Textdatenbanken

Sommersemester 2020

3. Vorlesung

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Zerlegung des Textes in Teile

0. Schritt: Entfernen von HTML-Markup o.ä.
1. Schritt: Zerlegung des Textes in Sätze (s. letzte Vorlesung)
2. Schritt: Zerlegung des Textes in Wörter (heute)

Segmentierung in Wörter I

Naiver Ansatz:

- Im Deutschen sind die Wörter eines Textes durch Leerzeichen getrennt (anders als z.B. im Chinesischen).
- Also zerlegen wir einfach einen Satz an den Leerzeichen (und harten Zeilenumbrüchen) und erhalten die Wörter dieses Satzes.
- Eventuell müssen wir nach der Trennung bei den Leerzeichen noch Satzzeichen wie Punkt, Komma und Anführungszeichen entfernen.

Die Zwischenräume zwischen Wörtern werden auch als *white space* bezeichnet.

Segmentierung in Wörter II

Dieser Algorithmus verhält sich merkwürdig bei den folgenden Beispielen:

- Solche abgelegenen Airports haben Vor- und Nachteile.
- Die Prinz zu Hohenlohe-Jagstberg & Banghard Beratungs GmbH & Co Vermittlungs-KG, Sitz Berlin, ist als persönlich haftende Gesellschafterin eingetreten.
- Ab Sonntag, 20.1. präsentiert der PRO 7-Moderator Aiman Abdallah sein neues Format Galileo The Game, ein Quiz über Wissen, Logik und Strategie.

Denn:

- Im ersten Beispiel ist *Vor-* kein vollständiges Wort, sondern nur in der Fügung *Vor- und Nachteile* verständlich.
- Im zweiten Beispiel ist *Beratungs* kein wohlgeformtes Wort, sondern nur ein Namenbestandteil.
- Im dritten Beispiel schließlich ist *PRO 7-Moderator* ein „Wort“, welches ein Leerzeichen enthält. Bei einer weiteren Trennung an dem Leerzeichen erhielte man *7-Moderator*, was kein sinnvolles Wort ist.

Segmentierung in Wörter III

Trivialerweise erzeugen Fehler in den Texten fehlerhafte Wörter. Neben Rechtschreibfehlern wirken sich auch Worttrennungen aus, die ehemals am Zeilenende standen, aber bei der Konvertierung nicht rückgängig gemacht wurden. Dies passiert besonders bei Originaldokumenten im Format PDF:

- "Hier ist die Luft prickelnd wie Champagner", wirbt die Kurverwal- tung.

Struktur von Wörtern

Wann ist eine Zeichenkette ein (deutsches) Wort? Leider gibt es keine Informatik-nahe Definition der Linguisten.

1. Ansatz: Wir beschreiben die erlaubten Zeichen.
   - Also Buchstaben a-z (groß und klein), dazu ÄÖÜäöüß. Und der Bindestrich.
   - Was ist mit Ziffern (Audi A4), é (Café), Ø (Øre) usw.? Das ist Geschmackssache.
   - Auch Abweichungen von der natürlichen Reihenfolge von einem möglichen Großbuchstaben am Wortanfang und nachfolgenden Kleinbuchstaben sind möglich (pH-Wert).

Jede solche Beschreibung trifft auf Ausnahmen, die nicht erfasst werden.

2. Ansatz: Wir definieren Wortgrenzen. Alles, was zwischen zwei Wortgrenzen steht, ist ein Wort. Eventuell werden unschöne Wörter einer bestimmten Struktur als „Nicht-Wörter“ von der Bearbeitung ausgeschlossen.

Schwer zu separierender Text I

Die schwierig zu separierenden Stellen sind bunt hervorgehoben!

¿Cuál fue la mejor decada musical?

I simply can't owe you 30-40,000 bucks!

Dr. Gerhard Schröder (SPD), der 61jährige Ex-Kanzler, meinte am 18.12.2002 zum Ver.di-Vorstand: "Es ist vielleicht eine große Chance für die F.D.P.!"

Ihr Hotel an der Scater- und Fahrradstreckestrecke in Jüterbog mit 3,4-dihydrit(bi)methen hat bis 12.3. geöffnet...

Videre er det opprettet samarbeid med Logon on å kunne bruke grammatikken fra NorGram og XLE (http://www.ling.uib.no/~victoria/NorGram/).

Schwer zu separierender Text II

'Dies ist,ein Testsatz'.

Dies ist ein , Testsatz .Prof. Heyer und die ASV.

Dies ist ein , Testsatz.Dies ist ein zweiter ,der nicht erkannt worden ist.

Prof. Gruhn ist der Chef des Lehrstuhls für angew. Informatik.

Als Millionär hat man mehr als 1.000.000,00 Euro.

Er schaffte den neuen Weltrekord in einer Zeit von 12:34,567 Minuten und der langsamste benötigte eine Zeit von 1:12:34,567.

Mehr als 70 Prozent der Antwortenden (bisher knapp 1 200 Einsendungen) befürworten die Wiederauflage der «Großen Koalition». 15 Prozent sind für eine CDU-Minderheitsregierung, 5 Prozent für Rot-Grün und der Rest machte keine Angaben.

Schwer zu separierender Text III

```
Die BLZ ist 123 456 789.
Er belegte den 1. Platz.
Andrea's Auto bleibt stehen.
Er ist als Nummer 3 2 Runden zurück.
("TEST")
#ver.di#
Nietzsche formulierte das hymnisch: "Rechnen wir aus der
   Lyrik in Ton und Wort die Suggestion jenes...Fiebers
   ab: was bleibt von der Lyrik und Musik übrig?...Das
   virtuose Gequak kaltgestellter Frösche, die in ihrem
   Sumpf desperieren.
```

Wohlgeformte Wörter

Nicht alle so erhaltenen Zeichenketten sind wohlgeformte Wörter.

Kriterien für wohlgeformte Wörter:

- Bestehen aus lateinischen Buchstaben a-z und ä,ö,ü,ß,é in Groß- oder Kleinschreibung
- Können im Inneren einen oder mehr Bindestriche enthalten
- Dürfen bis zu zwei Ziffern enthalten, nicht am Anfang.

Nicht aufgenommen werden:

- Zahlen- und Datumsangaben
- Wörter mit Bindestrich, bei denen das entsprechende Wort ohne Bindestrich wesentlich häufiger ist (z.B. *Umlei-tung*)
- Wörter mit Punkten im Inneren (z.B. URLs; Ausnahme: *Ver.di*)
- Wörter mit Ziffern am Anfang (damit entfallen auch *3er*, *18jährige*)

Abgelehnte Wörter

Zurecht abgelehnt

- 90/Die
- z.B
- Bild"-Zeitung
- sind&ldquo
- Industrie-
- Don-ners-tag
- BoerseGo.de
- DE0005108401
- ?Das
- Ã¼ber

Wären akzeptabel gewesen

- Frankfurt/Main
- Dipl.-Ing.
- "Bild"-Zeitung
- S&P-500
- m^2
- afro-amerikanische

Wie findet man das Alphabet?

- Zeichen (ohne Unterscheidung von Groß-/Kleinschreibung) mit Ihrer Häufigkeit ermitteln
- Abbruch bei geeignetem Schwellwert

Beobachtung: Schwellwert 0.01% liefert genau die Zeichen für die ersten 10.000 Wörter (zumindest für deutsch, englisch und isländisch)

Character	Frequency in ‰ without repetition of words	with repetition of words
'	0.0	0.2
-	9.1	1.2
.	0.1	0.4
a	64.3	57.3
b	23.7	20.8
c	29.0	27.5
d	24.5	49.1
e	141.5	163.8
f	20.9	17.6
g	35.6	29.9
h	39.9	42.2
i	64.5	79.6
j	2.0	2.5
k	22.9	14.4
l	46.3	36.8
m	25.6	27.2
n	85.3	100.4
o	34.8	26.3
p	18.4	9.7
q	0.7	0.3
r	79.2	74.7
s	71.5	62.0
t	69.1	62.7
u	38.0	38.8
v	9.0	9.4
w	10.8	14.6
x	1.5	0.6
y	3.2	1.1
z	11.8	12.3
ß	1.8	1.9
ä	6.4	5.4
é	0.1	0.0
ö	2.8	2.7
ü	5.0	6.6

Speicherung in einer Datenbank

In einer relationalen Datenbank sollen die anfallenden Daten gespeichert werden:

- Sätze
- Wörter (mit Anzahl im Korpus)
- Inverse Liste zum Nachschlagen der Wörter
- Angaben zu Einzelwörtern: Grammatik, Sachgebiet, ...
- Wortpaare mit Angaben, z.B. Synonyme und Kookkurrenzen

Speicherung der Sätze

Tabelle **`sentences`** mit **`s_id, sentence, source`**

`s_id`: Id des Satzes, fortlaufend nummeriert
`sentence`: Der Volltext des Satzes als String.
`source`: (Verweis auf) die Quelle (URL) des Satzes

Index auf: **`s_id`**
Index auf Sätze?

Evtl. weitere Angaben wie z.B. Sprache

Mögliche Indexe auf Sätze

Typischerweise unterscheiden sich aufeinanderfolgende Sätze in einer alphabetischen Satzliste nach wenigen Wörtern.

Möglichkeiten:

- Index auf left(sentence, 32). Erlaubt schnelle Suche nach allen Sätzen mit gegebenem Satzanfang. Suche mit Wildcard am Anfang verlangt komplette sequenzielle Suche.
- Volltextindex (ähnlich Suchmaschinen). Aufwändige Indexerstellung bei MySQL, merkwürdiges Ranking, Probleme mit Stoppwörtern.
- Index mit Hashing (z.B. md5). Erlaubt nur Suche nach vollständigen Sätzen, extrem schnell. Sinnvoll, um Dubletten zu verhindern.

Speicherung der Wörter

Tabelle **`words`** mit **`w_id, word, freq`**

`w_id`: Id des Wortes, Nummerierung s. nächste Folie
`word`: Das Wort als String.
`freq`: Häufigkeit des Wortes, absolute Frequenz im Korpus

Index auf: **`w_id, word`**

Nummerierung der Wörter I

Problem: Wir finden ständig mehr Wörter, die Liste muss erweiterbar sein.

1. Möglichkeit: Alphabetisch. Schlecht wegen nicht vorhandener Erweiterbarkeit.
2. Möglichkeit: In der Reihenfolge ihres ersten Auftretens. Möglich, aber keinerlei sinnvolles Kriterium. Nachteil: Gleiche Wörter bekommen in verschiedenen Korpora verschiedene Wortnummern.
3. Möglichkeit: Nach Frequenz absteigend sortiert. Bei Erweiterung bleibt diese Eigenschaft wenigstens näherungsweise erhalten.
   Praktisch: Wenn nur mit hochfrequenten Wörtern gearbeitet werden soll, lässt sich dies über eine Einschränkung bei den Wortnummern (z.B. <1000) erreichen.

Nummerierung der Wörter II

Unser Vorgehen:

- Wortnummern <100 werden für Sonderzeichen reserviert und einmalig fest (für alle Sprachen!) vergeben.
- Für eine Sprache wird ein möglichst großes Korpus beschafft. Die Wörter daraus werden nach Häufigkeit sortiert und erhalten Wortnummern ab 101. Diese Nummern sind fest ab diesem Moment.
- Bei späteren Erweiterungen werden für neue Wörter nachfolgende größere Wortnummern vergeben. Die Wortliste bleibt näherungsweise sortiert nach Häufigkeit (für w_id>100)

Die Wortliste

Die Sonderzeichen

w_id	word	freq
1	!	6587
2	"	258528
3	#	89
4	$	60
5	%	868
6	&	1806
7	'	30390
8	(	57595
9	)	57241
10	*	0
11	+	369
12	,	871378
13	-	49692
14	.	1030519
15	/	2252
16	:	88111
17	;	8308
18	<	32
19	=	113
20	>	37
21	?	27494
22	@	5
23	[	173
24	\	4
25	]	179
...	..	...

Die häufigsten Wörter sind

w_id	word	freq
101	der	495584
102	die	467926
103	und	310533
104	in	262491
105	den	191241
106	von	149713
107	zu	140260
108	das	135021
109	mit	129344
110	sich	121794
111	ist	119115
112	auf	116042
113	im	112622
114	nicht	111996
115	für	110971
116	Die	110476
117	des	103899
118	dem	101168
119	ein	97642
120	eine	86336
121	es	78517
122	als	75285
123	auch	73717
124	an	72307
125	hat	63207
...	...	...

Längste Wörter unter den top-1M -ohne Bindestrich-

Finanzmarktstabilisierungsergänzungsgesetz	42
Frequenzbereichszuweisungsplanverordnung	40
Telekommunikationsüberwachungsverordnung	40
Schwangerschaftskonfliktberatungsstellen	40
Vermögensschadenhaftpflichtversicherung	39
Unternehmensbeteiligungsgesellschaften	38
Gemeindeverkehrsfinanzierungsgesetzes	37
Verkehrsordnungswidrigkeitenverfahren	37
Aufstiegsfortbildungsförderungsgesetz	37
Beschäftigungssicherungsvereinbarung	36

Längste Wörter unter den top-1M -mit Bindestrich-

mathematisch-naturwissenschaftlich-technischen	46
Aufmerksamkeitsdefizit-Hyperaktivitätsstörung	45
Hohensaaten-Friedrichsthaler-Wasserstraße	41
Telekommunikations-Überwachungsverordnung	41
Telekommunikations-Kundenschutzverordnung	41
Schleswig-Holstein-Sonderburg-Glücksburg	40
Europameisterschafts-Qualifikationsspiel	40
Vermögensschaden-Haftpflichtversicherung	40
Gewässerschaden-Haftpflichtversicherung	39
Einkommensteuer-Durchführungsverordnung	39

Wortgruppen

In die Wortliste aufgenommen werden auch Wortgruppen bestehend aus zwei oder mehr Wörtern, die dann in dieser Reihenfolge vorkommen müssen.

Wie werden die Anzahlen der Wörter aus Wortgruppen gezählt? Wir zählen zunächst die Einzelwörter ohne Berücksichtigung der Wortgruppen, danach noch einmal nur die Wortgruppen. Wörter in Wortgruppen werden also mehrfach gezählt.
Vorteil: Das Hinzufügen oder Weglassen von Wortgruppen ändert die Anzahlen der Einzelwörter nicht!

Die häufigsten Wortgruppen

w_id	word	freq
290	vor allem	8017
604	zum Beispiel	2644
650	unter anderem	2544
668	immer wieder	2405
853	Vor allem	1849
890	am Ende	1753
923	New York	1681
1157	Ende des	1323
1182	Gerhard Schröder	1306
1205	am Wochenende	1275
1245	DIE WELT	1240
1326	Jahre alt	1200
1371	in Höhe von	1071
1399	nach wie vor	1068
1430	immer mehr	1066
1596	ersten Mal	981
1524	in der Nacht	957
1622	kurz vor	950
1640	im Vergleich	937
1627	zu Hause	922
1617	so viel	919
1669	zwei Wochen	896
1751	dieses Jahres	885
1725	im Osten	880
1752	im Sommer	842
...	...	...

Sinnvolle Wortgruppen

Was sind sinnvolle Wortgruppen? Linguisten kennen

- Eigennamen (*Steffi Graf*, *Haus des Buches*)
- Wendungen (*vom Zaun brechen*, *hin und wieder*)

Viele dieser Wortgruppen können flektiert werden:

Steffi Grafs, *vom Zaun gebrochen*

Manche Wendungen sind diskontinuierlich (das können wir momentan nicht speichern)

brach einen Streit vom Zaun,

Quellen für Wortgruppen

Handarbeit

- Listen aus Wörterbüchern, z.B. Suchformen für Phraseologismen (z.B. „Unmögliche möglich“ für „das Unmögliche möglich machen/gemacht“); im Deutschen Suchformen für 5000 Phraseologismen

Crowdsourcing

- Wikipedia-Artikeltitel bestehend aus mehreren Wörtern
  - Prominente Personen u.a. Eigennamen, Fachterminologie, ...
  - In allen Sprachen
  - Aktuell

Automatisch erzeugt

- Häufiges Auftreten in dieser Form
- Bestimmte Wortartenmuster (nicht „hatte gestern in“)
- Qualität teilweise unbefriedigend

Häufige Wortgruppen der Längen >=3 und >=4

w_id	word	freq
1371	in Höhe von	1071
1399	nach wie vor	1068
1524	in der Nacht	957
2039	in der Nähe	698
2191	in der Stadt	664
2435	in der Lage	619
2416	Bundeskanzler Gerhard Schröder	610
2632	in der Regel	583
2887	in erster Linie	521
3098	George W. Bush	507
3233	im Alter von	495
3493	im Zusammenhang mit	440
3489	auf jeden Fall	431
3512	auf den Markt	421
3262	vor einem Jahr	420
3712	auf der Straße	413
3873	in den letzten Jahren	393
3819	in der Hand	390
3733	von Anfang an	386
3912	so gut wie	376
4318	in diesen Tagen	368
4233	im Vergleich zu	360
4267	mit Blick auf	329
4422	auf den Weg	321
4787	in der Tat	292
4856	auf diese Weise	288

w_id	word	freq
3873	in den letzten Jahren	393
5428	auf der Suche nach	249
6612	US-Präsident George W. Bush	225
6701	auf den ersten Blick	223
6980	auf der anderen Seite	212
7099	in den nächsten Tagen	206
6744	in der Nähe von	195
11793	Gesellschaft mit beschränkter Haftung	112
11765	Tag der offenen Tür	106
14301	unter Dach und Fach	89
13756	Papst Johannes Paul II	82
14991	in Ost und West	79
15382	auf den Markt kommen	79
15284	Präsident George W. Bush	74
15467	Auf den ersten Blick	73
16766	eine Frage der Zeit	68
17205	in der Lage sein	64
20085	unter der Leitung von	64
20565	mit von der Partie	60
20828	für die kommende Saison	59
24356	liegt auf der Hand	56
19456	die Art und Weise	54
19982	Bürgermeister Ole von Beust	53
19410	Regierende Bürgermeister Eberhard Diepgen	53
22193	ein Dorn im Auge	49
22809	für den Fall, dass	49

Häufige Wortgruppen der Länge >=5

w_id	word	freq
27162	im wahrsten Sinne des Wortes	49
27928	in den letzten zehn Jahren	47
28266	Bund für Umwelt und Naturschutz	40
30181	Haus der Kulturen der Welt	35
29629	Schritt in die richtige Richtung	33
32403	Regulierungsbehörde für Telekommunikation und Post	32
31105	alle Hände voll zu tun	31
35099	einen Strich durch die Rechnung	27
35860	mit dem Rücken zur Wand	26
35026	Organisation für Sicherheit und Zusammenarbeit in Europa	26
44940	Gewerkschaft Öffentliche Dienste, Transport und Verkehr	25
39847	nicht von heute auf morgen	25
43080	Gewerkschaft Handel, Banken und Versicherungen	24
40261	Trennung von Amt und Mandat	24
49968	Tropfen auf den heißen Stein	22
49411	Berlins Regierender Bürgermeister Klaus Wowereit	21
49845	für sich in Anspruch nehmen	20
57404	tief in die Tasche greifen	19
54060	kein Blatt vor den Mund	18
52240	auf der Höhe der Zeit	17
49307	auf dem Weg der Besserung	15
50926	Bund für Umwelt und Naturschutz Deutschland	15
48526	an allen Ecken und Enden	15
62370	ein Tropfen auf den heißen Stein	14
71433	ein Schritt in die richtige Richtung	14
57109	die Klinke in die Hand	14

Häufige Wortgruppen der Länge >=6

w_id	word	freq
35026	Organisation für Sicherheit und Zusammenarbeit in Europa	26
44940	Gewerkschaft Öffentliche Dienste, Transport und Verkehr	25
50926	Bund für Umwelt und Naturschutz Deutschland	15
62370	ein Tropfen auf den heißen Stein	14
71433	ein Schritt in die richtige Richtung	14
60458	von einem Tag auf den anderen	12
92220	mit Rat und Tat zur Seite	10
69276	nicht von der Hand zu weisen	10
87371	nur ein Tropfen auf den heißen Stein	9
124080	mit einem lachenden und einem weinenden Auge	8
104939	Zentralstelle für die Vergabe von Studienplätzen	8
109577	Bundesamt für Sicherheit in der Informationstechnik	8
107766	liegt in der Natur der Sache	8
112747	Denkmal für die ermordeten Juden Europas	8
129293	von der Hand in den Mund	8
124022	mit dem Kopf durch die Wand	6
102324	Bundesamt für die Anerkennung ausländischer Flüchtlinge	6
170557	Deutschen Forschungsanstalt für Luft- und Raumfahrt	5
113378	ohne mit der Wimper zu zucken	5
149485	Proceedings of the National Academy of Sciences	5
137921	Kommission zur Ermittlung des Finanzbedarfs der Rundfunkanstalten	5
156302	alle Hände voll zu tun haben	5
174998	Grün Berlin Park und Garten GmbH	5
115145	Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung	5
200401	Bundesverband der deutschen Gas- und Wasserwirtschaft	4
250511	Jim Knopf und Lukas der Lokomotivführer	4