mCLOUD: Beschreibungstexte

Adrian Wilke, Marleen Wagner

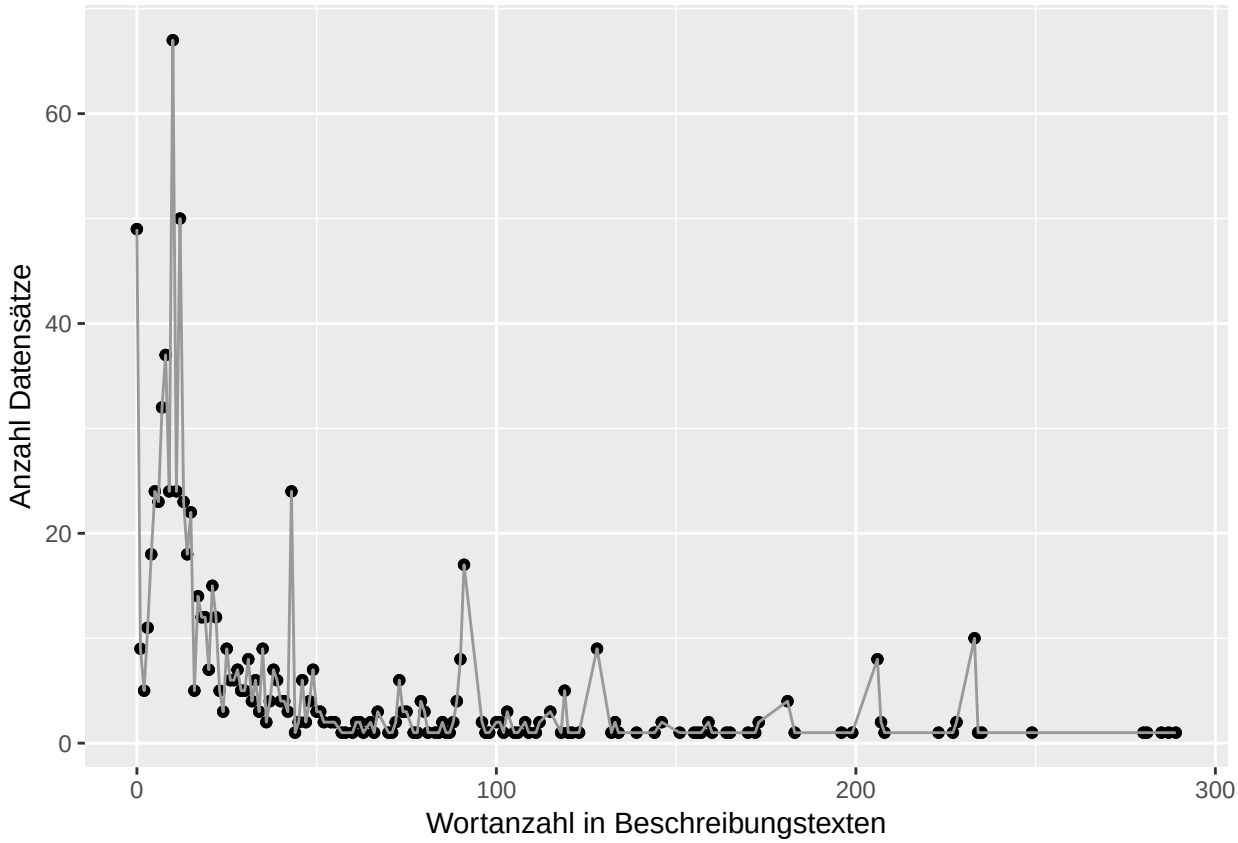
2018-03-28

Auf der Plattform mCLOUD stellt das Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) offene Daten aus dem Bereich Mobilität zur Verfügung. Die verfügbaren Datensätze werden von verschiedenen Anbietern bereitgestellt und sind damit zunächst voneinander unabhängig. Zur Suche relevanter Daten bietet mCloud die Möglichkeit zur Filterung nach Kategorien, Datenanbietern, Lizenzen und Datenformaten an. Eine Möglichkeit zur Verbesserung einer solchen Suchfunktion ist die Aufbereitung und Vernetzung der Metadaten.

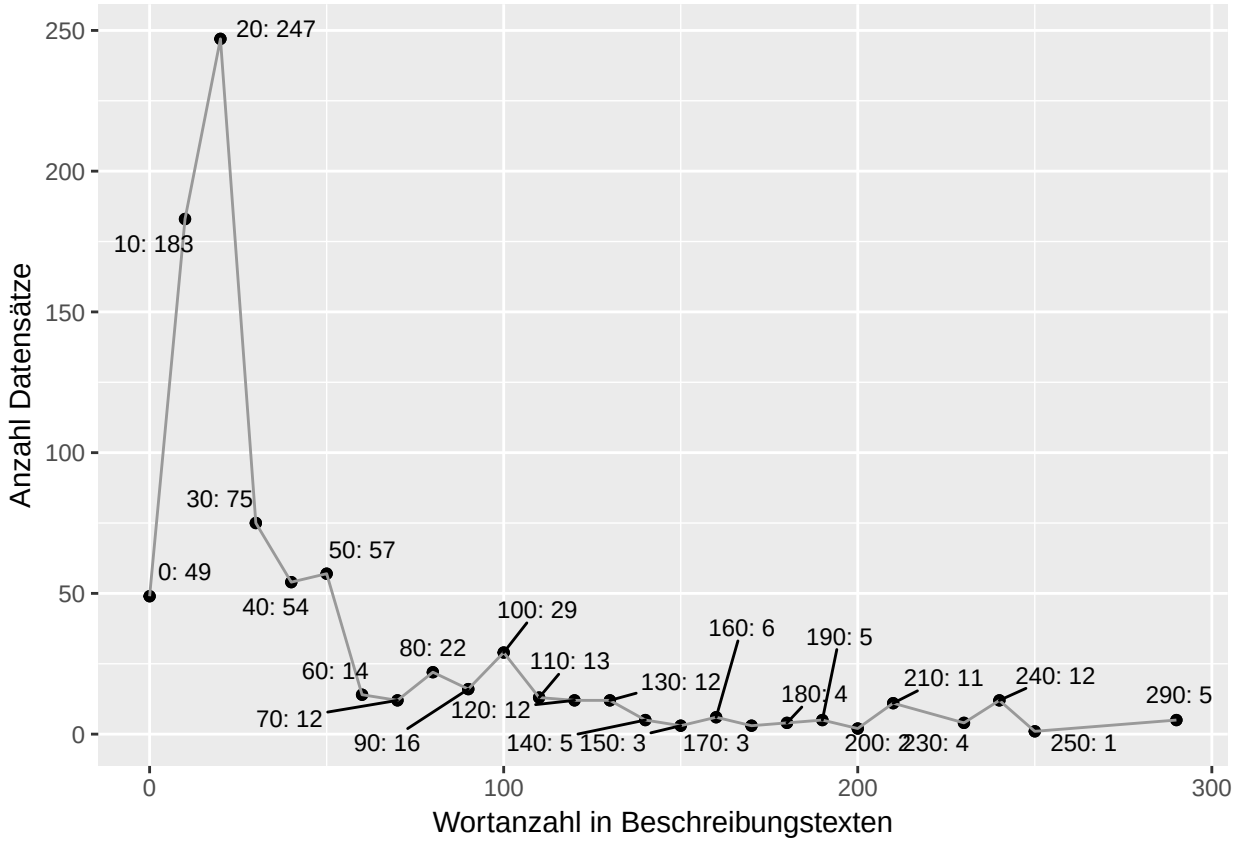
Das Projekt Open Data Portal Germany (OPAL) verfolgt das Ziel der Veredlung und Verknüpfung von Metadaten, um die Auffindbarkeit von Datensätzen zu verbessern. Dabei stehen öffentliche Datenquellen aus Deutschland im Mittelpunkt. Eine Besonderheit des Projekts ist die Fokussierung auf Metadaten. Zur Analyse werden nicht die eigentlichen Datensätze, sondern deren Beschreibungen verwendet.

Eine Ansatz zur Gewinnung von Informationen aus Beschreibungen ist die Extraktion verlinkter, semantischer Daten aus unstrukturierten, natürlichsprachlichen Texten. Hierzu können bestehende Lösungen weiterentwickelt werden, z.B. AGDISTIS, FOX oder REX. Hier kommt die Frage auf, welche Daten dazu genutzt werden können.

Offene Daten, die auf mCLOUD bereitgestellt werden, sind überwiegend mit Beschreibungstexten versehen. Um eine erste Übersicht der verfügbaren Daten zu erhalten, haben wir einen pragmatisches Vorgehen gewählt. Es wurden 856 Beschreibungstexte heruntergeladen. Jeder Text wurde durch Leerzeichen getrennt, so dass sich eine Annäherung der Anzahl verwendeter Wörter ergibt. Eine Übersicht zeigt die folgende Grafik.



Am häufigsten werden Beschreibungstexte mit 10 Wörtern verwendet. Zur besseren Unterscheidung der Punkte wurden die Häufigkeiten aggregiert. Auf der folgenden Grafik sind die Häufigkeiten nach einer Zusammenfassung der Größe 10 dargestellt. Es stehen z.B. 183 Datensätze zur Verfügung, deren Wortanzahl zwischen 1 und 10 liegt. Zwischen 251 und 260 Wörtern wurde kein Datensatz gefunden.



Zusammenfassend stellt mCLOUD aus dieser quantitativen Sichtweise (ohne die Semantik der Daten zu betrachten) eine erste Grundlage zur Analyse unstrukturierter Daten zur Verfügung. Für rund 89 Prozent der Datensätze wird ein Beschreibungstext von mindestens 5 Wörtern bereitgestellt. Rund 6 Prozent der Datensätze sind nicht mit Beschreibungen versehen. In 9 Fällen besteht die Beschreibung aus einem Wort; häufig ist dies der Name einer Applikation. Die umfangreichste Beschreibung mit 289 Wörtern ist der Datensatz “Parkdaten Stadt Düsseldorf”.

Im Web finden Sie auch zukünftig Neuigkeiten zu OPAL.

Auf Twitter informieren @DiceResearch und @akswgroup über Neuigkeiten in den Bereichen Data Science und Semantic Web.

Wenn Ihnen dieser Artikel gefallen hat, empfehlen Sie ihn auf Twitter.

Table 1: Wortanzahl in Beschreibungstexten

Von Wortanzahl	Bis Wortanzahl	Anzahl Datensätze	Prozent
0	0	49	5.72
1	10	183	21.38
11	20	247	28.86
21	30	75	8.76
31	40	54	6.31
41	50	57	6.66
51	60	14	1.64

Von Wortanzahl	Bis Wortanzahl	Anzahl Datensätze	Prozent
61	70	12	1.40
71	80	22	2.57
81	90	16	1.87
91	100	29	3.39
101	110	13	1.52
111	120	12	1.40
121	130	12	1.40
131	140	5	0.58
141	150	3	0.35
151	160	6	0.70
161	170	3	0.35
171	180	4	0.47
181	190	5	0.58
191	200	2	0.23
201	210	11	1.29
221	230	4	0.47
231	240	12	1.40
241	250	1	0.12
281	290	5	0.58

Table 2: Wortanzahl in Beschreibungstexten

Wortanzahl	Anzahl Datensätze
0	49
1	9
2	5
3	11
4	18
5	24
6	23
7	32
8	37
9	24
10	67
11	24
12	50
13	23
14	18
15	22
16	5
17	14
18	12
19	12
20	7
21	15
22	12
23	5
24	3
25	9
26	6
27	6

Wortanzahl	Anzahl Datensätze
28	7
29	5
30	5
31	8
32	4
33	6
34	3
35	9
36	2
37	4
38	7
39	6
40	4
41	4
42	3
43	24
44	1
45	2
46	6
47	2
48	4
49	7
50	3
51	3
52	2
54	2
55	2
57	1
58	1
60	1
61	2
62	2
63	1
65	2
66	1
67	3
70	1
71	1
72	2
73	6
74	3
75	3
77	1
78	1
79	4
80	3
81	1
83	1
84	1
85	2
86	1
87	1

Wortanzahl	Anzahl Datensätze
88	2
89	4
90	8
91	17
96	2
97	1
98	1
100	2
101	2
102	1
103	3
105	1
106	1
108	2
109	1
111	1
112	2
115	3
118	1
119	5
120	1
121	1
123	1
128	9
132	1
133	2
134	1
139	1
144	1
146	2
151	1
155	1
156	1
157	1
159	2
160	1
164	1
165	1
170	1
172	1
173	2
181	4
183	1
196	1
199	1
206	8
207	2
208	1
223	1
227	1
228	2
233	10

Wortanzahl	Anzahl Datensätze
234	1
235	1
249	1
280	1
281	1
285	1
287	1
289	1

Alle Angaben ohne Gewähr.