PROBABILITY AND INFORMATION THEORY

Outline

- Intro
- Basics of probability and information theory
  - Probability space
  - Rules of probability
  - Useful distributions
  - Zipf’s law & Heaps’ law
  - Information content
  - Entropy (average content)
  - Lossless compression
  - Tf-idf weighting scheme
- Retrieval models
- Retrieval evaluation
- Link analysis
- From queries to top-k results
- Social search

Set-theoretic view of probability theory

- Probability space
  - (Ω, E, P) with
  - Ω: sample space of elementary events
  - E: event space, i.e. subsets of Ω, closed under $\cap$, $\cup$, and $\neg$, usually $E = 2^{\Omega}$
  - P: $E \rightarrow [0, 1]$, probability measure

Properties of P (set-theoretic view):

1. $P(\emptyset) = 0$ (impossible event)
2. $P(\Omega) = 1$
3. $P(A) + P(\neg A) = 1$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(\bigcup_i A_i) = \sum_i P(A_i)$ for pairwise disjoint A_i

Sample space and events: examples

- Rolling a die
  - Sample space: {1, 2, 3, 4, 5, 6}
  - Probability of even number: looking for events $A = \{2\}$, $B = \{4\}$, $C = \{6\}$,
    $P(A \cup B \cup C) = 1/6 + 1/6 + 1/6 = 0.5$

- Tossing two coins
  - Sample space: {HH, HT, TH, TT}
  - Probability of HH or TT: looking for events $A = \{TT\}$, $B = \{HH\}$, $P(A \cup B) = 0.5$

- In general, when all outcomes in Ω are equally likely, for an $e \in E$ holds:

$$P(e) = \frac{\#\text{ outcomes in } e}{\#\text{ outcomes in sample space}}$$

Calculating with probabilities

- Total/marginal probability
  - $P(B) = \sum_j P(B \cap A_j)$ for any partitioning of Ω in $A_1, \dots, A_n$ (sum rule)
- Joint and conditional probability
  - $P(A, B) = P(A \cap B) = P(B|A)\,P(A)$ (product rule)
  - Bayes' theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

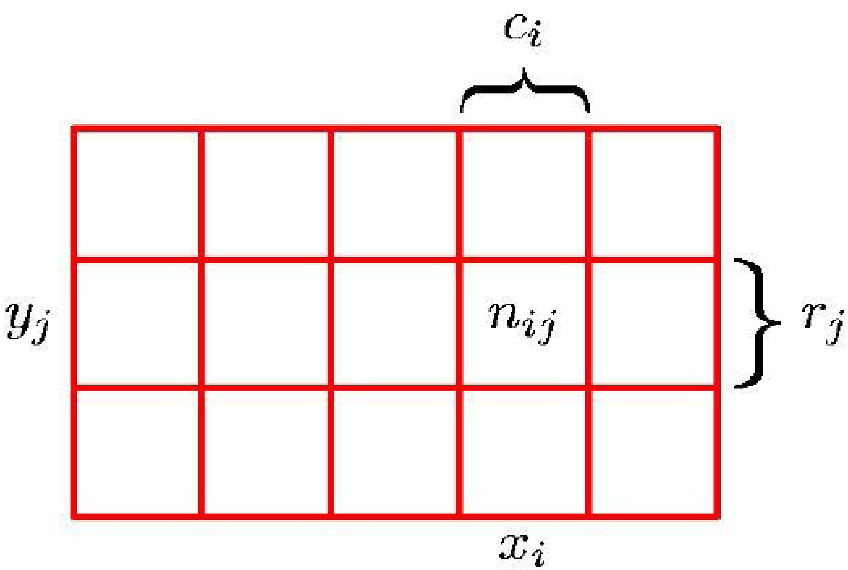
Thomas Bayes

- Independence
  - $P(A_1, \dots, A_n) = P(A_1 \cap \dots \cap A_n) = P(A_1)\,P(A_2)\,\dots\,P(A_n)$, for independent events $A_1, \dots, A_n$
- Conditional Independence
  - A is independent of B given $C \Leftrightarrow P(A|B, C) = P(A|C)$
  - If A and B are independent, are they also independent given C?

Discrete and continuous random variables

- Random variable on probability space (Ω, E, P)
  - $X\colon \Omega \rightarrow M \subseteq \mathbb{R}$ (numerical representations of outcomes) with $\{e | X(e) \leq x\} \in E$ for all $x \in M$
  - Examples
    - Rolling a die: $X(i) = i$
    - Rolling two dice: $X(a, b) = 6(a-1) + b$
    - If M is countable X is called discrete, otherwise continuous

Calculating probabilities: example (1)



Example from C. Bishop: PRML

Marginal probability

$$P(X = x_i) = \frac{c_i}{N}$$

Sum rule

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$
$$= \frac{1}{N} \sum_j n_{ij} = \frac{c_i}{N}$$

Joint probability

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

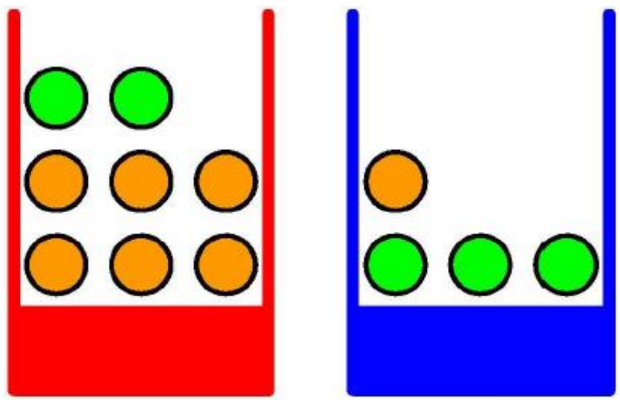
Product rule

$$P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i) P(X = x_i)$$
$$= \frac{n_{ij}}{c_i} \frac{c_i}{N} = \frac{n_{ij}}{N}$$

Calculating probabilities: example (2)

Suppose: $P(B = r) = 2/5$

Apples and Oranges



Fruit is orange, what is probability that box was blue?

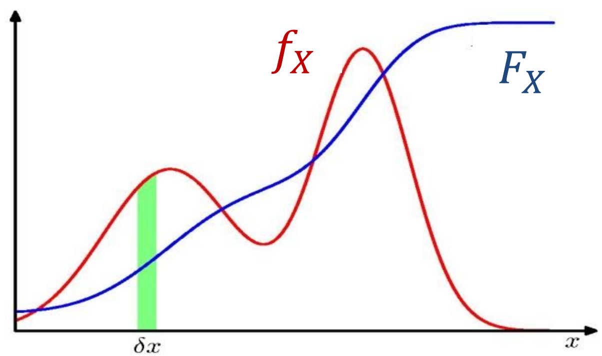
$$P(B = b \mid F = o) = \frac{P(F = o \mid B = b)\, P(B = b)}{P(F = o)}$$

$$P(F = o) = P(F = o \mid B = r)\, P(B = r) + P(F = o \mid B = b)\, P(B = b) = 9/20$$

Example from C. Bishop: PRML

Pdfs, cdfs, and quantiles

- Probability density function (pdf)
  - $f_X : M \rightarrow [0,1]$ with $f_X(x) = P(X = x)$
- Cumulative distribution function (cdf)
  - $F_X : M \rightarrow [0,1]$ with $F_X(x) = P(X \leq x)$

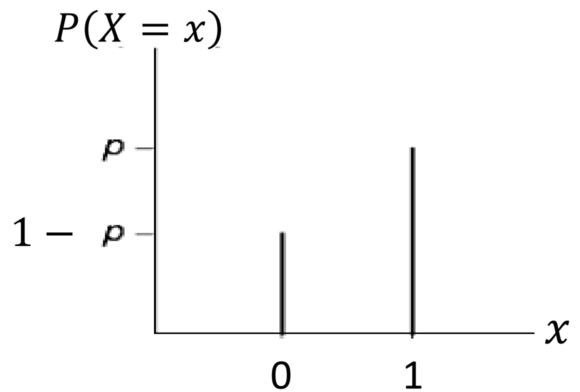


From C. Bishop: Pattern Recognition and Machine Learning

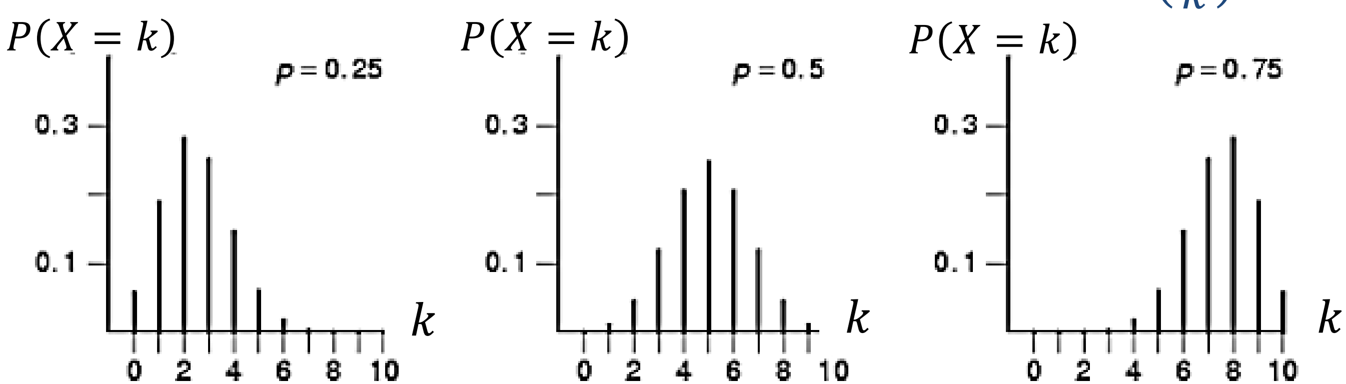
- Quantile function
  - $F^{-1}(q) = \inf\{x | F_X(x) > q\},\ q \in [0,1]$ (for $q = 0.5$, $F^{-1}(q)$ is called **median**)

Useful distributions (1)

- Examples of discrete distributions
  - Uniform distribution over {1, 2, …, m}: $P(X=k)=f_X(k)=\frac{1}{m}$
  - Bernoulli distribution with parameter p: $P(X=x)=f_X(x)=\ p^x(1-p)^{1-x}$



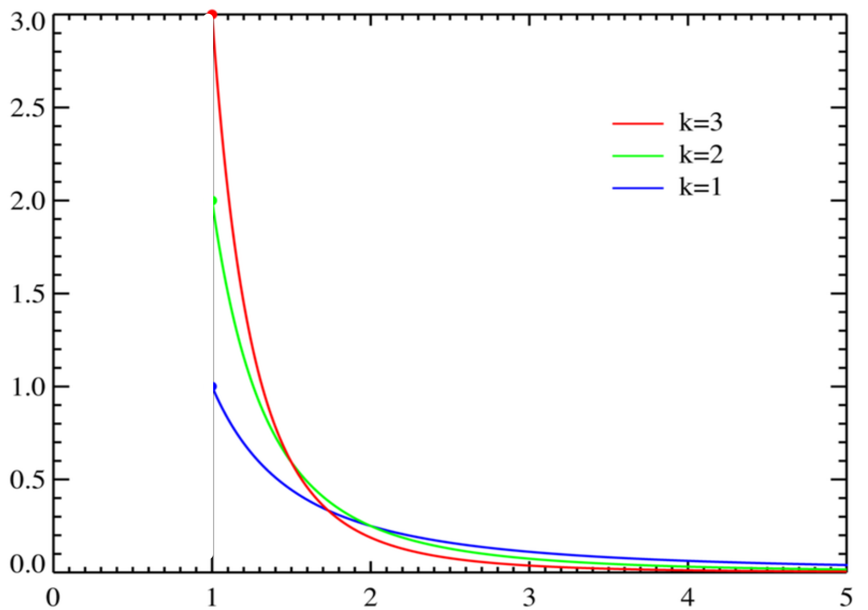
  - Binomial distribution with parameter p: $P(X=k)=f_X(k)=\binom{m}{k}p^k(1-p)^{m-k}$



Useful distributions (2)

- Examples of continuous distributions

  - Uniform distribution over [a, b] : $P(X = x) = f_X(x) = \frac{1}{b-a}$ for $a < x < b$
  - Pareto distribution: $P(X = x) = f_X(x) = \frac{k}{b}\left(\frac{b}{x}\right)^{k+1}$ for $x > b$



Source: Wikipedia

Multivariate distributions

- Let $X_1, \dots, X_m$ be random variables over the same prob. space with domains $dom(X_1), \dots, dom(X_m)$.

  The ***joint distribution*** of $X_1, \dots, X_m$ has a pdf $f_{X_1,\dots,X_m}(x_1, \dots, x_m)$ with

  $\sum_{x_1 \in dom(X_1)} \cdots \sum_{x_m \in dom(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m) = 1$, or

  $\int_{x_1 \in dom(X_1)} \cdots \int_{x_m \in dom(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m)\ dx_1 \dots dx_m = 1$

  The ***marginal distribution*** of X_i is $F_{X_1,\dots,X_m}(x_i) =$

$\sum_{x_1 \in dom(X_1)} \cdots \sum_{x_{i-1} \in dom(X_{i-1})} \sum_{x_{i+1} \in dom(X_{i+1})} \cdots \sum_{x_m \in dom(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m)$
or

$\int_{x_1 \in dom(X_1)} \cdots \int_{x_{i-1} \in dom(X_{i-1})} \int_{x_{i+1} \in dom(X_{i+1})} \cdots \int_{x_m \in dom(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m)\ dx_1 \dots dx_m$

Multivariate distribution: example

- **Multinomial distribution** with parameters n, m (rolling n m-sided dice)

  $$P(X_1 = k_1 \dots X_m = k_m) = f_{X_1,\dots,X_m}(k_1, \dots, k_m) = \frac{n!}{k_1! \dots k_m!} {p_1}^{k_1} \dots {p_m}^{k_m}$$

  with $k_1 + \cdots + k_m = n$ and $p_1 + \ \dots + \ p_m = 1$

  Note: in information retrieval, the multinomial distribution is often used to model the following case:

  - document d with n terms from the alphabet $\{w_1, \dots, w_m\}$, where each w_i occurs k_i times in d

Expectation, variance, and covariance

- **Expectation**
  - For discrete variable X: $E(X) = \sum_x x\, f_X(x)$
  - For continuous variable X: $E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$
  - Properties
    - $E(X_i + X_j) = E(X_i) + E(X_j)$
    - $E(X_i X_j) = E(X_i)E(X_j)$ for independent, identically distributed (i.i.d.) variables X_i, X_j
    - $E(aX + b) = aE(x) + b$ for constants a, b

- **Variance**
  - $Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$, $StDev(X) = \sqrt{Var(X)}$
  - Properties
    - $Var(X_i + X_j) = Var(X_i) + Var(X_j)$ for i.i.d. variables X_i, X_j
    - $Var(aX + b) = a^2 Var(x)$ for constants a, b

- **Covariance**
  - $Cov(X_i, X_j) = E[(X_i - E[X_i])\,(X_j - E[X_j])]$
  - $Var(X) = Cov(X, X)$

Statistical parameter estimation through MLE

- Maximum Likelihood Estimation (MLE)
  - After tossing a coin n times, we have seen k times head.
    Let p be the unknown probability of the coin showing head.
    Is it possible to estimate p?
  - We know observation corresponds to Binomial distribution, hence:

$$L(p;k,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

  - Maximizing $L(p;k,n)$ is equivalent to maximizing $\log L(p;k,n)$
    $\log L(p;k,n)$ is called ***log-likelihood function***

$$\log L(p;k,n) = \log \binom{n}{k} + k \log p + (n-k) \log (1-p)$$

$$\frac{\partial \log L}{\partial p} = \frac{k}{p} - \frac{(n-k)}{(1-p)} = 0 \Rightarrow p = \frac{k}{n}$$

Formal definition of MLE

- Maximum Likelihood Estimation (MLE)

  Let $x_1, \dots, x_n$ be a random sample from a distribution $f(\boldsymbol{\theta}, x)$
  (Note that $x_1, \dots, x_n$ can be viewed as the values of i.i.d. random variables $X_1, \dots, X_n$)
  $L(\boldsymbol{\theta}; x_1, \dots, x_n) = P[x_1, \dots, x_n \text{ originate from} f(\boldsymbol{\theta}, x)]$
  Maximizing $L(\boldsymbol{\theta}; x_1, \dots, x_n)$ is equivalent to maximizing $\log L(\boldsymbol{\theta}; x_1, \dots, x_n)$, i.e., the log-likelihood function: $\log P(x_1, \dots, x_n | \boldsymbol{\theta})$.

  - If $\frac{\partial \log L}{\partial p}$ is analytically intractable, use iterative numerical methods, e.g. Expectation Maximization (EM)
    (More on this, in the Data Mining lecture…)

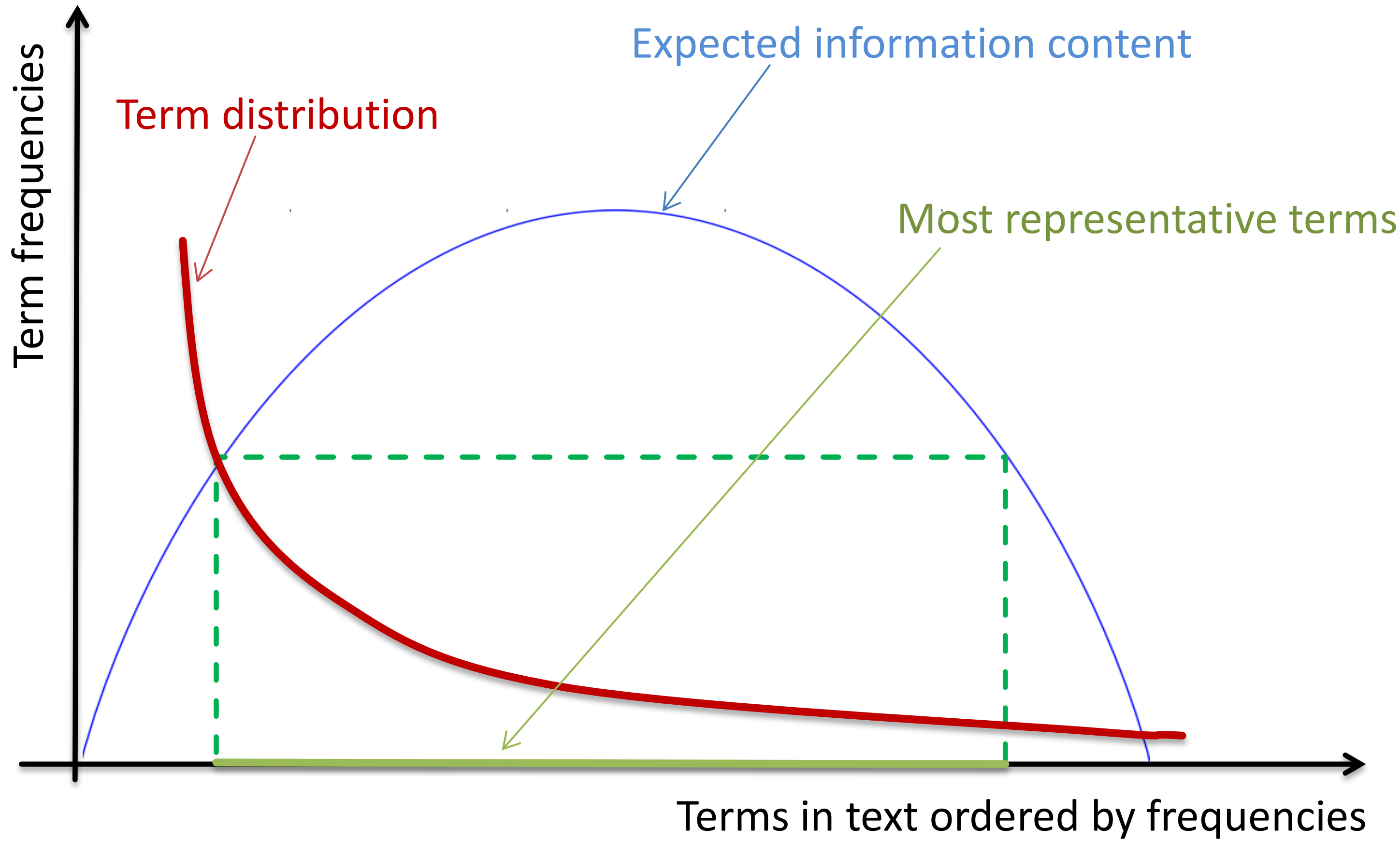
Modeling natural language: three questions

1. Is there a general model for the distribution of terms in natural language?

2. Given a term in a document, what is its information content?

3. Given a document, by which terms is it best represented?

Modeling natural language



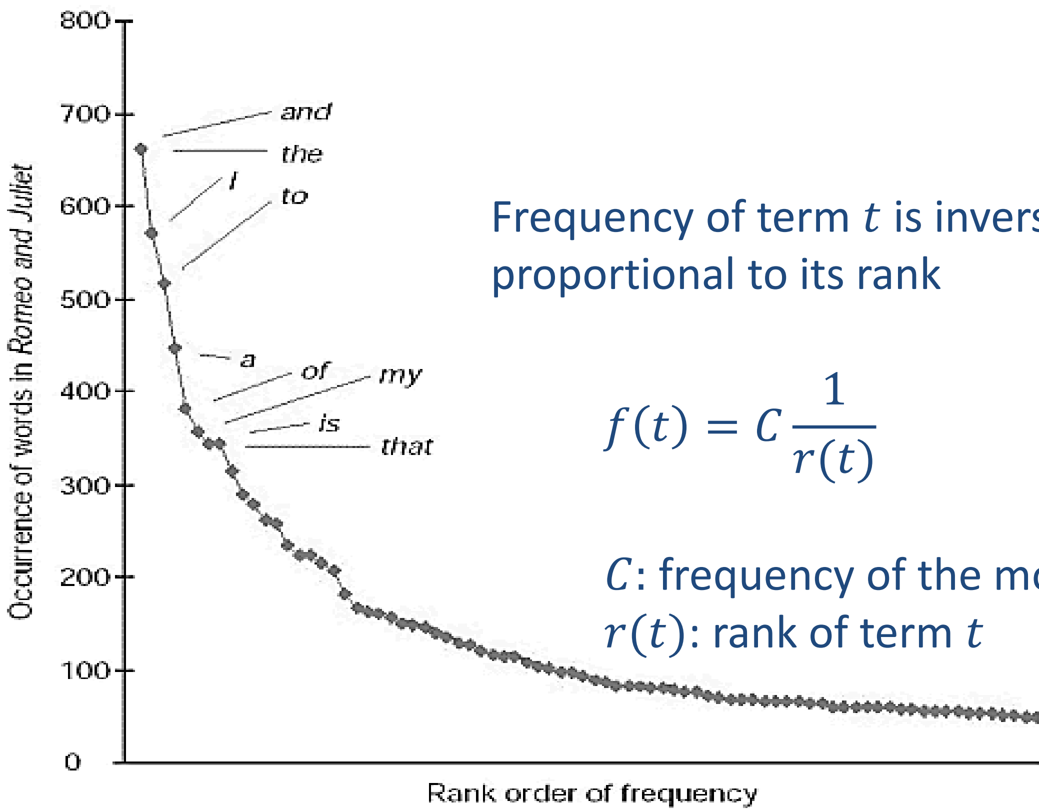
Is there a weighting scheme that gives higher weights to representative terms?

Zipf's law

- **Linguistic observation**

  In large text corpus

  - few terms occur ***very frequently***
  - many terms occur ***infrequently***



Frequency of term t is inversely proportional to its rank

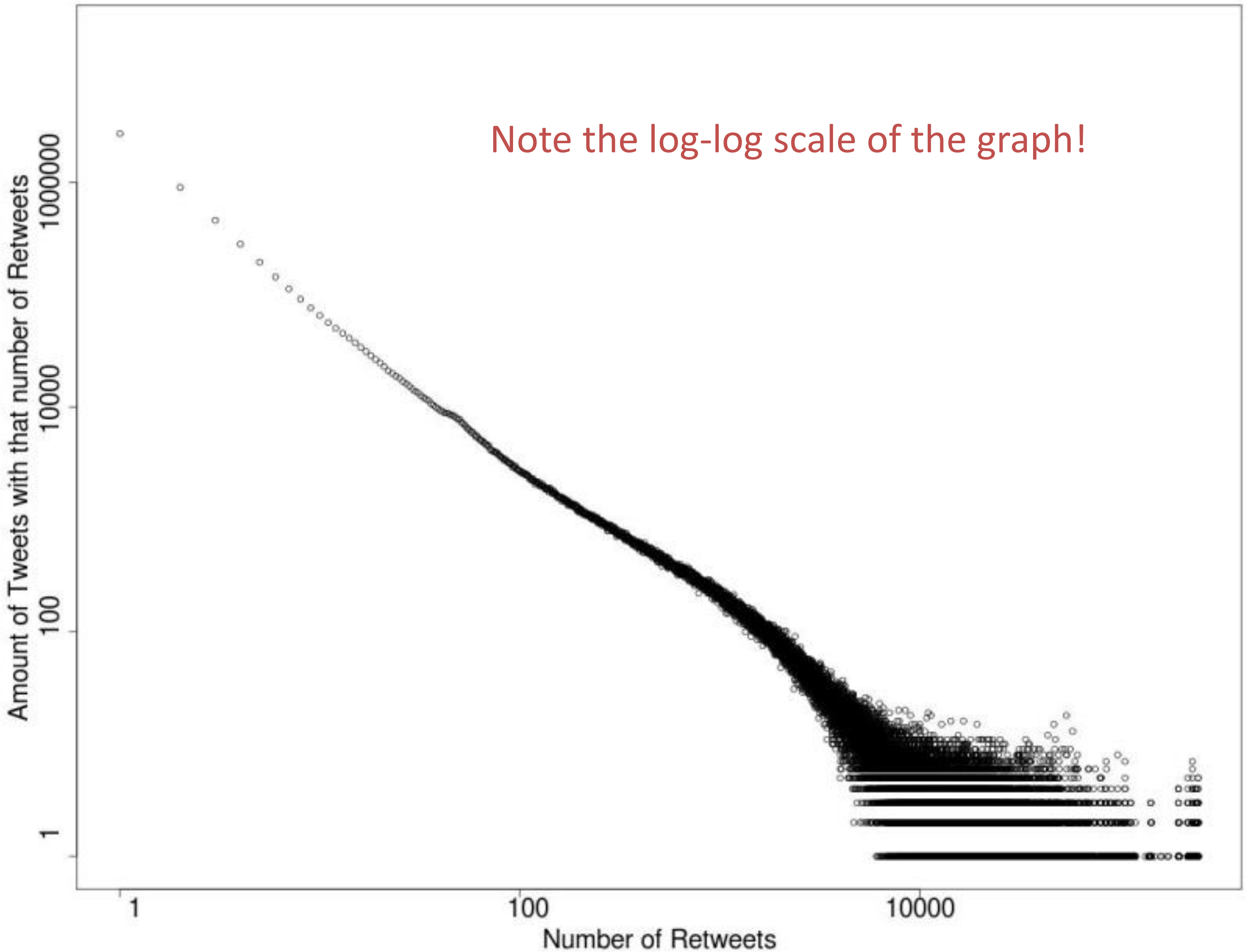
$$f(t) = C\frac{1}{r(t)}$$

C: frequency of the most frequent term

$r(t)$: rank of term t

Source: http://www.ucl.ac.uk/~ucbplrd/language_page.htm

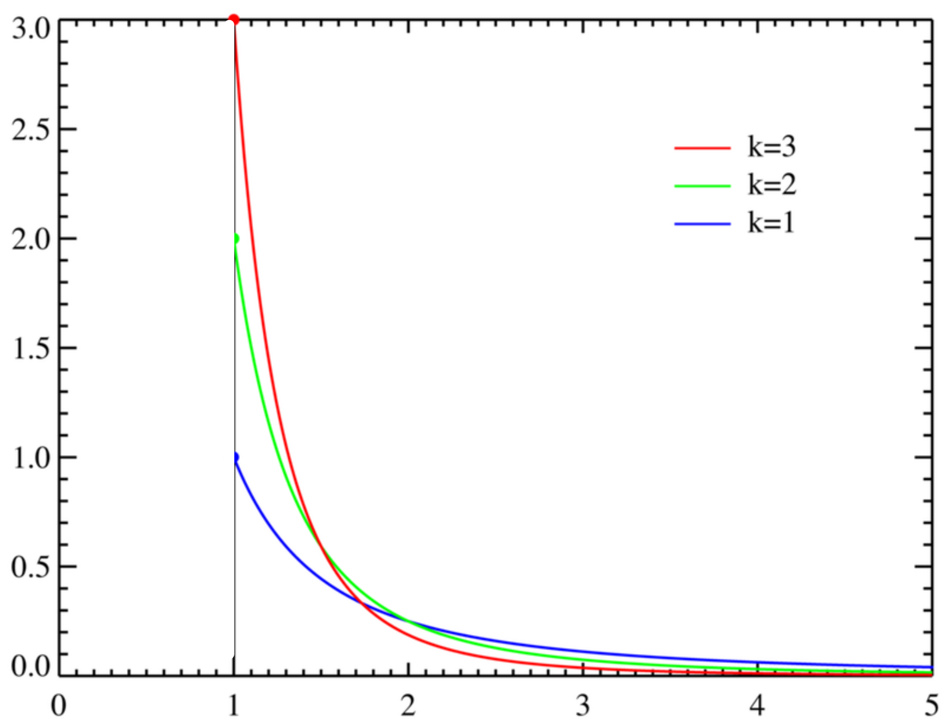
Example: retweets on Twitter



Source: Master’s thesis by M. Jenders (2012)

Pareto distribution

- Probability that continuous random variable X is equal to some value x is given by $f_X(x;k,\theta) = P(X = x) = \begin{cases} \frac{k}{\theta}\left(\frac{\theta}{x}\right)^{k+1} & for\ x \geq \theta \\ 0 & for\ x < \theta \end{cases}$

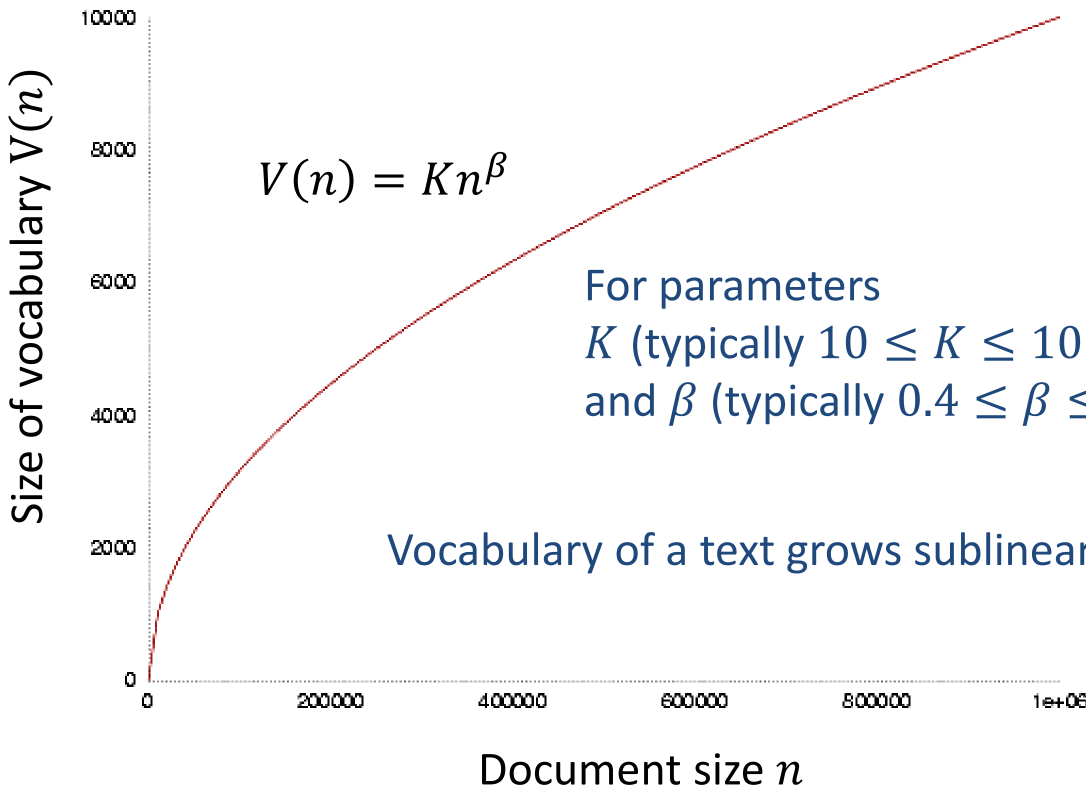


Source: Wikipedia

- Pareto principle
  - 80% of the effects come from 20% of the causes
- Family of distributions
  - Power law distributions
- Examples
  - Distribution of populations over cities
  - Distribution of wealth
  - Degree distribution in web graph (or social graphs)

Heap's law

- Empirical law describing the portion of vocabulary captured by a document

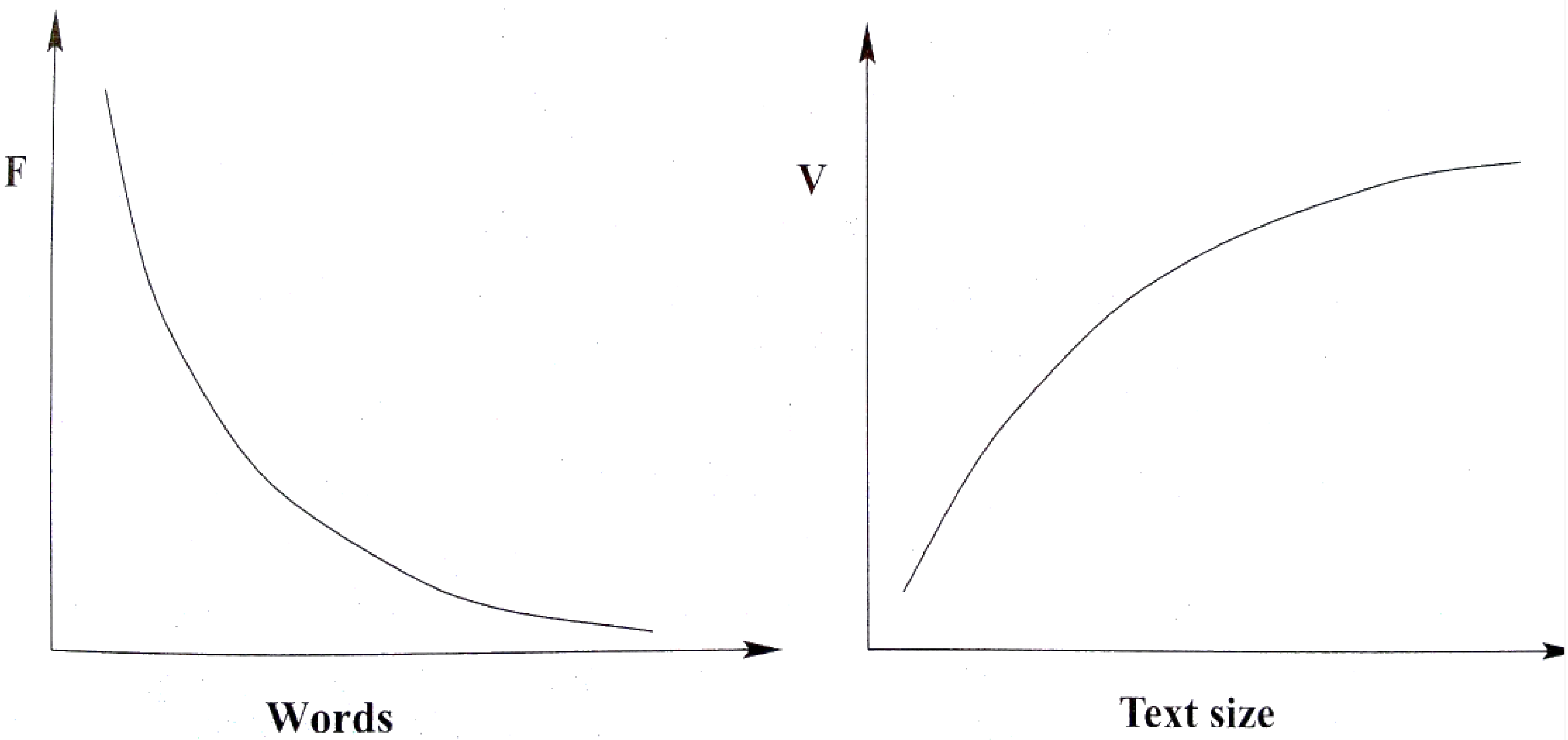


For parameters
K (typically $10 \leq K \leq 100$)
and β (typically $0.4 \leq \beta \leq 0.6$)

Vocabulary of a text grows sublinearly with its size!

See also: Modern Information Retrieval, 6.5.2

Zipf's law & Heaps' law



Source: Modern Information Retrieval

- Two sides of the same coin …
- Both laws suggest **opportunities for compression** (more on this, later)
- How to **compress as much as possible without loosing information**?

From information content to entropy

- **Information content**

  Can we formally capture the content of information?

  - 1. Intuition: the more surprising a piece of information (i.e., event), the higher its information content should be.

    $$h(x)\uparrow \quad P(x)\downarrow$$

  - 2. Intuition: the information content of two independent events x and event y should simply add up (additivity).

    $$h(x+y) = h(x) + h(y)$$

  Define $h(x) := -\log_2 P(x)$

- **Entropy (expected information content)**

  Let X be a random variable with 8 equally possible states.

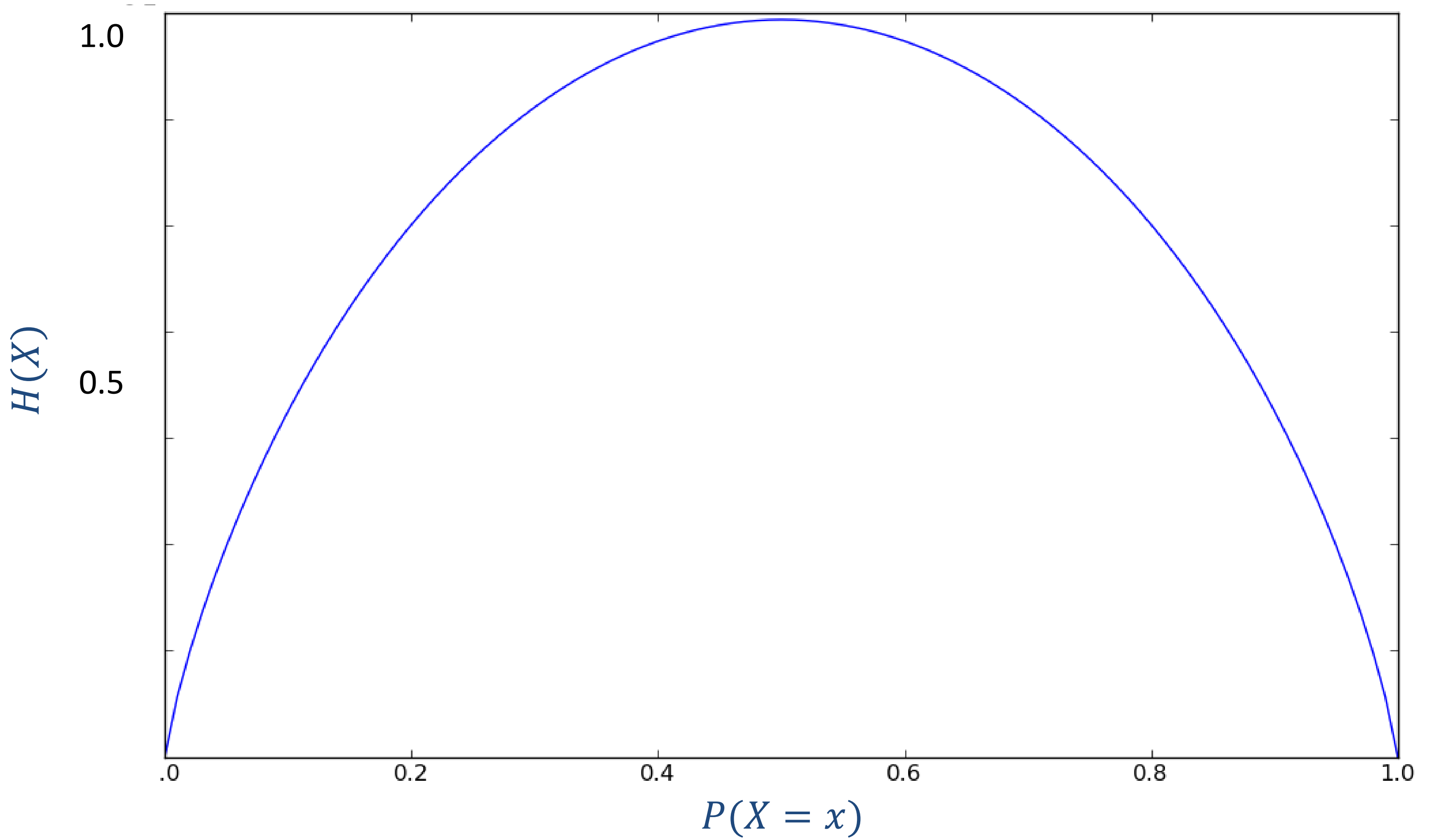
  What is the average number of bits needed to encode a state of X?

  $$H(X) = -\sum_{x \in dom(X)} P(x) \log P(x) \text{ (i.e. the entropy of } X\text{)}$$

  $$= -8\,\frac{1}{8}\log\frac{1}{8} = 3$$

  Also: entropy is a lower bound on the average number of bits needed to encode a state of X.

Entropy function



Relative entropy

- Relative entropy (Kullback-Leibler Divergence)

  Let f and g be two probability density functions over random variable X. Assuming that g is an approximation of f, the additional average number of bits to encode a state of X through g is given by

$$KL(f \parallel g) = \int_x f(x) \log \frac{f(x)}{g(x)} \, dx$$

- Properties of relative entropy
  - $KL(f \parallel g) \geq 0$ (Gibbs' inequality)
  - $KL(f \parallel g) \neq KL(g \parallel f)$ (asymmetric)

- Related symmetric measure: Jensen-Shannon Divergence
  - $JS(f, g) = \alpha\, KL(f \parallel g) + \beta\, KL(g \parallel f)$ with $\alpha + \beta = 1$

Mutual information

- Mutual information

  Let X and Y be two random variables with a joint distribution function P. The degree of their independence is given by

$$I[X,Y] = KL\big(P(X,Y) \parallel P(X)P(Y)\big) = \iint p(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \, dX \, dY$$

- Properties of mutual information
  - $I[X,Y] \geq 0$
  - $I[X,Y] = 0$ if and only if X and Y are independent
  - $I[X,Y] = H[X] - H[X|Y] = H[Y] - H[Y|X]$ (also known as: information gain)
    (i.e., the entropy reduction of X by being told the value of Y)

Lossless compression (1)

- **Huffman compression**

Let X be a random variable with 8 possible states

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$$

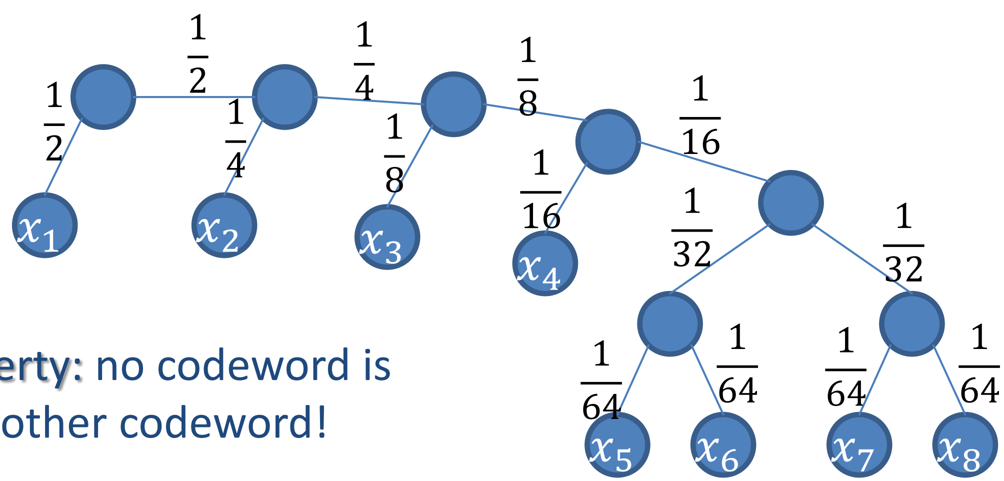
with occurrence probabilities

$$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$$

In any case: 3 bits would be sufficient to encode any of the 8 states.

Can we do better?

$$encoding: 0,10,110,1110,111100,111101,111110,111111$$



Bottom-up tree construction by combining lowest-frequency subtrees

Prefix property: no codeword is prefix of another codeword!

Lossless compression (2)

- **Shannon's noiseless coding theorem**

  Let X be a random variable with n possible states. For any noiseless encoding of the states of X, $H(X)$ is a lower bound on the average code length of a state of X.

- **Theorem**

  The Huffman compression is an **entropy encoding** algorithm (i.e., it achieves the lower bound estimated by entropy)

- **Corollary**

  The Huffman compression is optimal for lossless compression

Lossless compression (3)

- Ziv-Lempel compression (e.g., LZ77)
  - Use **lookahead window** and **backward window** to scan text
  - Identify in lookahead window the longest string that occurs in backward window
  - Replace the string by a pointer to its previous occurrence
  - Text is encoded in triples $(previous, length, new)$

    $previous$: distance to previous occurrence

    $length$: length of the string

    new: symbol following the string

More advanced variants use adaptive dictionaries with statistical occurrence analysis!

Lossless compression (4)

- **Ziv-Lempel compression (e.g., LZ77)**
  - Example

    Text: $A\ A\ B\ A\ B\ B\ B\ A\ B\ A\ A\ B\ A\ B\ B\ B\ A\ B\ B\ A\ B\ B$

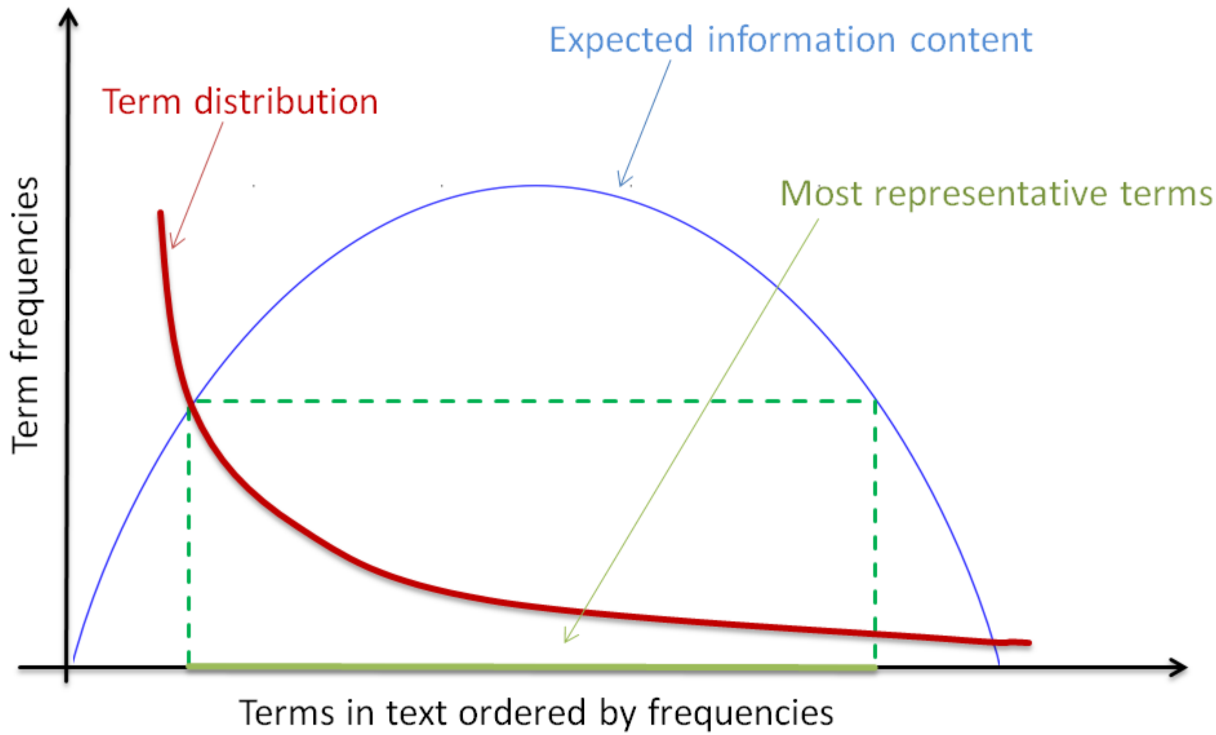
    Code: $(\emptyset, 0, A)(-1,1, B)(-2,2, B)(-4, 3, A)(-9, 8, B)(-3,3, \emptyset)$

- Note that LZ77 and other sophisticated lossless compression algorithms (e.g. LZ78, Lempel-Ziv-Welch,…) encode several states at the same time.

- With appropriately generalized notions of variables and states, Shannon's lossless coding theorem still holds!

Tf-idf weighting scheme (1)

- Given a document, **by which terms is it best represented**?
  - Is there a weighting scheme that gives higher weights to representative terms?



Tf-idf weighting scheme (2)

- Given a document, **by which terms is it best represented**?
  - Is there a weighting scheme that gives higher weights to representative terms?
  - Consider corpus with documents $D = \{d_1, \dots, d_n\}$ with terms from a vocabulary $V = \{t_1, \dots, t_m\}$.
  - The term frequency of term t_i in document d_j is measured by

$$tf(t_i, d_j) = \frac{freq(t_i, d_j)}{max_k\, freq(t_k, d_j)}$$

Normalisation makes estimation independent of document length.

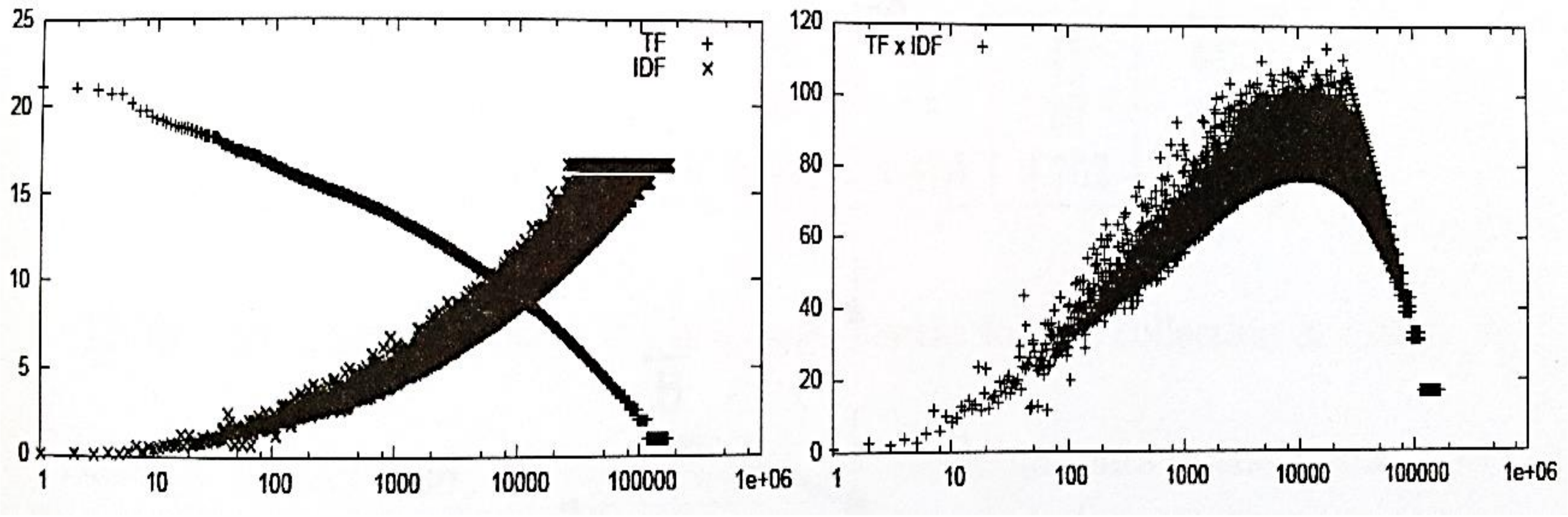
  - The inverse document frequency for a term t_i is measured by

$$idf(t_i, D) = \log \frac{|D|}{|\{d \in D;\, t_i\ occurs\ in\ d\}|}$$

  - Central weighting scheme for scoring and ranking

Downweights terms that ocurr in many documents (i.e., **stop words:** the, to, from, if, …).

Tf, idf, and tf-idf



Tf, idf, and tf-idf weights (plotted in log-scale) computed on a collection from Wall Street Journal (~99,000 articles published between 1987 and 1989)

Source: Modern Information Retrieval

Various tf-idf weighting schemes

- Different weighting schemes based on the tf-idf model, implemented in the SMART system

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$ (Section 6.4.4)
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/\mathit{CharLength}^\alpha, \alpha < 1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

Source: Introduction to Information Retrieval

Probability theory

- **Summary**

  - Sample space, events random variables
  - Sum rule (for marginals), product rule (for joint distributions), Bayes' theorem (using conditionals)
  - Distributions (discrete, continuous, multivariate), pdfs, cdfs, quantiles
  - Expectation, variance, covariance
  - Maximum likelihood estimation

Information theory

- Summary
  - Information content
  - Entropy, relative entropy (= KL divergence), mutual Information
  - Lossless compression, Lempel-Ziv and Huffman compression (entropy encoding algorithm)
  - Shannon's noiseless coding theorem
  - Tf-idf