Aufgabenblatt 6 — 28.11.2019

Einführung in die Computerlinguistik (WS 19/20) — Abgabe bis 7.12.2019

Aufgabe 1) Naive Bayes Classification **Punkte: 8**

Gegeben ist das folgende Trainingsset und Testdokument:

	ID	Dokument	c=Veganismus
Traininsset	3	Tofu Margarine	+
	4	Hefe Seitan Seitan Tofu	+
	1	Milch Käse Joghurt	-
	2	Käse Butter Käse	-
Testset	5	Käse Hefe Seitan Hefe	?

a) Multinomial Naive Bayes *Punkte: 4*

Schätzen Sie einen **multinomialen** Naive Bayes Klassifizier auf Basis der Trainingsdaten. Welche Klasse weist Ihr Klassifizier dem Dokument aus dem Testset zu? Geben Sie alle Berechnungen als Teil der Lösung mit ab. Benutzen Sie Laplace-Smoothing.

b) Binomial Naive Bayes *Punkte: 4*

Schätzen Sie einen **binomialen** Naive Bayes Klassifizier auf Basis der Trainingsdaten. Welche Klasse weist Ihr Klassifizier dem Dokument aus dem Testset zu? Geben Sie alle Berechnungen als Teil der Lösung mit ab. Benutzen Sie Laplace-Smoothing. Woraus erklären sich die Unterschiede zum multinomialem Modell?

Aufgabe 2) Evaluation **Punkte: 4**

Sie richten ein geselliges Abendessen aus und möchten den Ernährungsgewohnheiten all Ihrer Gäste entgegenkommen. Dazu konsultieren Sie einen Multiclass-Klassifizierer, um zwischen vegetarischen (aber nicht veganen), veganen und fleischhaltigen Rezepten zu unterscheiden. Nach manueller Auswertung der Ergebnisse erhalten Sie diese Konfusionsmatrix (hierbei steht z.B. *meat* für die wirklich richtige Lösung und *meat'* für das Resultat des Klassifizierers).

system labels \ gold labels	**meat**	**veggy**	**vegan**	
meat′	109	5	0	114
veggy′	20	65	7	92
vegan′	6	2	7	15
	135	72	14	

Berechnen Sie nun die Precision und Recall-Werte für jede einzelne Klasse, sowie die Micro- & Macroaverage Precision.

Aufgabe 3) Textklassifikation Fragen **Punkte: 8**

1. Wir haben im multinomialen Modell und im binomialen Modell, die a-priori Wahrscheinlichkeit der Klasse c mit

   $P(c) = (Anzahl\ der\ Dokumente\ der\ Klasse\ c)/(Gesamtanzahl\ Dokumente)$

   geschätzt. Im multinomialem Modell hätten wir ja auch die folgende Formel wählen können

   $(Anzahl\ der\ Worte\ in\ Dokumenten\ der\ Klasse\ c)/(Gesamtanzahl\ Worte\ in\ allen\ Dokumenten)$

   Dies wäre analog zu den Featureschätzungen. Warum wird dies wohl nicht getan? (2 Punkte)

2. Die Verwendung von Mutual Information als Merkmalsselektion ist (theoretisch) am geeignetsten für das Bernouillimodell. Warum ist dies so? Wie könnte man die Formel und Berechnungen modifizieren, so dass sie auch für das Multinomialmodell geeignet ist? (2 Punkte)

3. Was ist der Wert von Mutual Information, wenn Klasse und Wort/Term komplett abhängig sind? Was ist der Wert von MI, wenn Klasse und Wort/Term komplett unabhängig sind? (4 Punkte)

Aufgabe 4) Bonusaufgabe: Textklassifikation mit non-word features **Punkte: 4**

Sie möchten einen Spamklassifizierer schreiben. Ihnen stehen folgende Merkmale und Trainingsdaten zur Verfügung.

1. Anzahl der Ausrufezeichen
2. Menge an "Bild" im Vergleich zu "Text"
3. Ist die *subject line* in *all caps*?
4. Vorkommen des Wortes *Gewinn*
5. Ist die subject line all caps?

	ID	Ausrufezeichen	Bild/Text	all caps	"Gewinn"	Spam
Trainingsset	1	viel	niedrig	ja	ja	+
	2	wenig	niedrig	ja	ja	+
	3	wenig	mittel	ja	nein	+
	4	viel	mittel	nein	nein	+
	5	mittel	hoch	nein	ja	+
	6	viel	hoch	nein	ja	+
	7	viel	mittel	nein	nein	+
	8	mittel	niedrig	nein	nein	-
	9	wenig	niedrig	ja	nein	-
	10	wenig	niedrig	nein	ja	-

Beachten Sie, dass es Sie hier ihr Modell leicht abwandeln müssen, denn Sie müssen nun ja Wahrscheinlichkeiten wie die untigen schätzen:

- $P(Ausrufezeichen = wenig|+)$
- $P(Ausrufezeichen = mittel|+)$

1. Wie würden Sie diese Wahrscheinlichkeiten schätzen? (Smoothing können Sie hier einmal ignorieren.). (1 Punkt)
2. Berechnen Sie nun die Bewertung für eine email mit den Features **[mittel, mittel, ja, ja]** mit einem NB-Äquivalent (ohne Smoothing). Hierzu müssen Sie nur die Wahrscheinlichkeiten schätzen, die für genau diese Test-Email notwendig sind. (2 Punkte)
3. Was wäre die Wahrscheinlichkeit für $p(Ausrufezeichen = viel|ja)$ mit Laplace-Smoothing (1 Punkt)