

Web Mining im SoSe 2017 – Übung 1

Ingo Adrian und Steffen Pegenau
6. 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)

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 eine 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.

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.

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)

Considered stopwords: False						

frankenstein.txt			samsa.txt			
Total words: 77986			Total words: 25186			

Word	Abs.	Perc.		Word	Abs.	Perc.

the	4327	0.05548		the	1327	0.05269
and	3004	0.03852		to	822	0.03264
of	2754	0.03531		and	694	0.02755
i	2720	0.03488		he	571	0.02267
to	2160	0.0277		of	550	0.02184
my	1750	0.02244		his	550	0.02184
a	1438	0.01844		was	398	0.0158
in	1174	0.01505		in	392	0.01556
was	996	0.01277		had	348	0.01382
that	994	0.01275		a	342	0.01358
with	709	0.00909		that	331	0.01314
had	681	0.00873		it	296	0.01175
but	671	0.0086		as	250	0.00993
he	575	0.00737		with	246	0.00977
which	547	0.00701		she	196	0.00778
his	533	0.00683		not	194	0.0077
me	530	0.0068		for	189	0.0075
as	525	0.00673		at	184	0.00731
you	521	0.00668		would	184	0.00731
not	503	0.00645		her	169	0.00671
by	478	0.00613		but	169	0.00671
for	476	0.0061		gregor	168	0.00667
it	460	0.0059		they	158	0.00627
on	452	0.0058		on	155	0.00615
this	409	0.00524		be	135	0.00536
from	399	0.00512		all	134	0.00532
have	365	0.00468		this	133	0.00528
be	362	0.00464		from	132	0.00524
at	329	0.00422		him	129	0.00512
her	328	0.00421		if	119	0.00472

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

Considered stopwords: True

frankenstein.txt			samsa.txt			
Total words: 77986			Total words: 25186			
<hr/>						
Word	Abs.	Perc.		Word	Abs.	Perc.
<hr/>						
could	194	0.00249		would	184	0.00731
one	191	0.00245		gregor	168	0.00667
would	178	0.00228		could	118	0.00469
me,	147	0.00188		gregor's	99	0.00393
yet	138	0.00177		room	87	0.00345
upon	127	0.00163		project	83	0.0033
may	111	0.00142		-	83	0.0033
might	107	0.00137		even	82	0.00326
me.	107	0.00137		sister	77	0.00306
every	106	0.00136		back	74	0.00294
shall	106	0.00136		father	72	0.00286
first	102	0.00131		door	68	0.0027
even	96	0.00123		mother	61	0.00242
towards	94	0.00121		one	57	0.00226
saw	91	0.00117		way	54	0.00214
project	81	0.00104		work	54	0.00214
found	80	0.00103		gutenberg-tm	54	0.00214
time	76	0.00097		time	46	0.00183
man	76	0.00097		without	46	0.00183
must	73	0.00094		little	43	0.00171
father	73	0.00094		get	43	0.00171
felt	72	0.00092		said	41	0.00163
"i	71	0.00091		see	40	0.00159
many	69	0.00088		made	40	0.00159
said	68	0.00087		still	39	0.00155
life	67	0.00086		chief	38	0.00151
made	66	0.00085		first	37	0.00147
dear	65	0.00083		much	37	0.00147
still	65	0.00083		go	36	0.00143
thought	65	0.00083		like	34	0.00135

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