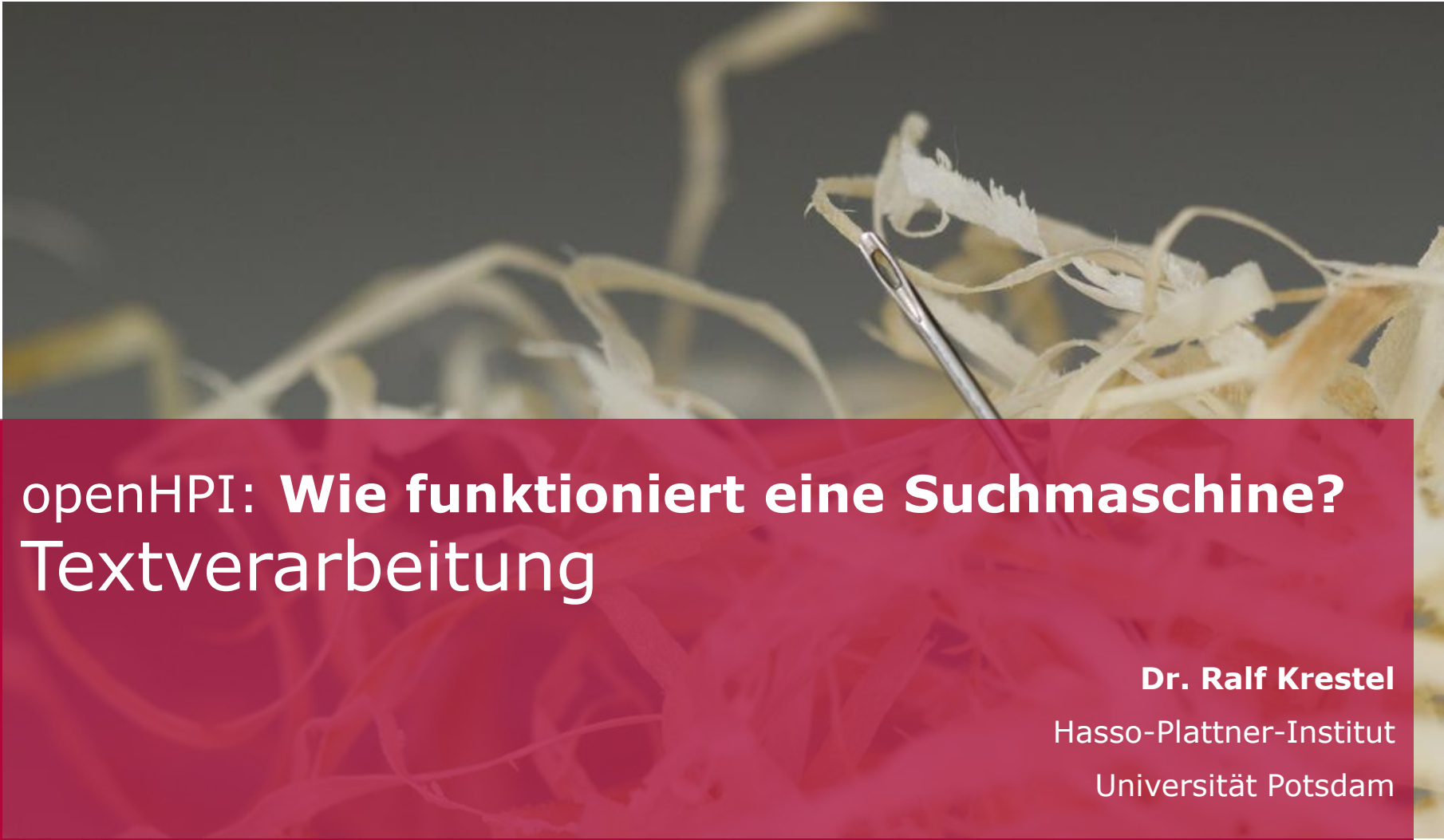




openHPI: **Wie funktioniert eine Suchmaschine?** Textverarbeitung

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

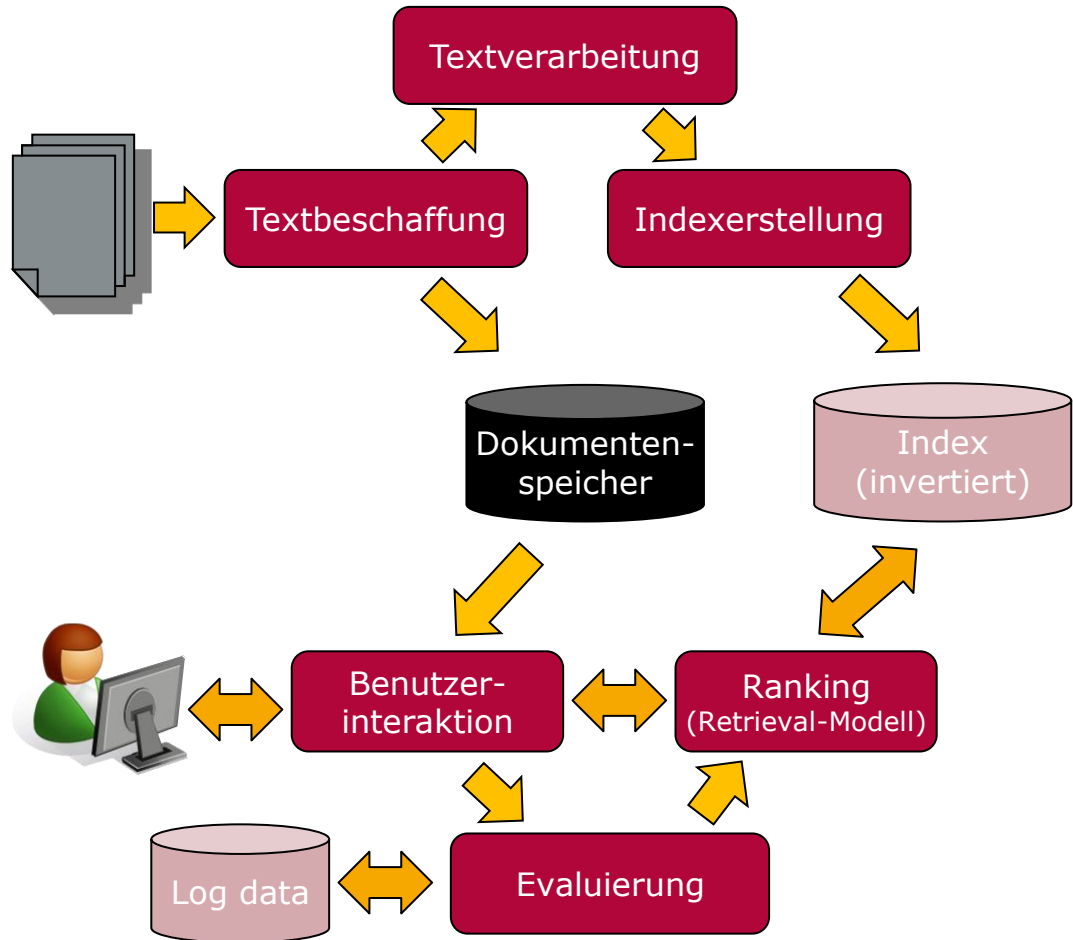
■ Grundlegende Bausteine

□ Indexierung

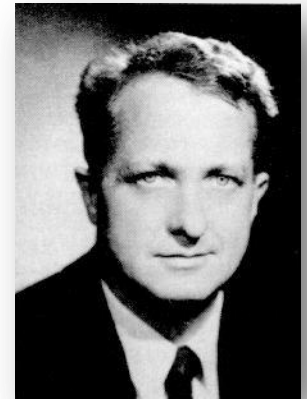
- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

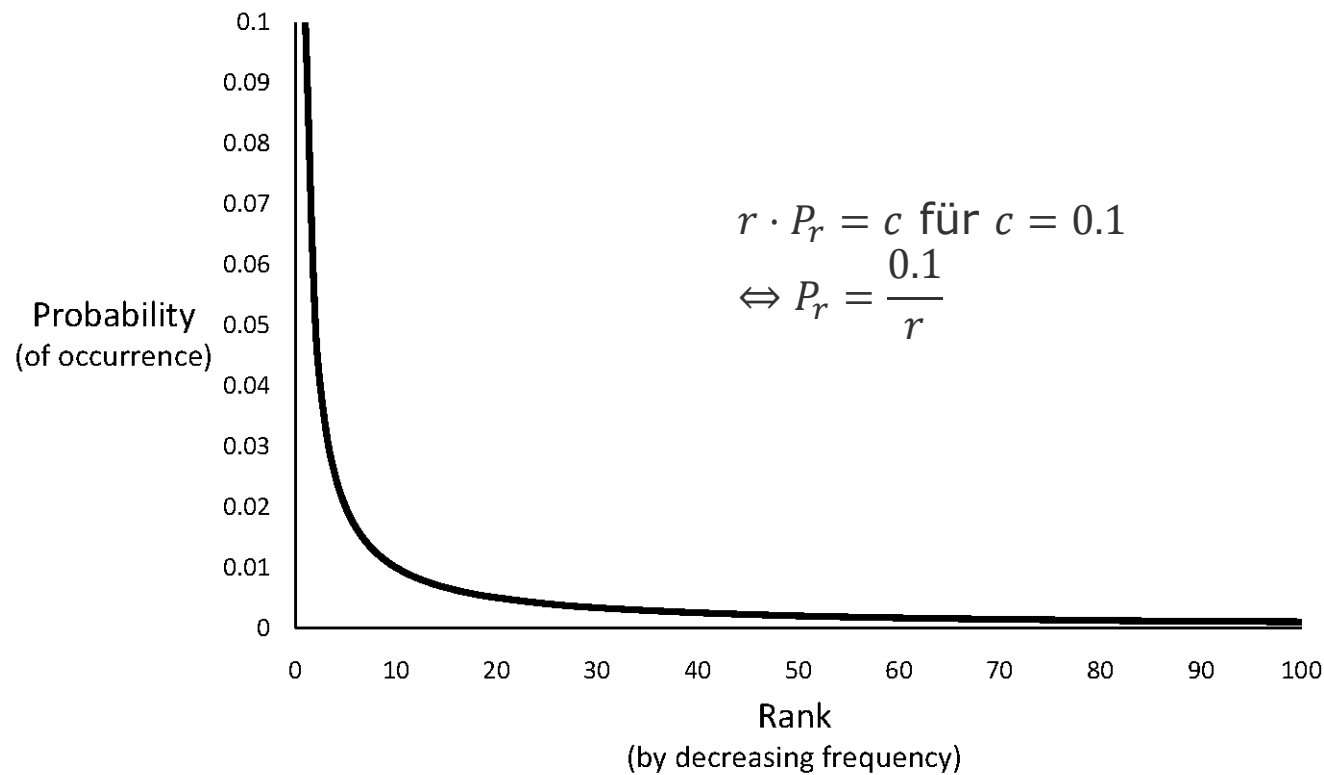
- Benutzerinteraktion
- Ranking
- Evaluierung



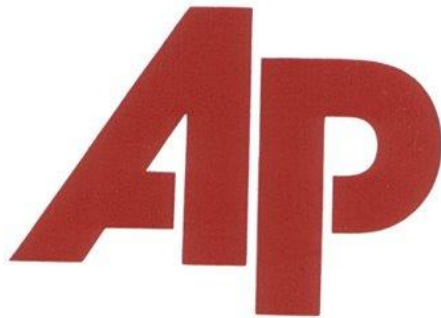
- Verteilung der Worthäufigkeiten ist asymmetrisch
 - Einige Wörter kommen sehr oft vor, sehr viele Wörter kommen sehr selten vor
 - Im Englischen machen die Wörter „the“ und „of“ 10% aller Wörter in Textdokumenten aus
 - Top 6 Wörter machen 20% aus
 - Top 50 Wörter machen 40% aus
 - Und: 50% aller Wörter einer großen Textsammlung kommen nur einmal vor
- Zipfsches Gesetz
 - Beobachtung:
Der **Rang** r eines Wortes mal
seiner **Auftrittswahrscheinlichkeit** P_r ist
ungefähr **konstant**
 - $r \cdot P_r \approx c$



George Kingsley Zipf
(1902–1950)



- Sammlung von Zeitungsartikeln der Associated Press von 1989

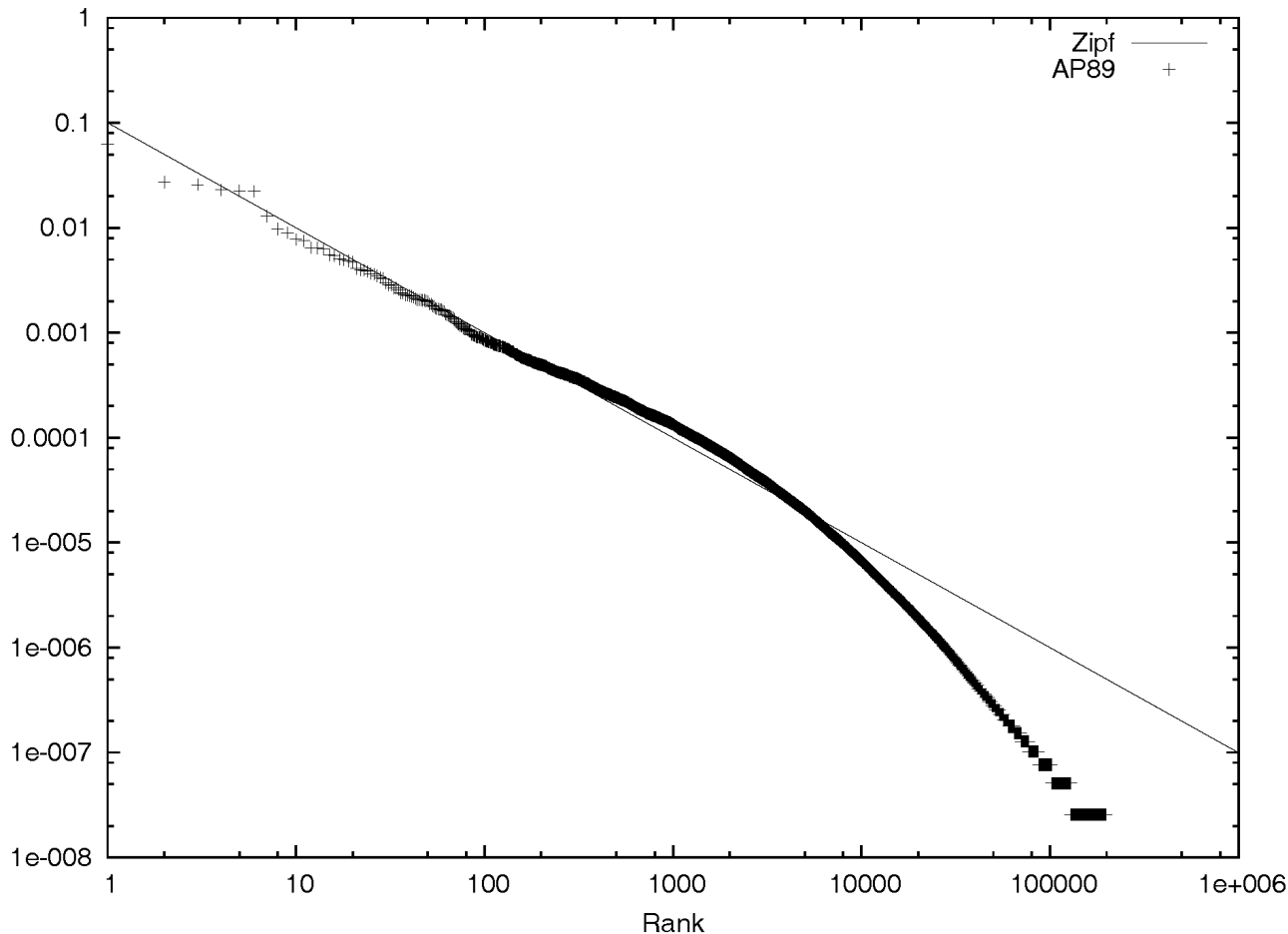


Anzahl Dokumente	84.678
Anzahl Wörter	39.749.179
Größe des Vokabulars	198.763
Wörter die mehr als 1000 mal vorkommen	4,169
Wörter die genau einmal vorkommen	70,064

Word	Freq.	r	$P_r(\%)$	$r.P_r$	Word	Freq.	r	$P_r(\%)$	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

■ Log-Log-Graph

- Zipfsches Gesetz ungenau für sehr niedrige und sehr hohe Häufigkeiten



- Zipfsches Gesetz → Nicht alle Wörter sind **gleich** nützlich für Suche
 - Entfernen/Ignorieren von nicht bedeutungstragenden Wörtern, wie z.B.
 - Bestimmte Artikel („der“, „die“, „das“)
 - Unbestimmte Artikel („ein“, „eine“, ...)
 - Präpositionen („an“, „in“,...)
 - Konjunktionen („und“, „oder“,...)
 - Entfernen von Wörtern, die in fast allen Dokumenten einer Sammlung vorkommen
 - Patentsuchmaschine („Erfindung“, „Design“,...)
 - Implementiert durch **Stoppwortliste**
 - **Vorteile**
 - Index verbraucht weniger Speicher
 - Anfrage kann schneller beantwortet werden
 - Weniger irrelevanten Ergebnisse
 - **Aber:** Manche Anfragen liefern keine Ergebnisse: „Wo ist DER“, „C und A“, ...
- openHPI | Wie funktioniert eine Suchmaschine? | Dr. Ralf Krestel

- Macht aus Texten eine Sequenz von Wörtern
 - E.g. alle alphanumerischen Zeichen einer Sequenz größer drei
 - Umwandlung der Großbuchstaben in Kleinbuchstaben
 - Getrennt durch Leerzeichen oder Sonderzeichen („.", „?", „ ",...)
- Nicht einfach, da...
 - Kurze Wörter auch wichtig sein können („xp“, „a5“, „2. Weltkrieg“,...)
 - Groß- und Kleinschreibung manchmal wichtig ist („Apple“ vs. apple,)
 - Bindestriche manchmal wichtig sind („t-shirt“, „cd-rom“, „e-bay“,...)
 - Apostrophe manchmal wichtig sind („O'Donnell“, 80's,...) oder falsch:



Bitte die Tür zum Hof
steht's verschlossen
halten!!!



- Umwandeln von Wörtern in ihren Wortstamm (**Stemming**)
 - Beliebt für Englisch: „runner“, „runs“, „running“ → „run“
- Umwandeln von Wörtern in ihre Grundform (**Lemmatisierung**)
 - Besser geeignet für Deutsch: „läuft“, „gelaufen“ → „laufen“
- Lemmatisierung ist teurer als Stemming
- Verbesserung der Suche für Englisch etwa 5-10%
- Aber: Für manche Sprachen deutlich mehr
 - Z.B. arabisch

kitab	<i>a book</i>
kitabī	<i>my book</i>
alkitab	<i>the book</i>
kitabuki	<i>your book (f)</i>
kitabuka	<i>your book (m)</i>
kitabuhu	<i>his book</i>
kataba	<i>to write</i>
maktaba	<i>library, bookstore</i>
maktab	<i>office</i>

Wörter mit der arabischen Grundform **ktb**