

**Aufgabe 1)** Heaps Gesetz und Zipfs Gesetz**Punkte: 5**

Im folgenden nehmen wir an, dass für das jeweilige Korpus das Zipfsche und Heapsche Gesetz gilt. Beim Zipfschen Gesetz betrachten wir hier keine Modifikationen oder Verfeinerungen.

1. In einem Korpus  $C1$  wird das häufigste Token 120mal gesehen. Wie häufig wird dann das dritthäufigste Token gesehen? (1 Punkt)
2. In einem Korpus  $C2$ , das aus 20 Millionen Tokens besteht, gilt das folgende. In den ersten 10 000 Tokens finden sich 3000 verschiedene Wörter (word types). In der ersten 1 Million (= 1 000 000) Tokens finden sich 30 000 verschiedene Wörter (types). Wie viele verschiedene Wörter (types) erwarten Sie dann im ganzen Korpus? Bitte berechnen Sie die Parameter von Heaps Law, die für dieses Korpus gelten, als Teil Ihrer Antwort. (3 Punkte)
3. Gegeben ein Korpus  $C3$ . Ab welchem Rang stehen die Wörter, die nur einmal gesehen wurden? Ab welchem Rang, die die höchstens zweimal gesehen wurden? (Die wirkliche Zahl ist natürlich korpusabhängig, Sie können aber den Rang mit Variablen ausdrücken). (1 Punkt).
4. Zum Nachdenken (unbenotet und unbepunktet): Wie sind Zipfs Law und Heaps Law wohl verwandt? Folgt eines aus dem anderen?

**Aufgabe 2)** Wahrscheinlichkeitsberechnungen**Punkte: 8**

1. Folgende Tabelle beinhaltet die häufigsten zeichenbasierten Uni-, Bi- und Trigramme aus dem 'Monty Python and the Holy Grail'-Buch aus nltk.book. Das Buch beinhaltet insgesamt 1.259.862 Zeichen. Der Underscore steht für das Leerzeichen.

Unigram	Bigram	Trigram
260818 _	42367 e _	24648 _ t h
115855 e	33867 _ t	18749 t h e
85539 t	29866 t h	18713 _ , _
75266 a	26833 s _	16143 h e _
68338 o	25979 h e	8491 e d _
64431 n	23293 d _	8371 n d _
62022 s	23082 t _	8010 n g _
61891 i	21912 _ a	7995 i n g
61434 h	19309 i n	7474 a n d
51311 r	19155 _ ,	7430 _ a n

Berechnen Sie auf Basis der Statistiken folgende Wahrscheinlichkeiten. Hierbei stehen die mit Komma ausgedrückten Verundungen für n-gramme, also Sequenzen (3 Punkte):

- a)  $P(-), P(r)$
- b)  $P(-, t), P(i, n, g)$
- c)  $P(h, e|t), P(g|i, n)$
2. Nach Ihrem Tod stehen Sie vorm Himmelstor. Da es leider schon recht voll im Himmel ist, bekommen Sie von einem Engel vier gleich aussehende, geschlossene Kästchen  $A, B, C, D$  zur Auswahl. In einem liegt ein kleiner goldener Schlüssel, der Sie in den Himmel lässt. In allen anderen liegt nichts und Sie kommen in einen Warteraum, in dem die ganze Zeit Musik von Lewis Capaldi läuft.<sup>1</sup> Sie wählen ein Kästchen (z.B.  $A$ ) aus. Nachdem Sie Ihre Wahl getroffen haben, öffnet der Engel, der natürlich den Inhalt aller Kästchen kennt, eines der anderen drei Kästchen (z.B.  $B$ ) und zeigt Ihnen, dass in diesem nichts ist. Es wird Ihnen nun angeboten, entweder Ihre ursprüngliche Wahl auf eines der beiden übrigen Kästchen zu ändern (also von z. B.  $A$  auf  $C$  oder  $D$ ), oder Sie können bei Ihrer ursprünglichen Wahl bleiben. Was tun Sie? Begründen Sie Ihre Antwort, indem Sie die jeweiligen Wahrscheinlichkeiten in den Himmel zu kommen, berechnen. Begründen Sie das Resultat Ihrer Berechnungen auch natürlichsprachlich. (4 Punkte)
3. Sandra benutzt zwei Maschinen zur Herstellung eines Produktes: 60% der Produktion werden mit  $M1$  hergestellt und 40% der Produktion mit  $M2$ . Die Wahrscheinlichkeit, dass  $M1$  fehlerhaften Output liefert, ist 10% und die Wahrscheinlichkeit, dass  $M2$  fehlerhaften Output liefert, ist 1%. Wieviel Prozent der Produktion ist im Durchschnitt fehlerhaft? (1 Punkt)

---

<sup>1</sup>Es steht also einiges auf dem Spiel.

**Aufgabe 3)** Analyse von bedingten Wahrscheinlichkeiten

**Punkte: 4**

Gegeben sei ein diskreter Wahrscheinlichkeitsraum und zwei Ereignisse  $A$  und  $B$ . Analysieren Sie das Verhältnis zwischen  $P(A|B)$  und  $P(A)$ . Insbesondere:

- Wann sind die beiden Wahrscheinlichkeiten gleich?
- Nennen Sie Fälle, in denen  $P(A|B)$  größer als  $P(A)$  ist (konstruieren Sie einen allgemeinen Fall sowie Beispiele aus der Münz- oder Würfelwelt)
- Nennen Sie einfache Fälle, in denen  $P(A|B)$  kleiner als  $P(A)$  ist.
- Leiten Sie allgemeine Größenverhältnisse ab.

**Aufgabe 4) Entropie****Punkte: 3**

1. Gegeben sind die folgenden drei Zufallsvariablen mit Wahrscheinlichkeitsverteilung. Bitte ordnen Sie diese nach Ihrer Entropie (von größter Entropie nach niedrigster). Begründen Sie Ihre Antwort kurz. Eine Berechnung der Entropie ist nicht erwünscht. (1 Punkt)

$$p(X) := \{1/3, 1/3, 1/3\}$$

$$p(Y) := \{1/2, 1/2\}$$

$$p(Z) := \{1/5, 1/5, 1/5, 1/5, 1/5\}$$

2. Gegeben sind die folgenden drei Zufallsvariablen mit Wahrscheinlichkeitsverteilung. Bitte ordnen Sie diese nach Ihrer Entropie (von größter Entropie nach niedrigster). Begründen Sie Ihre Antwort kurz. Eine Berechnung der Entropie ist nicht erwünscht. (2 Punkte)

$$p(A) := \{0.25, 0.25, 0.25, 0.25\}$$

$$p(B) := \{0.5, 0.4, 0.09, 0.01\}$$

$$p(C) := \{0.5, 0.4, 0.05, 0.05\}$$

**Aufgabe 5)** Bonusaufgabe: Zipfs Gesetz in der Praxis**Punkte: 4**

Sie setzen Ihre Arbeit auf dem Tweetcorpus aus Aufgabenblatt 3 fort. Jetzt schauen wir uns die positiv gelabelten Tweets an, indem wir die Token in eine Liste laden:<sup>2</sup>

```
import nltk
from nltk.corpus import twitter_samples
pos_tweets = twitter_samples.tokenized('positive_tweets.json')
pos_corpus = []
for tweet in pos_tweets:
    pos_corpus.extend(tweet)
print(pos_corpus[:10])
# ['#FollowFriday', '@France_Inte', '@PKuchly57', '@Milipol_Paris',
#  'for', 'being', 'top', 'engaged', 'members', 'in']
```

Erstellen Sie wie folgt eine Unigramm-Verteilung:

```
from nltk.util import ngrams # ein ngram-Generator

pos_dist = nltk.FreqDist(list(ngrams(pos_corpus , 1)))
```

Diese können Sie nun mit `pos_dist.plot(n)` plotten (hierzu muss das Python Modul `matplotlib` installiert sein), indem Sie der Funktion einen `int n` Wert geben, ab welchem Rang keine Wörter mehr geplottet werden sollen

1. Probieren Sie verschiedene Werte von `n` aus. Ab welchem Wert für `n` beobachten Sie eine Unigrammverteilung, wie Sie sie in der Vorlesung kennen gelernt haben? (Hinweis: für große `n` braucht die Methode `dist_plot()` sehr lange). (1 Punkt)
2. Erstellen Sie jetzt ein neues Minikorpus zusammen mit der zugehörigen Unigramm-Verteilung. In diesem neuen Korpus generieren wir "Wörter" rein zufällig aus den Charakteren `a, b, c, d, e` sowie einem Space.

```
import random
random_words = "".join([random.choice('abcde ') for _ in range(500000)]).split(" ")
random_dist = nltk.FreqDist(list(ngrams(random_words, 1)))
```

Plotten Sie auch diese Verteilung. Folgt Sie dem Zipf'schen Gesetz? Was war Ihre Erwartung? Oft wird gesagt, dass das Zipfsche Gesetz für natürliche Korpora dadurch zustande kommt, dass Menschen kürzere Wörter für oft verwendete Konzepte wählen (*principle of least effort*)— ist dies in Anbetracht des Resultats wahrscheinlich? (3 Punkte)

---

<sup>2</sup>Je nachdem, wo Sie gearbeitet haben, kann es sein, dass Sie auch noch den `download`-Befehl aus Aufgabenblatt 3, Aufgabe 3, ausführen müssen.

**Aufgabe 6)** Unbenotet und unkorrigiert: Verteilungen**Punkte:** –

In einer Versuchsreihe werden 10 Personen jeweils zwei verschiedenen Tests unterzogen. Der erste Test ist ein standardisierter Eignungstest einer Universität und liefert eine Bewertung zwischen 1 und 100. Sei  $X$  die Zufallsvariable, welche den ersten Test repräsentiert. Der zweite Test ist ein Gedächtnistest, bei dem die Probanden sich Zahlenketten fehlerfrei merken sollen, und ergibt pro Proband eine Anzahl von Zahlen, welche sich der Proband merken kann, wobei die Werte zwischen 1 und 5 schwanken. Sei  $Y$  die Zufallsvariable, welche den zweiten Test repräsentiert.

Die Versuchsergebnisse können folgender Tabelle entnommen werden:

Person	1	2	3	4	5	6	7	8	9	10
$x$	80	30	40	40	60	30	40	40	50	50
$y$	5	3	3	3	3	4	4	4	4	4

1. Berechnen Sie die Verteilungen von  $X$  und  $Y$ .
2. Berechnen Sie die multivariante/gemeinsame Verteilung (joint distribution) von  $X$  und  $Y$  sowie die Randverteilung/Marginalverteilung
3. Sind  $X$  und  $Y$  unabhängig verteilt?
4. Berechnen Sie die bedingten Wahrscheinlichkeiten (conditional probability) von  $X$ , gegeben  $Y = 3$ .