§ DAS ZIPFSCHE GESETZ §

Lina Nowacki, Maria Verbitski, Alexis Neumann unter Leitung von Susanne Lack

GRUNDPRINZIP

Mathematische Darstellung

$$p(n) \sim \frac{1}{n}$$

- p(n): Wahrscheinlichkeit oder relative Häufigkeit eines Elements
- n: Rang des Elements in der Häufigkeitsverteilung
- ~ bedeutet "proportional zu"
- Formel drückt aus, dass **die Wahrscheinlichkeit oder Häufigkeit eines Elements umgekehrt proportional zu seinem Rang** ist
- Normierung: Umwandlung der proportionalen Beziehung in eine exakte Wahrscheinlichkeitsverteilung
- Normierte Formel

$$p(n) = \frac{1}{n} \times \frac{1}{H_N}$$

- p(n): Wahrscheinlichkeit des Elements mit Rang n
- Normierungsfaktor H_N: Gegeben durch die harmonische Reihe

$$H_N = \sum_{n=1}^{N} \frac{1}{n}$$

- Da die Harmonische Reihe divergiert, muss eine Gesamtzahl an Elementen gegeben sein:
- N: Gesamtzahl der Elemente in der Verteilung
- Beispiel für N=4

$$H_4 = \sum_{n=1}^{4} \frac{1}{n} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12} = 2\frac{1}{12}$$

$$p(1) = \frac{1}{1} \times \frac{1}{\frac{25}{12}} = \frac{12}{25} \qquad p(2) = \frac{1}{2} \times \frac{1}{\frac{25}{12}} = \frac{6}{25}$$

$$p(3) = \frac{1}{3} \times \frac{1}{\frac{25}{12}} = \frac{4}{25} \qquad p(4) = \frac{1}{4} \times \frac{1}{\frac{25}{12}} = \frac{3}{25}$$

GRUNDGEDANKE

- Nicht alle Wörter werden gleich häufig benutzt, z.B. wird “die” häufiger benutzt als “Algebra”
- Nur wenige Wörter werden sehr häufig benutzt, viele Wörter aber werden selten benutzt
- Das Zipf-Gesetz beschreibt folgenden Zusammenhang zwischen der Auftrittswahrscheinlichkeit eines Wortes und seinem Rang in der nach Häufigkeit sortierten Liste:
  - Das häufigste Element tritt etwa doppelt so oft auf wie das zweithäufigste
  - Das häufigste Element tritt etwa dreimal so oft auf wie das dritthäufigste
  - Das häufigste Element tritt etwa viermal so oft auf wie das vierthäufigste

BEWEISE

- **Divergenz der harmonischen Reihe:**

$$\begin{aligned} H_n &= \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \ldots + \frac{1}{n} \\ &\geqslant \frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \ldots + \frac{1}{n} \\ &= \frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots + \frac{1}{n} \end{aligned}$$

$$\Rightarrow \lim_{n \to \infty} H_N = \infty$$

- **Der Normierungsfaktor:**

$$\sum_{n=1}^{N} p(n) = 1$$

$$p(1) + p(2) + \ldots + p(n) = 1$$

$$x \times \frac{1}{1} + x \times \frac{1}{2} + \ldots + x \times \frac{1}{n} = 1$$

$$x \times \left(\frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{n} \right) = 1$$

$$x \times H_N = 1 \quad \Rightarrow x = \frac{1}{H_N}$$

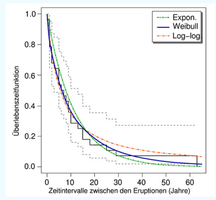
GESCHICHTE

George Kingsley Zipf (1902-1950)

- Amerikanischer **Linguist** und **Philologe**
- Entwickelte **statistische Methoden** für die Untersuchung von Sprachen
- Beginn der **quantitativen Linguistik**
- Versuch die Linguistik zu einer naturwissenschaftsorientierten Wissenschaft zu entwickeln
- Promotion bei der Harvard University
- **1930**: Zipfsches Gesetz
- **1935**: "The Psycho-Biology of Language"
- **1949**: "Human Behavior and the Principle of Least Effort"

VERWENDUNG

- **Universelles Prinzip** in Natur und menschlicher Aktivität
- Erklärt Effizienz in komplexen Systemen
- Zeigt das "Prinzip der geringsten Anstrengung"
- **Vereinfachung** komplexer Systeme
- Erkennung von **Mustern** in chaotischen Daten
- Ermöglichung von **Vorhersagen** in verschiedenen Bereichen
- **Optimierung** von Ressourcenallokation



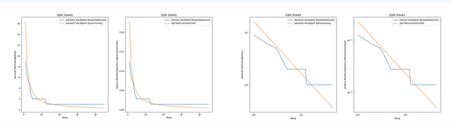
Zipfsche Vulkanausbrüche

ANWENDUNGSBEREICHE

- **Linguistik**: Analyse von Wortfrequenzen
- **IT**: Optimierung von Suchmaschinen und Datenbanken
- **Stadtplanung**: Analyse von Stadtgrößen und -verteilungen
- **Wirtschaft**: Untersuchung von Verkaufszahlen und Unternehmensgrößen
- **Sozialwissenschaften**: Analyse von Einkommens- und Vermögensverteilungen
- **Naturwissenschaften**: Beschreibung von Naturereignissen
- **Biologie**: Untersuchung genetischer Informationen
- **Wissenschaftliche Forschung**: Analyse von Zitationshäufigkeiten

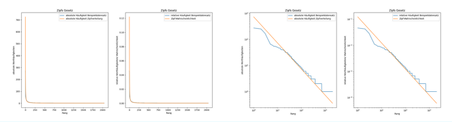
BEISPIELTEXTE

- **Die Zipf-Verteilung in einem Kurzgedicht:**



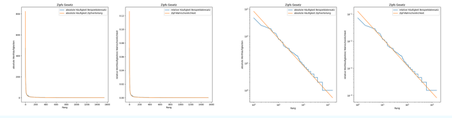
Häufigstes Wort: “die” 9x im Text;
entspricht 12,5% der Wörter
Insgesamt 72 Wörter

- **Die Zipf-Verteilung in einer wissenschaftlichen Veröffentlichung**



Häufigstes Wort: “die” 272x im Text;
entspricht ca. 4,59% der Wörter
Insgesamt 5927 Wörter

- **Die Zipf-Verteilung in einem Englischen Wikipedia-Artikel**



Häufigstes Wort: “the” 472x im Text;
entspricht ca. 7,24% der Wörter
Insgesamt 6519 Wörter
=> Je länger der Text, umso besser kann man die Zipf-Verteilung anwenden.
Vor allem die häufigsten Wörter kommen statistisch etwas zu selten vor

ERKLÄRUNGSANSÄTZE

- ,,**Principle of least effort**'' - ,,**Path of least resistance**''

Ein handelndes Subjekt/Gruppe organisiert seine Umgebung so, dass die zum Handeln erforderliche Energie generell minimiert wird, bezogen auf ein Optimum des zu erreichenden Effektes, der zeitlichen Perspektive und anderen Randbedingungen.
Nützliche Verhaltensweisen werden häufig ausgeführt, wodurch sie über Zeit schneller und einfacher auszuführen sind.
Natürliche Tendenz effizienter mit geringerem Aufwand zu kommunizieren, z.B. ,,Mathe'' statt ,,Mathematik''

- ,,**Preferential Attachement**''

Der bevorzugte Weg wird vereinfacht, sodass immer mehr Leute auch diesen Weg benutzen, vgl. Trampelpfade.
Wörter, die oft benutzt werden haben eine höhere Wahrscheinlichkeit wieder verwendet zu werden.

- **Verteilung zufällig generierter Wörter**

Prämisse: alle 26 Buchstaben und die Leertaste haben dieselbe Wahrscheinlichkeit getippt zu werden, nach einer Leertaste beginnt immer ein Wort mit einem Buchstaben
Wahrscheinlichkeit, dass ein Wort mit n Buchstaben generiert wird:

$$P(n) = \frac{\left(\frac{26}{27}\right)^{n-1}}{26^n \times 27}$$

=> kürzere Wörter sind viel wahrscheinlicher als längere Wörter
=> ergibt exakt die von Zipf vorhergesagte Verteilung

QUELLEN UND METHODEN

- Phyton, Excel, Word, Gillmeister-Software, ESKP
- Die linguistischen Hypothesen von G.K. Zipf - Claudia Prün (Einleitung)
- Quantitative Verteilungen im Wortschatz - Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons - 4 Selbstorganisation und Dynamik - Stefan Engelberg (Mannheim)
- Bernard Zitzer: https://bernardzitzer.com/de/zipfsche-gesetz-zipfs-law/
- Thought.Co: https://www.thoughtco.com/principle-of-least-effort-zipfs-law-1691104
- The Zipf-Mystery - Vsauce: https://www.youtube.com/watch?v=fCn8zs912OE
- Wikipedia: https://de.wikipedia.org/wiki/Zipfsches_Gesetz
- https://de.wikipedia.org/wiki/Pareto-Verteilung
- https://en.wikipedia.org/wiki/Super_Mario
- Über Projektionen: Weltkarten und Weltanschauungen - Julia Mia Stirnemann
- Texte und Daten wurden in Excel und, mithilfe eines passenden Programmes, in Phyton analysiert und in Graphen dargestellt. Desweiteren wurden diese mit der idealen Zipfschen Konstante verglichen und die Abweichungen berechnet.

ERKLÄRUNGSANSÄTZE 2.0

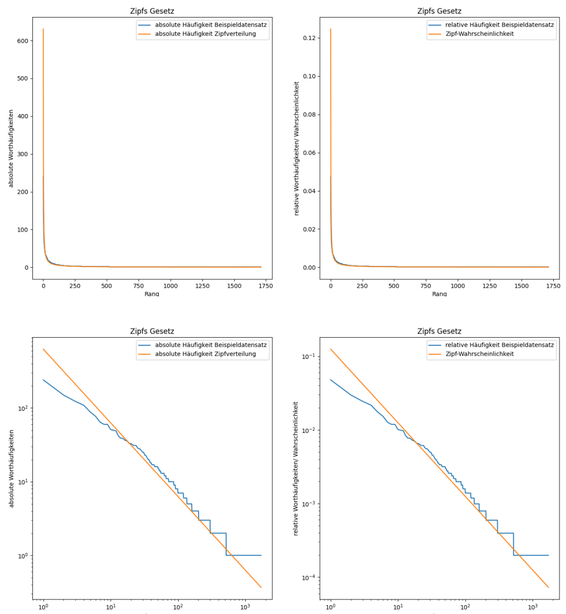
- ,,**Paretoprinzip**'' **- 80/20-Prinzip**

Vom Italiener Vilfredo Pareto erkannt ist die Paretoverteilung eine häufig aufkommende Wahrscheinlichkeitsverteilung, in der 80% der Fälle durch 20% der Elemente abgedeckt sind, z.B. 80% des Landes in Italien gehört 20% der Leute.
Die Zipf-Verteilung und die Paretoverteilung sind Umkehrfunktionen.
Wie bereits erwähnt, erstellt die Zipf-Verteilung einen Zusammenhang zwischen p(n) und dem Kehrwert von n: $p(n) \sim \left(\frac{1}{n}\right)^a$
Wir betrachten in unserem Beispiel den einfachen Fall $a = 1$
=> Die Paretoverteilung und die Zipf-Verteilung sind in unserem Fall identisch
=> 20% aller existierenden Wörter machen 80% der Sprache aus

- **Die Sprecher-Hörer-Beziehung**

Für einen Sprecher stellt das Sprechen die geringste Anforderung dar, wenn er häufig dieselben Wörter benutzt.
Für einen Zuhörer stellt das Verstehen des Gesprochenen die geringste Anforderung dar, wenn der Sprecher ein möglichst differnziertes Vokabular benutzt. Das Ergebnis daraus ist, dass wenige Wörter häufig benutzt werden (Sprecher), und viele Wörter selten (Hörer).

BAROCKE GEDICHTSAMMLUNG

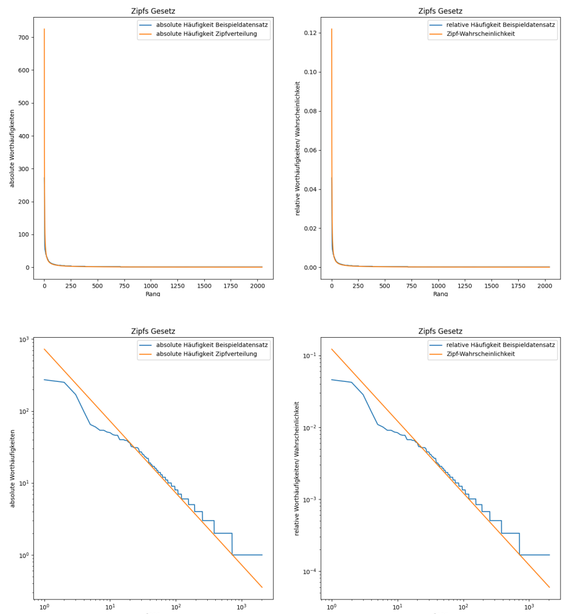


'und': 241
'die': 150
'der': 123
'ich': 109
'nicht': 88
'was': 77
'ist': 64
'in': 60
'ein': 60
'so': 51

5058 Wörter insgesamt

https://www.deutschelyrik.de/home.html

ZEITUNGSLIVEBLOCK

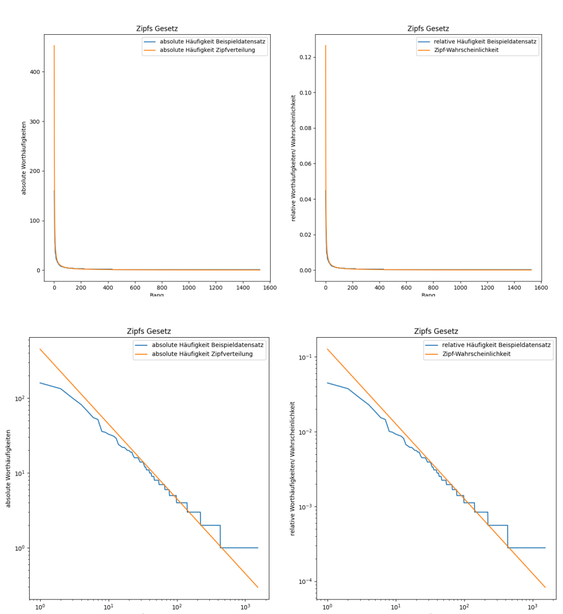


'der': 246
'die': 215
'und': 149
'in': 100
'den': 87
'mit': 82
'merz': 68
'auf': 68
'das': 64
'von': 63

6206 Wörter insgesamt

https://www.sueddeutsche.de/politik/bundestagswahl-2025-news-tv-duell-merz-scholz-umfrage-li.3194431

WIKIPEDIA-ARIKIKEL MATHE

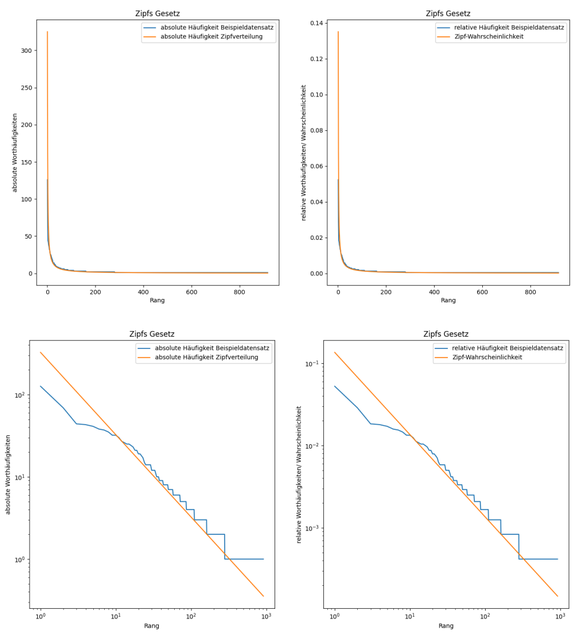


'die': 160
'der': 134
'und': 100
'mathematik': 82
'in': 66
'ist': 55
'von': 52
'das': 36
'im': 35
'eine': 33

3580 Wörter insgesamt

https://de.wikipedia.org/wiki/Mathematik

DAS GESPRACH MIT DEM BETER - KAFKA

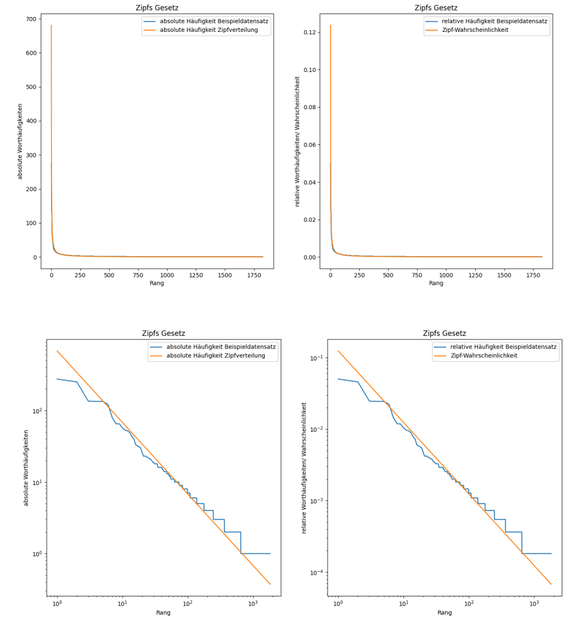


'ich': 126
'und': 69
'die': 44
'nicht': 43
'sie': 41
'der': 38
'in': 37
'er': 35
'es': 32
'den': 32

2394 Wörter insgesamt

https://www.reclam.de/data/media/978-3-15-011106-2.pdf

WIKIPEDIA-ARTIKEL AUF HINDI



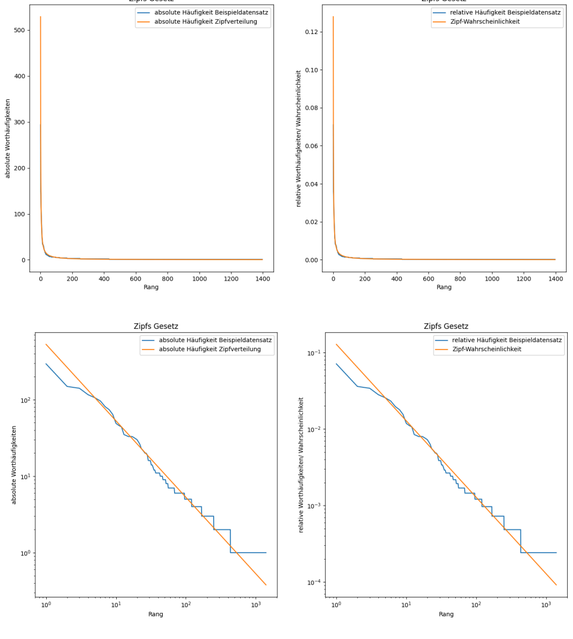
'के': 275
'में': 251
'से': 135
'और': 134
'फ्लोरिडा': 134
'की': 122
'का': 79
'को': 65
'राज्य': 65
'एक': 57

'von': 275
'in': 251
'aus': 135
'und': 134
'Florida': 134
'von': 122
'von': 79
'zu': 65
'Zustand': 65
'eins': 57

5573 Wörter insgesamt

https://hi.wikipedia.org/wiki/%E0%A4%AB%E0%A4%BC%E0%A5%8D%E0%A4%B2%E0%A5%8B%E0%A4%B0%E0%A4%BF%E0%A4%A1%E0%A4%BE

WIKIPEDIA-ARTIKEL AUF FRANZOSISCH



'de': 294
'la': 149
'les': 141
'frites': 116
'en': 106
'et': 96
'à': 81
'le': 74
'dans': 64
'des': 49

4135 Wörter insgesamt

https://fr.wikipedia.org/wiki/Frite