statistisch unabhängig*). Geben Sie an, wie die Wahrscheinlichkeit des Satzes *„heute schneit es"* berechnet wird (Wahrscheinlichkeit von Wortsequenzen).

(c) Berechnen Sie die Länge $|\vec{x}|$ des Vektors $\vec{x} = (4, 2, 4)$.

(d) Berechnen Sie das Skalarprodukt $\vec{x} \cdot \vec{y}$ zwischen den Vektoren $\vec{x} = (2, 4)$ und $\vec{y} = (3, 2)$.

(e) Wozu dient die sog. *Addiere-1 Glättung*?

(1+1+1+1+1 = 5 Punkte)

NAME:

Aufgabe 3 *Naive-Bayes* [8 Punkte]

Der Naive-Bayes Algorithmus ist ein Algorithmus zur Klassifikation. Gehen Sie für diese Aufgabe davon aus, dass wir den Naive-Bayes Algorithmus zur Klassifikation von E-Mails verwenden. E-Mails sollen dabei entweder in die Klasse `HAM` (E-Mail ist keine Spam-E-Mail) oder in die Klasse `SPAM` (E-Mail ist eine Spam-E-Mail) eingeteilt werden.

(a) Handelt es sich beim Naive-Bayes Algorithmus um einen *überwachten (supervised)* oder einen *unüberbwachten (unsupervised)* Algorithmus (begründen Sie Ihre Aussage)?

(b) Wie könnten mögliche Trainingsdaten für den Naive-Bayes Algorithmus zur Klassifikation von E-Mails aussehen?

(c) Beschreiben Sie, wie mit dem Naive-Bayes Algorithmus eine E-Mail klassifiziert wird.

(2+2+4 = 8 Punkte)

NAME:

NAME:

Aufgabe 4 *Objektorientierung* [4 Punkte]

```
import nltk
import os

class DocumentCollection:
    def __init__(self, docs):
        self.docs = docs
    @classmethod
    def from_dir(cls, d):
        cls([Document.from_file(p) for p in os.listdir(d) if p.endswith(".txt")])
    def x(self, term):
        return len([d for d in self.docs if term in d.counts])

class Document:
    def __init__(self, counts):
        self.counts = counts
    @classmethod
    def from_file(cls, path):
        with open(path, mode='r', encoding='utf-8') as f:
            return cls(nltk.FreqDist(nltk.word_tokenize(f.read())))
    def y(self, term):
        return self.counts.get(term, 0)
```

Betrachten Sie folgenden objektorientierten Python-Code.

(a) Was berechnet die Funktion `x` in der `DocumentCollection` Klasse?

(b) Was berechnet die Funktion `y` in der `Document` Klasse?

(c) Was enthält das Attribut `docs` in der `DocumentCollection` Klasse?

(d) Was ist die Bedeutung der `self` Variable in den beiden Klassen?

(1+1+1+1 = 4 Punkte)

NAME:

NAME:

Aufgabe 5 *Information Retrieval* [6 Punkte]

Erläutern Sie die Dokumentensuche mit dem Vektorraummodell (Vektor Space Model). Gehen Sie dabei auch auf die Dokumentenrepräsentation, die Gewichtungsfunktion, die Ähnlichkeitsberechnung und die Bearbeitung von Suchanfragen ein.

NAME:

NAME:

Aufgabe 6 *NLTK und lexikalische Information* [7 Punkte]

(a) Definieren Sie die Begriffe *Token*, *Type* und *Konkordanz*.

(b) Betrachten Sie folgenden Code. Was enthalten jeweils die Variablen `ws`, `bs`, `cfd` und `result`?

```
import nltk
ws = nltk.corpus.brown.words(categories="fiction")
bs = nltk.bigrams(ws)
cfd = nltk.ConditionalFreqDist(bs)
result = cfd["living"].max()
```

(3+4 = 7 Punkte)

NAME:

NAME:

Aufgabe 7 *POS Tagging* [4 Punkte]
Gegeben sei die Hypothese: *„Ein Satz endet niemals mit einer Präposition"*. Beschreiben Sie, wie man mit NLTK in einem Korpus diese Hypothese überprüfen kann.