Web Mining im SoSe 2017 – Übung 1

Ingo Adrian und Steffen Pegenau
7. Mai 2017

Aufgabe 1

Aufgabenstellung:
Überlegen Sie sich eine neuartige, originelle Web Mining Anwendung, die mit Text-Klassifikationsverfahren gelöst werden könnte. Skizzieren Sie eine mögliche Umsetzung (z.B. Sammlung der Trainingsdaten, Klassifikation der Trainingsdaten, Einsatz des gelernten Klassifikators in der Praxis, etc.) (2 Punkte)

Lösung:
Für die Qualität einer wissenschaftlichen Literaturrecherche ist unter anderem die Herkunft und Art der referenzierten Werke entscheidend. Um die Selektion zu unterstützen, sollen die Ergebnisse einer Suche auf Google Scholar klassifiziert werden.
Die Umsetzung soll folgendermaßen Ablaufen:

1. An einem Fachgebiet wird ein Ranking von Quellen festgelegt. Beispiel: Journal A ist besser als Journal B, aber schlechter als Konferenz C.

2. Quellen, die am Fachgebiet vorhanden sind dienen als Trainingsdaten

3. Die Quellen werden dem Ranking entsprechend klassifiziert.

4. Für einen Browser wird ein Plugin entwickelt, das sich bei zukünftigen Google Scholar Recherchen einklinkt. Dabei werden die ersten n Ergebnisse klassifziert und dem Nutzer nach absteigender Qualität neu sortiert angezeigt.

Aufgabe 2

Aufgabenstellung:
Schreiben Sie ein einfaches Programm, das eine sortierte Liste der in einem Text vorkommenden Worte (im weitesten Sinn alles was durch Leerzeichen begrenzt wird) mit den assoziierten Häufigkeiten (absolut und prozentual) erstellt und sortiert ausgibt. (2 Punkte)

Vergleichen Sie anhand der Ausgabe Ihres Programms die 30 am häufigsten vorkommenden Worte in zwei oder mehreren längeren Texten der gleichen Sprache (z. B. E-books, Projekt Gutenberg, etc.). Wählen Sie einge geeignete Darstellung für Ihren Vergleich. Sind diese Worte als Merkmale für Text-Klassifizierungs-Aufgaben geeignet? Warum? Modifizieren Sie Ihr Programm dahingehend, daß es eine Liste von Stoppwörtern erhalten kann, die ignoriert werden. Wiederholen Sie die vorherige Aufgabe, indem Sie

jedoch diesmal die Stoppwörter der jeweiligen Sprache ignorieren (eine Auswahl finden Sie unter `http://www.nltk.org/nltk_data/packages/corpora/stopwords.zip`). Wie würden Sie nun die Eignung der 30 häufigsten Wörter einschätzen?

Lösung:
Als zu vergleichende Texte wurden *Frankenstein* von Mary Shelley und *Die Verwandlung* von Franz Kafka in der englischen Übersetzung gewählt. Beide Werke wurden als Textdatei vom Projekt Gutenberg bezogen. Generische Textpassagen, die beispielweise Lizenzinformationen beinhalten wurden manuell entfernt.
Beim Betrachten der Liste (Abb. 1) fällt auf, dass die 30 häufigsten Wörter beider Texte zum größten Teil Pronomen (wie *I*, *he* oder *you*) oder Konjunktionen (*and*, *for*) und Artikel (*the*) sind. Da sich diese Wörter in quasi jedem englischen Text finden, sind sie nahezu bedeutungslos im Sinne der Text-Klassifizierung. Nur durch Kenntnis dieser Wörter ist es praktisch unmöglich, Rückschlüsse auf den Inhalt des Textes zu ziehen.
Die Einbeziehung einer Liste mit Stopwords soll genau solche Fälle verhindern. In einer solchen Liste sind Wörter enthalten, die keinerlei Aussagekraft über den Inhalt des Textes liefern und deshalb bei der Analyse außen vor gelassen werden sollen. Unter Nichtbeachtung dieser Wörter stellen sich die 30 häufigsten Wörter beider Texte wie in Abb. 2 dar.
Nun befinden sich unter den 30 Wörtern auch solche, die zumindest grob Rückschlüsse auf den Inhalt der Texte zulassen, wie z. B. *saw*, *time*, *father* (Frankenstein) oder *gregor*, *room*, *sister* (Die Verwandlung).

Aufgabe 3

Aufgabenstellung:
Die Auftrittswahrscheinlichkeiten von Worten in Texten folgen einer sogenannten Zipf-Verteilung, d. h. einer Verteilung, die doppelt logarithmisch ist. Überprüfen Sie das anhand der gewählten Texte. (2 Punkte)
Plotten Sie die Häufigkeiten (y-Achse) über den Rang (x-Achse), also die Anzahl der Vorkommnisse des häufigsten Wortes zuerst, dann die Anzahl des zweithäufigsten Wortes, etc. Betrachten Sie sowohl eine absolute als auch eine logarithmische Skalierung beider Achsen. Was können Sie beobachten?
Bestimmen Sie die Anzahl der Worte, die mit einer gegebenen Häufigkeit vorkommen (also, wie viele Wörter gibt es, die mit Häufigkeit 1 vorkommen, wie viele mit Häufigkeit 2, etc.). Produzieren Sie ähnliche Grafiken (Anzahl der Worte mit einer gewissen Häufigkeit über die Häufigkeit) und interpretieren Sie diese.
Lösung:
Die Zipf-Verteilung beschreibt, dass die Auftrittswahrscheinlichkeit eines Elementes umgekehrt proportional von seiner Position n in einer absteigend nach Häufigkeit geordneten Liste von N Elementen abhängt:[1]

$$p(n) = \frac{1}{H_N} \cdot \frac{1}{n}$$

[1] `https://de.wikipedia.org/wiki/Zipfsches_Gesetz`

mit

$$H_N = \sum \frac{1}{n}$$

Inwiefern diese Verteilung Gültigkeit für die Worthäufigkeiten besitzt, wurde anhand der beiden Texte aus Aufgabe 2 geprüft. Die mit einem Python-Skript[2] gesammelten Daten sind in den Tabellen 7 und 8 für $n = 40$ dargestellt. Ob die Worthäufigkeiten Zipf-verteilt sind, lässt sich anhand der letzten Spalte der Tabellen abschätzen, die die prozentuale Abweichung des Erwartungswertes vom tatsächlichen Wert angibt. Beispiel: Das Wort „the“ kommt im Frankenstein 4195 Mal vor, während die geschätze Häufigkeit 7935,879 beträgt. Damit kommt das Wort 89,17% weniger häufig vor, als von der Verteilung prognostiziert.
Für eine solide Aussage über die Verteilung wären Methoden der Statistik nötig. Es fällt aber auf, dass unter den 20 häufigsten Wörtern im Frankenstein bei 19 der Erwartungswert mehr als 10% vom tatsächlichen Wert abweicht; in der Verwandlung sind es 18.
Zur grafischen Analyse wurden in den Abbildungen 3 und 5 die Worthäufigkeit (y-Achse) in Abhängigkeit der Position in der absteigend sortierten Liste (x-Achse) abgetragen. Da es sich bei der Zipfverteilung um eine doppeltlogarithmische Verteilung handelt, wurden die selben Zahlen in den Abbildungen 4 und 6 mit logarithmierten Skalen dargestellt. Für die Zipf-Verteilung spricht, dass beide Datensätze die Charakteristik einer $1/n$ aufweisen und sich in den logarithmierten Grafiken einer fallenden Geraden annähern.

[2]Zu finden in `aufg03/main.py`

Aufgabe 4

Aufgabenstellung:
Modifizieren Sie das Programm, so daß es nicht Worte sondern a) Buchstaben bzw. b) Buchstabenpaare zählt. Vergleichen Sie deren Häufigkeitsverteilung sowohl zweier in der gleichen Sprache verfassten Texte als auch zweier in verschiedenen Sprachen abgefasster Texte. (2 Punkte)

Lösung:
Vergleicht man die Erscheinungshäufigkeit der Buchstaben in der deutschen Fassung von Kafkas Verwandlung mit der der englischen Fassung (Abb. 9), so fällt vor allem auf, dass in beiden Sprachen der Buchstabe e am häufigsten vorkommt, wenngleich im deutschen Text mit 13,58 % ein wenig öfter als im englischen (9,98 %).

Richtet man den Blick nun auf die Buchstabenpaare, so finden sich in beiden Sprachen die Paare *ll* und *ss* jeweils recht weit oben in den jeweiligen Listen (Abb. 10). Bei anderen Paaren, wie z. B. *ee* oder *nn* finden sich dagegen deutliche Unterschiede in den auftretenden Häufigkeiten (40 zu 393 bzw. 389 zu 33).

Bei der Häufigkeitsverteilung der Buchstaben sowie der Buchstabenpaare lässt sich in den logarithmierten Abbildungen 11 bis 16 ebenfalls eine fallende Gerade erkennen, was auf eine Befolgung des Zipfschen Gesetzes hindeutet.

Considered stopwords: False

frankenstein.txt				samsa.txt		
Total words: 74952				Total words: 22095		
Word	Abs.	Perc.		Word	Abs.	Perc.
the	4153	0.0		the	1146	0.0
and	2935	0.0		to	746	0.0
i	2720	0.0		and	626	0.0
of	2636	0.0		he	569	0.0
to	2084	0.0		his	550	0.0
my	1750	0.0		of	427	0.0
a	1382	0.0		was	397	0.0
in	1118	0.0		had	348	0.0
was	995	0.0		in	335	0.0
that	979	0.0		that	316	0.0
had	681	0.0		a	285	0.0
but	666	0.0		it	284	0.0
with	662	0.0		as	241	0.0
he	573	0.0		with	198	0.0
which	540	0.0		she	196	0.0
his	533	0.0		would	184	0.0
me	530	0.0		not	173	0.0
as	516	0.0		her	169	0.0
not	482	0.0		at	168	0.0
by	454	0.0		gregor	168	0.0
you	453	0.0		but	164	0.0
for	450	0.0		for	162	0.0
it	448	0.0		they	155	0.0
on	442	0.0		on	143	0.0
from	385	0.0		him	129	0.0
this	362	0.0		from	118	0.0
have	360	0.0		could	117	0.0
be	343	0.0		be	116	0.0
her	328	0.0		all	116	0.0
at	314	0.0		have	107	0.0

Abbildung 1: Liste der 30 am häufigsten vorkommenden Wörter in Frankenstein und Die Verwandlung

```
Considered stopwords: True
-----------------------------------------------------
frankenstein.txt        || samsa.txt
Total words: 74952      || Total words: 22095
-----------------------------------------------------
Word    Abs.    Perc.   ||      Word    Abs.    Perc.
-----------------------------------------------------
could   193     0.0     ||      would   184     0.0
one     190     0.0     ||      gregor  168     0.0
would   178     0.0     ||      could   117     0.0
me,     147     0.0     ||      gregor's        99      0.0
yet     138     0.0     ||      room    87      0.0
upon    125     0.0     ||      even    80      0.0
me.     107     0.0     ||      sister  77      0.0
might   107     0.0     ||      -       76      0.0
every   106     0.0     ||      back    74      0.0
shall   104     0.0     ||      father  72      0.0
first   102     0.0     ||      door    68      0.0
may     96      0.0     ||      mother  61      0.0
towards 94      0.0     ||      one     55      0.0
even    94      0.0     ||      way     52      0.0
saw     91      0.0     ||      time    46      0.0
found   77      0.0     ||      little  43      0.0
time    76      0.0     ||      get     43      0.0
man     76      0.0     ||      said    41      0.0
father  73      0.0     ||      made    40      0.0
felt    72      0.0     ||      still   39      0.0
"i      71      0.0     ||      without 38      0.0
said    68      0.0     ||      see     37      0.0
life    67      0.0     ||      chief   37      0.0
many    67      0.0     ||      first   36      0.0
made    66      0.0     ||      go      36      0.0
still   65      0.0     ||      much    36      0.0
dear    65      0.0     ||      like    34      0.0
thought 65      0.0     ||      head    33      0.0
soon    64      0.0     ||      quite   30      0.0
must    64      0.0     ||      came    29      0.0
-----------------------------------------------------
```

Abbildung 2: Liste der 30 am häufigsten vorkommenden Wörter in Frankenstein und Die Verwandlung unter Nichtbeachtung von Stopwords

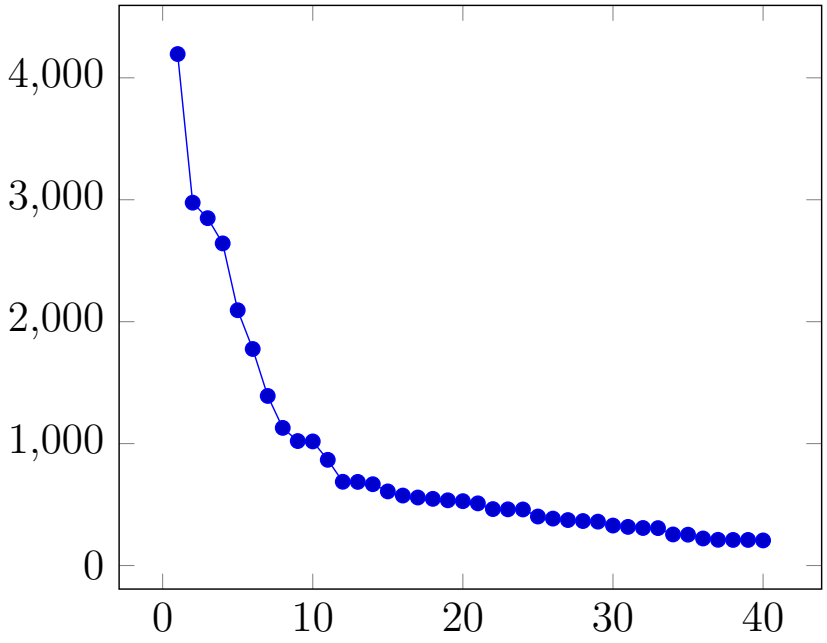

Abbildung 3: Die Häufigkeit (y-Achse) der 50 häufigsten Wörter in Frankenstein

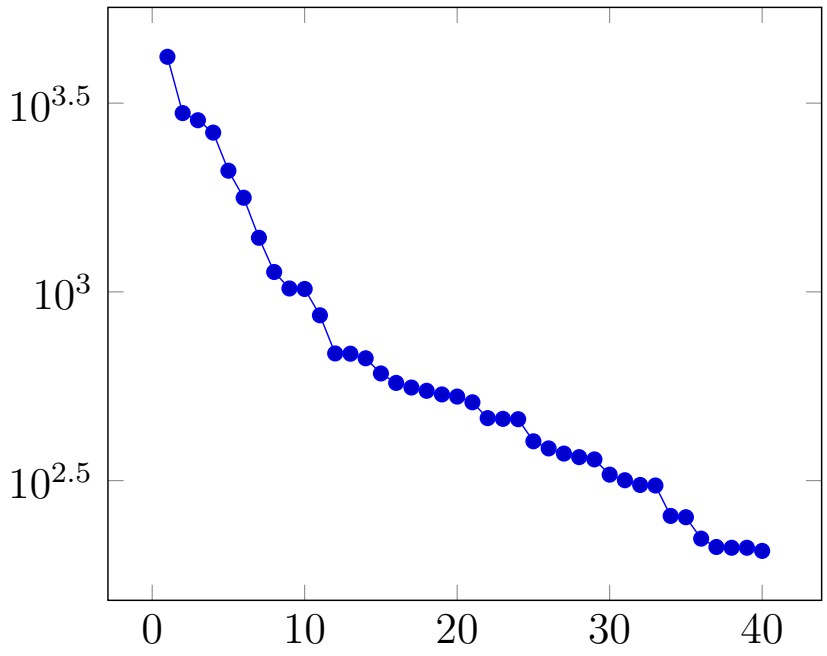

Abbildung 4: Die Häufigkeit (y-Achse) der 50 häufigsten Wörter in Frankenstein (logarithmiert)

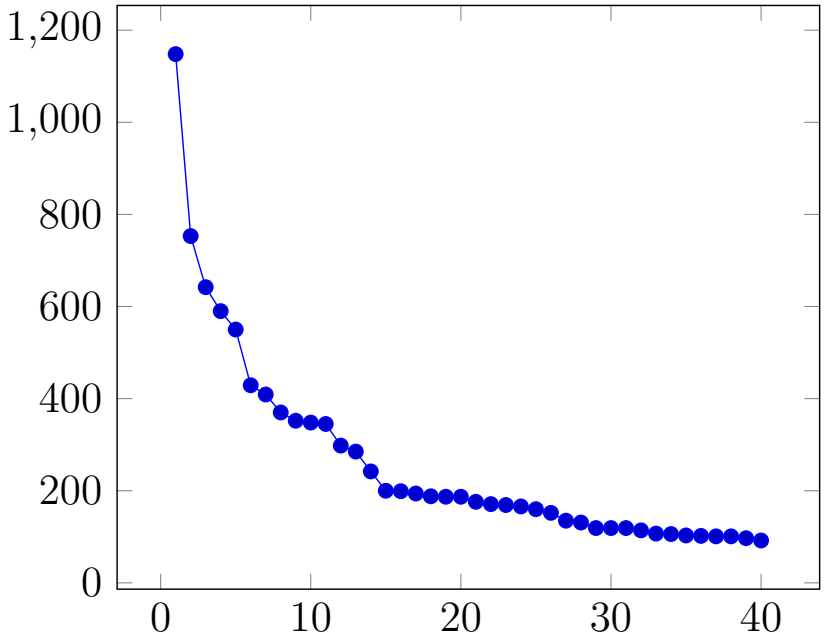

Abbildung 5: Die Häufigkeit (y-Achse) der 50 häufigsten Wörter in Die Verwandlung

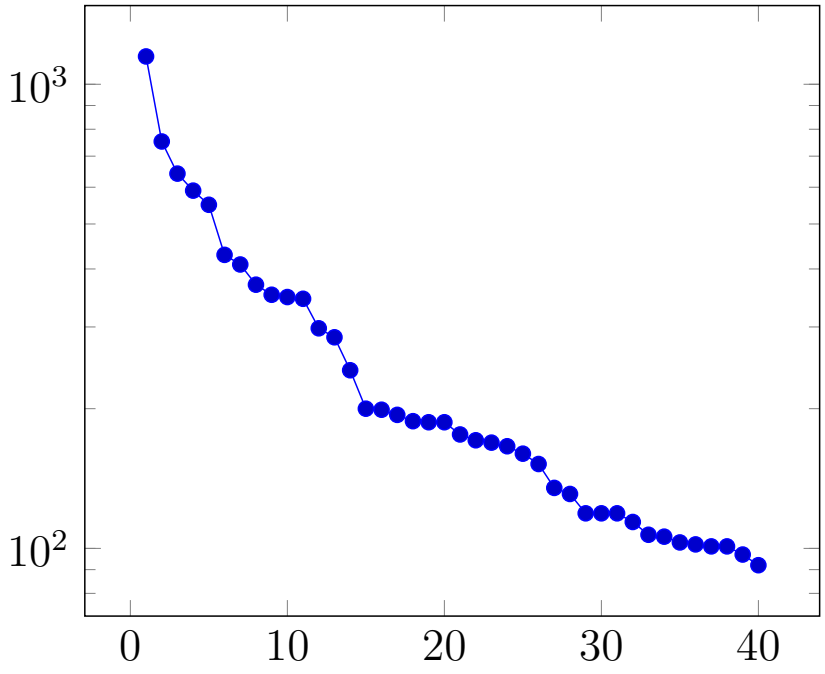

Abbildung 6: Die Häufigkeit (y-Achse) der 50 häufigsten Wörter in Die Verwandlung (logarithmiert)

n	Wort	$A \equiv$ Anzahl	$1/n$	$p(n) \equiv 1/(nH_N)$	$E \equiv p(n)N$	$1 - E/A$
1	the	4195.0	1.0	0.2337	7935.879	-0.8917
2	and	2976.0	0.5	0.1169	3967.9395	-0.3333
3	i	2849.0	0.3333	0.0779	2645.293	0.0715
4	of	2642.0	0.25	0.0584	1983.9698	0.2491
5	to	2094.0	0.2	0.0467	1587.1758	0.242
6	my	1776.0	0.1667	0.039	1322.6465	0.2553
7	a	1391.0	0.1429	0.0334	1133.697	0.185
8	in	1129.0	0.125	0.0292	991.9849	0.1214
9	was	1021.0	0.1111	0.026	881.7643	0.1364
10	that	1018.0	0.1	0.0234	793.5879	0.2204
11	me	867.0	0.0909	0.0212	721.4435	0.1679
12	but	687.0	0.0833	0.0195	661.3233	0.0374
13	had	686.0	0.0769	0.018	610.4522	0.1101
14	with	667.0	0.0714	0.0167	566.8485	0.1502
15	he	608.0	0.0667	0.0156	529.0586	0.1298
16	you	574.0	0.0625	0.0146	495.9924	0.1359
17	which	558.0	0.0588	0.0137	466.8164	0.1634
18	it	547.0	0.0556	0.013	440.8822	0.194
19	his	535.0	0.0526	0.0123	417.6778	0.2193
20	as	528.0	0.05	0.0117	396.794	0.2485
21	not	510.0	0.0476	0.0111	377.899	0.259
22	for	463.0	0.0455	0.0106	360.7218	0.2209
23	by	461.0	0.0435	0.0102	345.0382	0.2515
24	on	460.0	0.0417	0.0097	330.6616	0.2812
25	this	402.0	0.04	0.0093	317.4352	0.2104
26	from	385.0	0.0385	0.009	305.2261	0.2072
27	her	373.0	0.037	0.0087	293.9214	0.212
28	have	365.0	0.0357	0.0083	283.4243	0.2235
29	be	360.0	0.0345	0.0081	273.651	0.2399
30	when	328.0	0.0333	0.0078	264.5293	0.1935
31	at	317.0	0.0323	0.0075	255.9961	0.1924
32	were	308.0	0.0313	0.0073	247.9962	0.1948
33	is	307.0	0.0303	0.0071	240.4812	0.2167
34	she	255.0	0.0294	0.0069	233.4082	0.0847
35	your	253.0	0.0286	0.0067	226.7394	0.1038
36	him	222.0	0.0278	0.0065	220.4411	0.007
37	an	211.0	0.027	0.0063	214.4832	-0.0165
38	so	210.0	0.0263	0.0062	208.8389	0.0055
39	they	210.0	0.0256	0.006	203.4841	0.031
40	one	206.0	0.025	0.0058	198.397	0.0369

Abbildung 7: Untersuchung der Gültigkeit des Zipfschen Gesetzes für den Text „Frankenstein“

n	Wort	$A \equiv$ Anzahl	$1/n$	$p(n) \equiv 1/(nH_N)$	$E \equiv p(n)N$	$1 - E/A$
1	the	1148.0	1.0	0.2337	2443.8226	-1.1288
2	to	753.0	0.5	0.1169	1221.9113	-0.6227
3	and	642.0	0.3333	0.0779	814.6075	-0.2689
4	he	590.0	0.25	0.0584	610.9556	-0.0355
5	his	550.0	0.2	0.0467	488.7645	0.1113
6	of	429.0	0.1667	0.039	407.3038	0.0506
7	was	409.0	0.1429	0.0334	349.1175	0.1464
8	it	370.0	0.125	0.0292	305.4778	0.1744
9	had	352.0	0.1111	0.026	271.5358	0.2286
10	in	348.0	0.1	0.0234	244.3823	0.2978
11	that	345.0	0.0909	0.0212	222.1657	0.356
12	gregor	298.0	0.0833	0.0195	203.6519	0.3166
13	a	285.0	0.0769	0.018	187.9864	0.3404
14	as	242.0	0.0714	0.0167	174.5588	0.2787
15	she	200.0	0.0667	0.0156	162.9215	0.1854
16	with	199.0	0.0625	0.0146	152.7389	0.2325
17	s	194.0	0.0588	0.0137	143.7543	0.259
18	him	188.0	0.0556	0.013	135.7679	0.2778
19	would	187.0	0.0526	0.0123	128.6222	0.3122
20	her	187.0	0.05	0.0117	122.1911	0.3466
21	not	176.0	0.0476	0.0111	116.3725	0.3388
22	but	171.0	0.0455	0.0106	111.0828	0.3504
23	at	169.0	0.0435	0.0102	106.2532	0.3713
24	for	166.0	0.0417	0.0097	101.8259	0.3866
25	they	160.0	0.04	0.0093	97.7529	0.389
26	on	152.0	0.0385	0.009	93.9932	0.3816
27	all	135.0	0.037	0.0087	90.5119	0.3295
28	room	131.0	0.0357	0.0083	87.2794	0.3337
29	be	119.0	0.0345	0.0081	84.2697	0.2919
30	from	119.0	0.0333	0.0078	81.4608	0.3155
31	could	119.0	0.0323	0.0075	78.833	0.3375
32	out	114.0	0.0313	0.0073	76.3695	0.3301
33	have	107.0	0.0303	0.0071	74.0552	0.3079
34	there	106.0	0.0294	0.0069	71.8771	0.3219
35	if	103.0	0.0286	0.0067	69.8235	0.3221
36	father	102.0	0.0278	0.0065	67.884	0.3345
37	been	101.0	0.027	0.0063	66.0493	0.346
38	sister	101.0	0.0263	0.0062	64.3111	0.3633
39	so	97.0	0.0256	0.006	62.6621	0.354
40	this	92.0	0.025	0.0058	61.0956	0.3359

Abbildung 8: Untersuchung der Gültigkeit des Zipfschen Gesetzes für den Text „Die Verwandlung“

Charpairs considered: False

samsa_german.txt				samsa.txt		
Total chars: 124068				Total chars: 119232		
Char	Abs.	Perc.		Char	Abs.	Perc.
	17635	0.14214			20850	0.17487
e	16844	0.13576		e	11900	0.09981
n	9737	0.07848		t	9023	0.07568
r	7703	0.06209		o	7468	0.06263
i	7206	0.05808		h	7099	0.05954
s	5913	0.04766		a	7052	0.05915
t	5837	0.04705		i	5996	0.05029
a	5348	0.04311		n	5916	0.04962
h	5235	0.04219		s	5680	0.04764
d	4603	0.0371		r	5462	0.04581
u	3816	0.03076		d	4185	0.0351
l	3373	0.02719		l	3930	0.03296
g	3360	0.02708		u	2467	0.02069
c	3322	0.02678		w	2370	0.01988
m	2766	0.02229		g	2360	0.01979
o	2221	0.0179		m	2291	0.01921
ã	2211	0.01782		f	2103	0.01764
\n	1920	0.01548		\n	1956	0.0164
,	1868	0.01506		c	1923	0.01613
w	1755	0.01415		y	1628	0.01365
b	1711	0.01379		b	1380	0.01157
f	1653	0.01332		,	1293	0.01084
z	1385	0.01116		p	1262	0.01058
k	1210	0.00975		v	824	0.00691
v	824	0.00664		k	767	0.00643
↳	816	0.00658		.	737	0.00618
.	630	0.00508		'	331	0.00278
¤	581	0.00468		"	262	0.0022
p	485	0.00391		;	170	0.00143
ÿ	460	0.00371		-	115	0.00096

Abbildung 9: Liste der 30 am häufigsten vorkommenden Buchstaben in Die Verwandlung (deutsch und englisch)

Charpairs considered: True

samsa_german.txt				samsa.txt		
Total chars: 124068				Total chars: 119232		
Char	Abs.	Perc.		Char	Abs.	Perc.
tt	487	0.00393			595	0.00499
ll	436	0.00351		ll	525	0.0044
nn	389	0.00314		oo	493	0.00413
mm	316	0.00255		ee	393	0.0033
ss	264	0.00213		ss	192	0.00161
	150	0.00121		tt	140	0.00117
rr	127	0.00102		\n\n	110	0.00092
\n\n	126	0.00102		pp	96	0.00081
ff	97	0.00078		ff	88	0.00074
--	92	0.00074		rr	83	0.0007
ee	40	0.00032		mm	40	0.00034
pp	32	0.00026		dd	37	0.00031
aa	16	0.00013		nn	33	0.00028
gg	7	6e-05		cc	25	0.00021
kk	4	3e-05		gg	15	0.00013
uu	3	2e-05		bb	8	7e-05
ii	3	2e-05		ii	3	3e-05
hh	2	2e-05		..	2	2e-05
bb	2	2e-05		zz	1	1e-05
dd	1	1e-05		--	1	1e-05

Abbildung 10: Liste der 30 am häufigsten vorkommenden Buchstabenpaare in Die Verwandlung (deutsch und englisch)

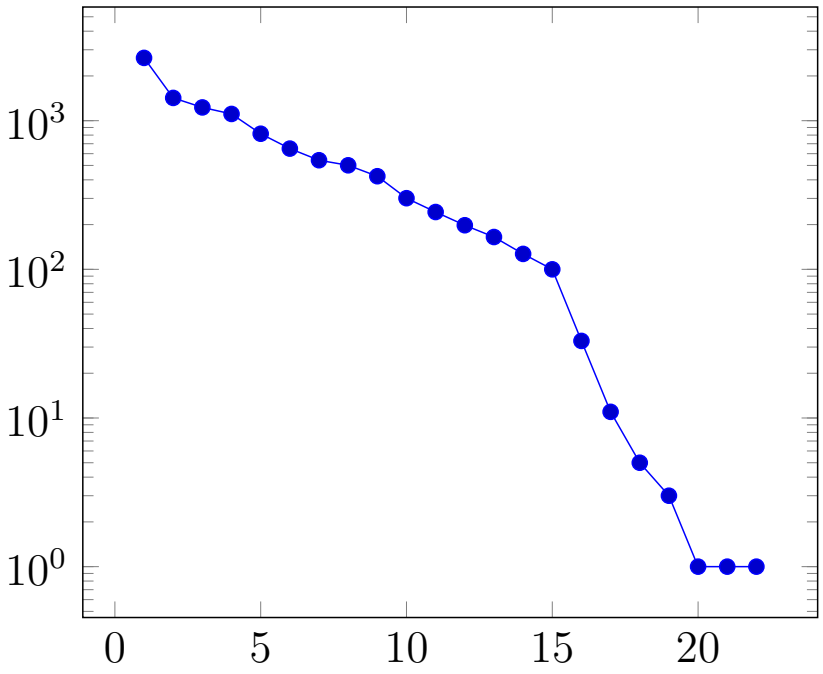

Abbildung 11: Die Häufigkeit (y-Achse) der 30 häufigsten Buchstabenpaare in Frankenstein (logarithmiert)

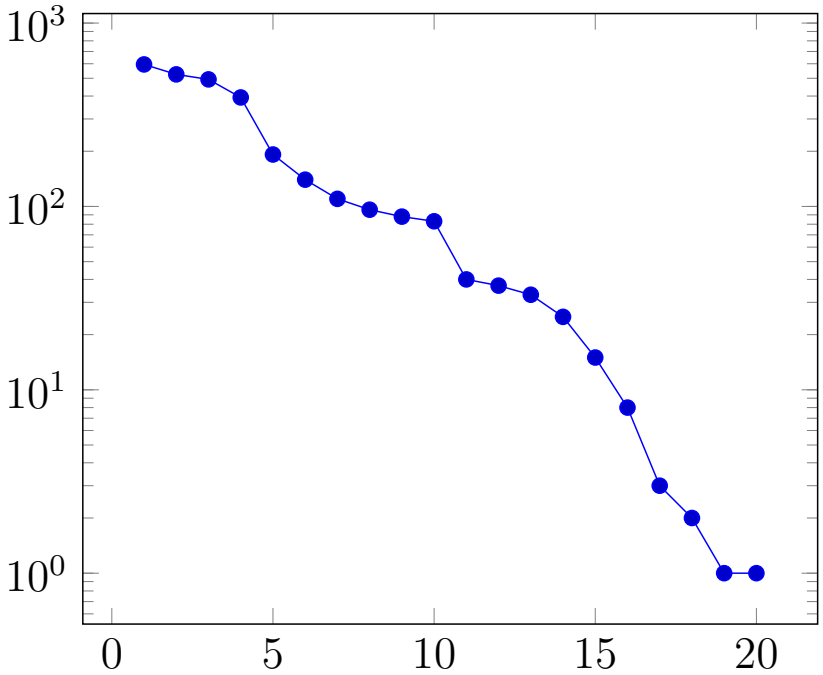

Abbildung 12: Die Häufigkeit (y-Achse) der 30 häufigsten Buchstabenpaare in Die Verwandlung (englisch) (logarithmiert)

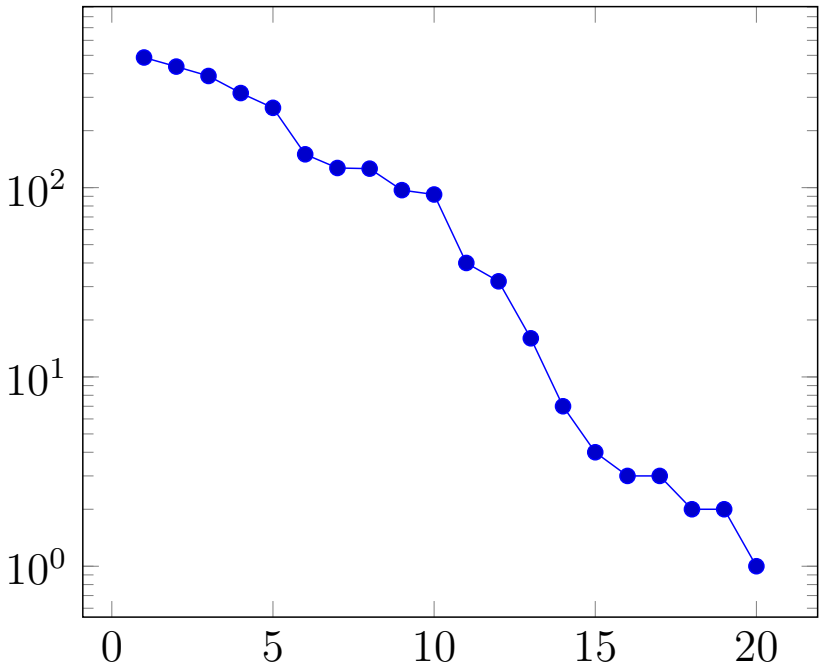

Abbildung 13: Die Häufigkeit (y-Achse) der 30 häufigsten Buchstabenpaare in Die Verwandlung (deutsch) (logarithmiert)

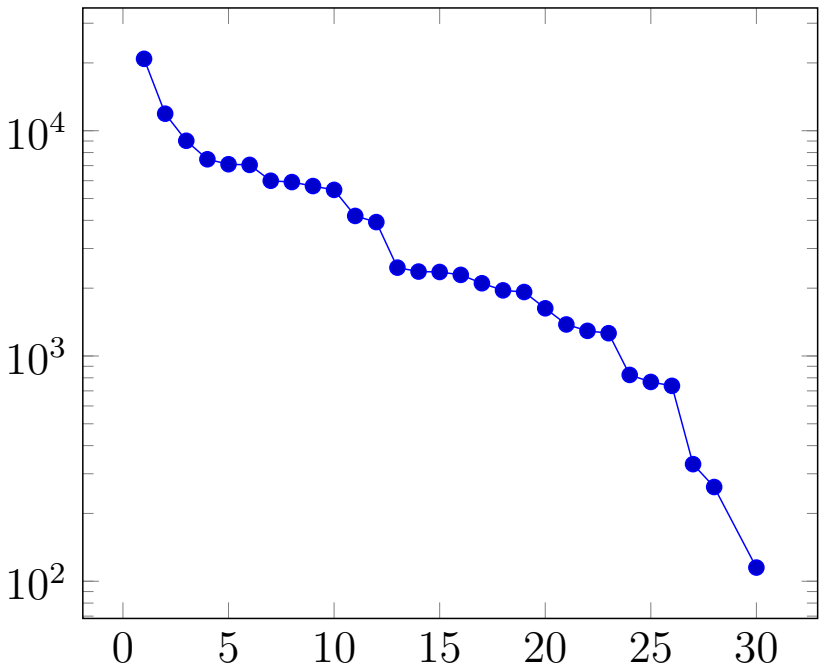

Abbildung 14: Die Häufigkeit (y-Achse) der 30 häufigsten Buchstaben in Die Verwandlung (englisch) (logarithmiert)

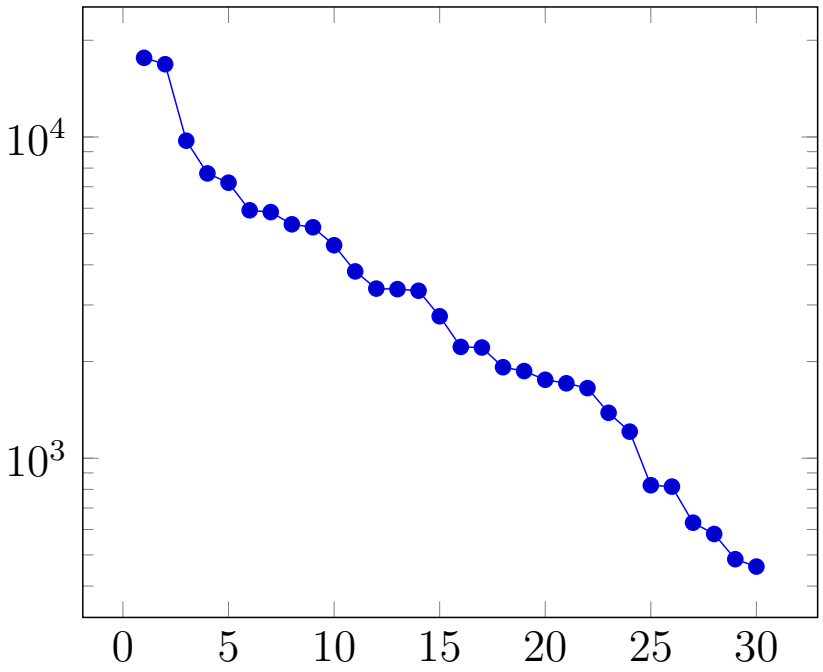

Abbildung 15: Die Häufigkeit (y-Achse) der 30 häufigsten Buchstabenpaare in Die Verwandlung (deutsch) (logarithmiert)

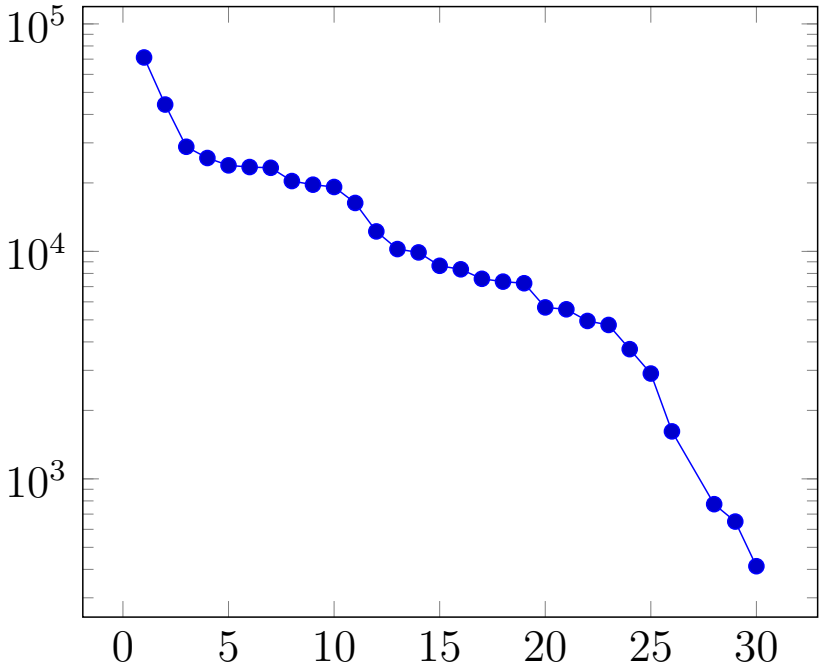

Abbildung 16: Die Häufigkeit (y-Achse) der 30 häufigsten Buchstaben in Frankenstein (logarithmiert)