Introduction
(SNLP tutorial)

Vilém Zouhar

March 16, 2021

Overview

- Languages
- TODO

Languages

Language

$L \subseteq \Sigma^*$ (all possible substrings by elements of alphabet Σ)

- $\Sigma_1 = \{\texttt{a},\texttt{b},\ldots,\texttt{z},\texttt{ü},\texttt{ä},\texttt{ö}\}$
- $\Sigma_2 = \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}$
- $\Sigma_3 = \{\texttt{def},\texttt{True},\texttt{:},\texttt{print},\ldots\}$
- $\Sigma_4 = \{\texttt{SELECT},\texttt{INSERT},\texttt{DROP},\ldots\}$
- $\Sigma_5 = \{\texttt{hallo},\texttt{ja},\texttt{nein},\ldots\}$
- $\Sigma_6 = \{\texttt{+},\texttt{-},\texttt{=},\texttt{1},\texttt{2},\texttt{3}\ldots\}$
- $\Sigma_7 = \{\texttt{+},\texttt{-},\texttt{=},\texttt{1},\texttt{2},\texttt{3}\ldots\}$

- '`Oberfläche`' $\in L_1$ (German words)
- '`..GATTCCAATCAG`' $\in L_2$ (DNA)
- '`while True: f()`' $\in L_3$ (Python)
- '`SELECT * FROM tbl;`' $\in L_4$ (SQL)
- '`Wie geht's dir?`' $\in L_5$ (German)
- '`4=5`' $\in L_6$ (arithmetics)
- '`1=2+=3333=`' $\in L_7$ (???)

Usually defined by the alphabet and production rules (Automata and Grammar).

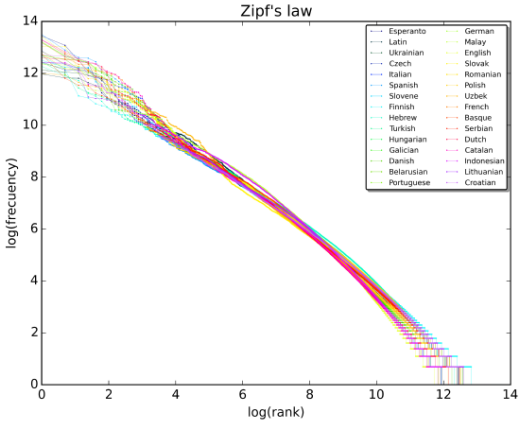
Zipf's Law

1. Sort words/entries by frequency f(x)
2. r(x) = position in the sorted list
   - Then $\exists\gamma : f(x) \propto \frac{1}{r(x)^\gamma}$
3. Most common entry m.
   - Then $\exists\gamma : f(x) = \frac{f(m)}{r(x)^\gamma}$

Rank	Frequency	Predicted ($\gamma = 0.7$)
4	606	$1507/4^{0.7} = 571$
5	490	$1507/5^{0.7} = 488$
...		
11	261	$1507/11^{0.7} = 281$
12	252	$1507/12^{0.7} = 264$

Rank	Word [3]	Frequency
1	the	1507
2	and	714
3	to	703
4	a	606
5	of	490
6	she	484
7	said	416
8	it	346
9	in	345
10	was	328
11	I	261
12	you	252
13	as	237
14	Alice	221

Zipf's Law



We are plotting:

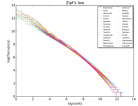
$$\log\left(\frac{C}{\exp(x)^{\gamma}}\right)$$
$$\log(C) - \gamma \log \exp(x)$$
$$\log(C) - \gamma x$$

Figure 1: log-log plot of Zipf's Law applpied to natural languages [4]

Introduction

Zipf's Law

Zipf's Law



We are plotting:
$\log\left(\frac{\text{rank}}{\text{total}}\right)$
$\log(C) - \gamma \log \exp(x)$
$\log(C) - \gamma x$

Figure 1: log-log plot of Zipf's Law appliped to natural languages [4]

Multiple ranks share same (low) frequency $\rightarrow$ stairs at the end

Zipf's Law Notes

- Works beyond natural languages
- DNA subsequences of fixed lengths
- Code (programming languages)
- Population of cities (frequency is city population) [5]

Rank	City	Population
1	New York, N.Y.	8,491,079
2	Los Angeles, Calif.	3,928,864
3	Chicago, Ill.	2,722,389
4	Houston, Tex.	2,239,558
5	Philadelphia, Pa.	1,560,297
6	Phoenix, Ariz.	1,537,058
7	San Antonio, Tex.	1,436,697
8	San Diego, Calif.	1,381,069
9	Dallas, Tex.	1,381,069

Probability Theory

TODO

Word probability

TODO

Conditional probability

TODO

Marginalization

TODO

Independent variables

TODO

Probability Theory

TODO

Bayes rule/decomposition

Expected value

Homework

TBD

Resources

1. UdS SNLP Class: https://teaching.lsv.uni-saarland.de/snlp/
2. Tutorial repository: https://github.com/zouharvi/uds-snlp-tutorial
3. Alice in Wonderland, Lewis Caroll
4. SergioJimenez, CC BY-SA 4.0
   https://commons.wikimedia.org/wiki/File:Zipf_30wiki_en_labels.png
5. https://www.futurelearn.com/info/courses/maths-linear-quadratic/0/steps/12150