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.

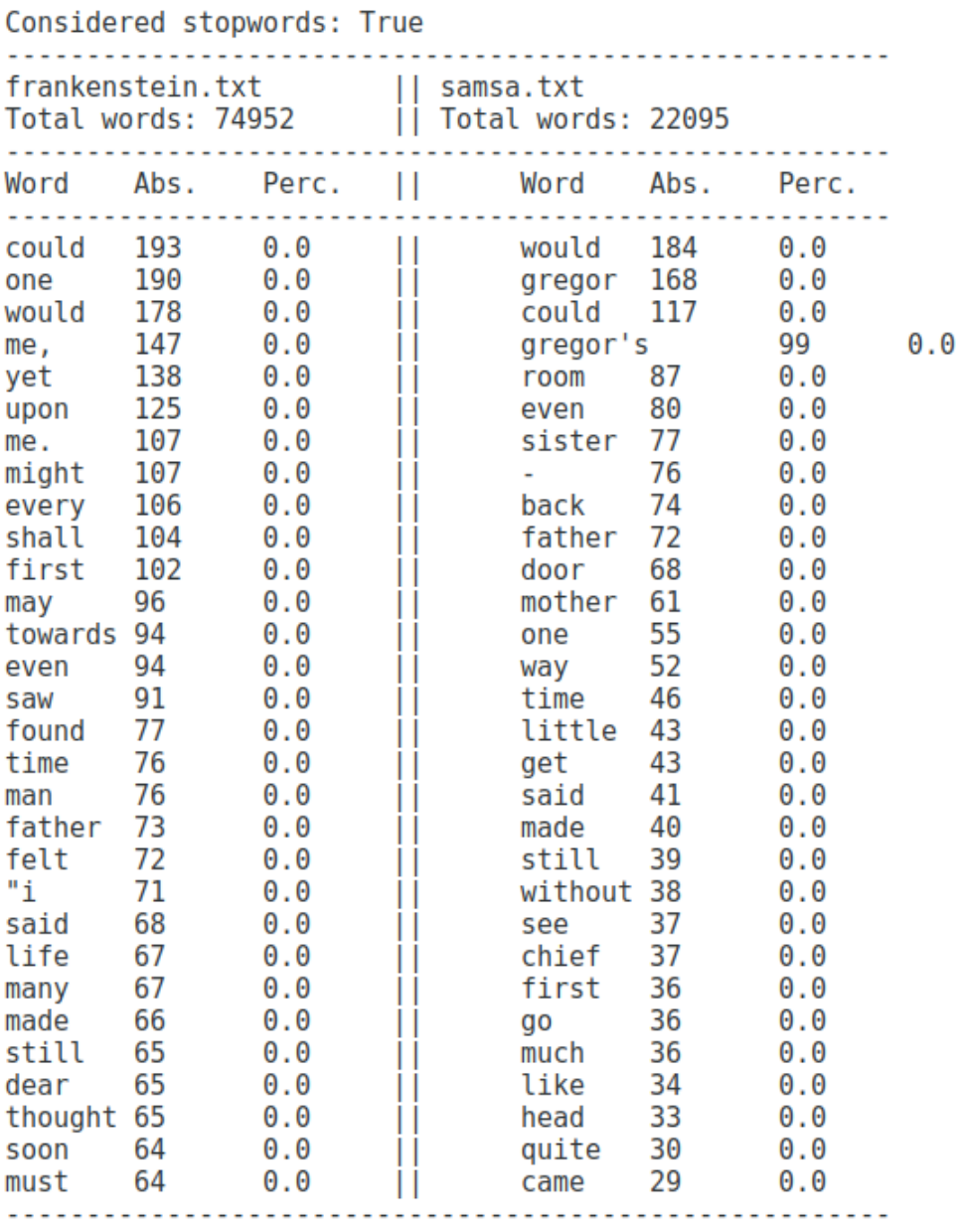
Aufgabe 4

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: 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