INFORMATION THEORY AND PROBABILITIES

Hervé Blanchon
Laurent Besacier
Laboratoire LIG
Équipe GETALP
herve.blanchon@univ-grenoble-alpes.fr
laurent.besacier@univ-grenoble-alpes.fr

Language as data

- Large amounts of texts available in digital form
- Billions of documents available on the Web
- Tens of thousands of annotated sentences (syntax trees)
- Hundred million words translated between English and other languages

Statistics on the novel "Tom Sawyer" by M. Twain

Count Words

Function words at the top

Exception: **Tom**

word tokens: 73,370

word types: 8,018

tokens to types ratio: 8.9

Word	Count	Use
the	3332	determiner (article)
and	2973	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Distributions

- 3993 singletons
  - Most words appear somewhat rarely
- The main part of the text corresponds to the hundred most frequent words

Count	Count of count
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
> 100	102

Zipf's Law

George Kingsley Zipf
(1902–1950)

- Law : $f \times r = k$

Rank r	Word	Count f	$f \times r$
1	the	3332	3332
2	and	2973	5944
3	a	1775	5235
10	he	887	8770
20	but	410	8400
30	be	294	8820
100	two	104	10400
1000	family	8	8000
8000	applausive	1	8000

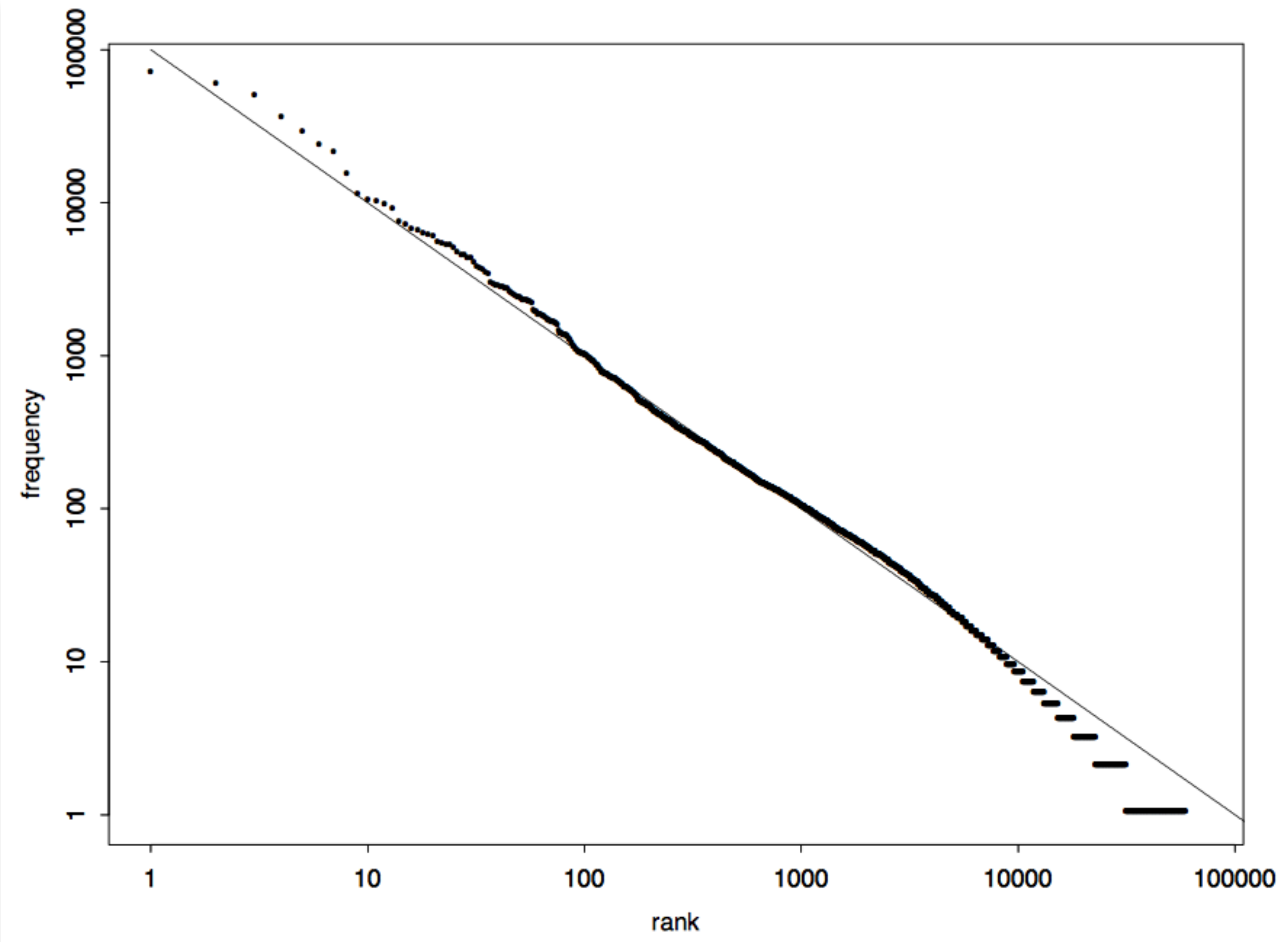
Zipf's Law for the Brown corpus

Brown Univ. Standard Corpus of Present-Day American English

Compiled
in the 60s

1M words

k = 100,000



https://en.wikipedia.org/wiki/Brown_Corpus

Probability distribution

- The probability (distribution) $p(w)$ of a word w in a corpus with s distinct words is:

$$p(w) = \frac{\text{count}(w)}{\sum_{i=1}^{s} \text{count}(w_i)}$$

- This estimation is referred to as "maximum likelihood"
- Distribution which answers the question:
  - "If I select randomly a word from a text, what is the probability that this word is the word w?"

Formalization

- Let W be a random variable
- We define the probability distribution p,
  - which indicates how likely the variable W takes the 'value' w ("is the word w")

$$prob(W = w) = p(w)$$

Joint Probability

- Goal
  - Study of two random variables at the same time
  - Example:
    - the words w_1 and w_2 that appear one after the other (a bigram), we model this with the distribution $p(w_1, w_2)$
    - If the occurrence of two words in bigrams is independent, we can write:
      - $p(w_1, w_2) = p(w_1)p(w_2)$, this assumption is probably wrong!
- Estimating the joint probability of two variables:
  - the same way this is done for a single variable

$$p(w_1, w_2) = \frac{\text{count}(w_1, w_2)}{\sum_{w_1', w_2'} \text{count}(w_1', w_2')}$$

Conditional probability

- Written $p(w_2|w_1)$
- Goal
  - answer the question: if the random variable $W_1 = w_1$, what is the probability that the variable W_2 takes the 'value' w_2
- Mathematically: $p(w_2|w_1) = \frac{p(w_1,w_2)}{p(w_1)}$
  - $p(w_1, w_2)$ *joint probability*
- Note
  - if W_1 and W_2 are independent then $p(w_2|w_1) = p(w_2)$

Rule 1: "Chain rule"

- We have
  - $p(w_1, w_2) = p(w_1)p(w_2|w_1)$
  - $p(w_1, w_2, w_3) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)$
  - etc.

Rule 2: "Bayes rule"

- The rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- Obtained from:

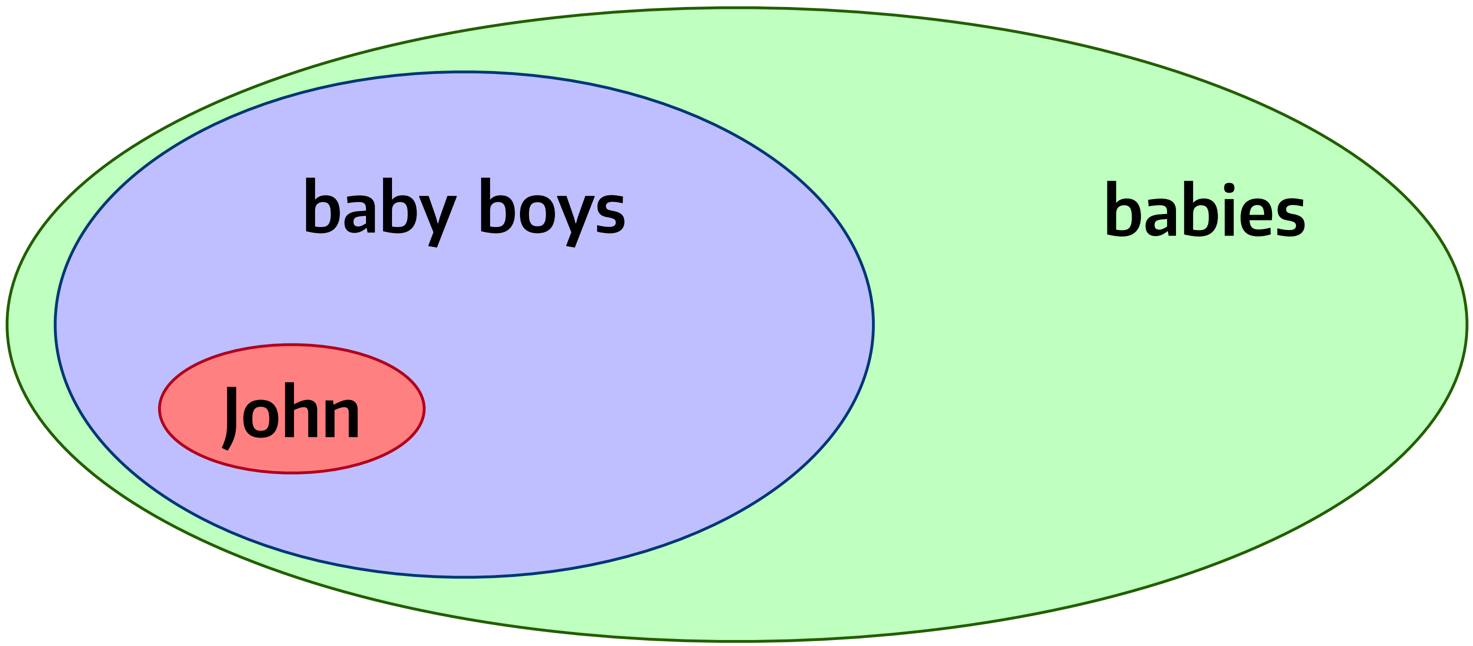
$$p(x,y) = p(y,x)$$

$$p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

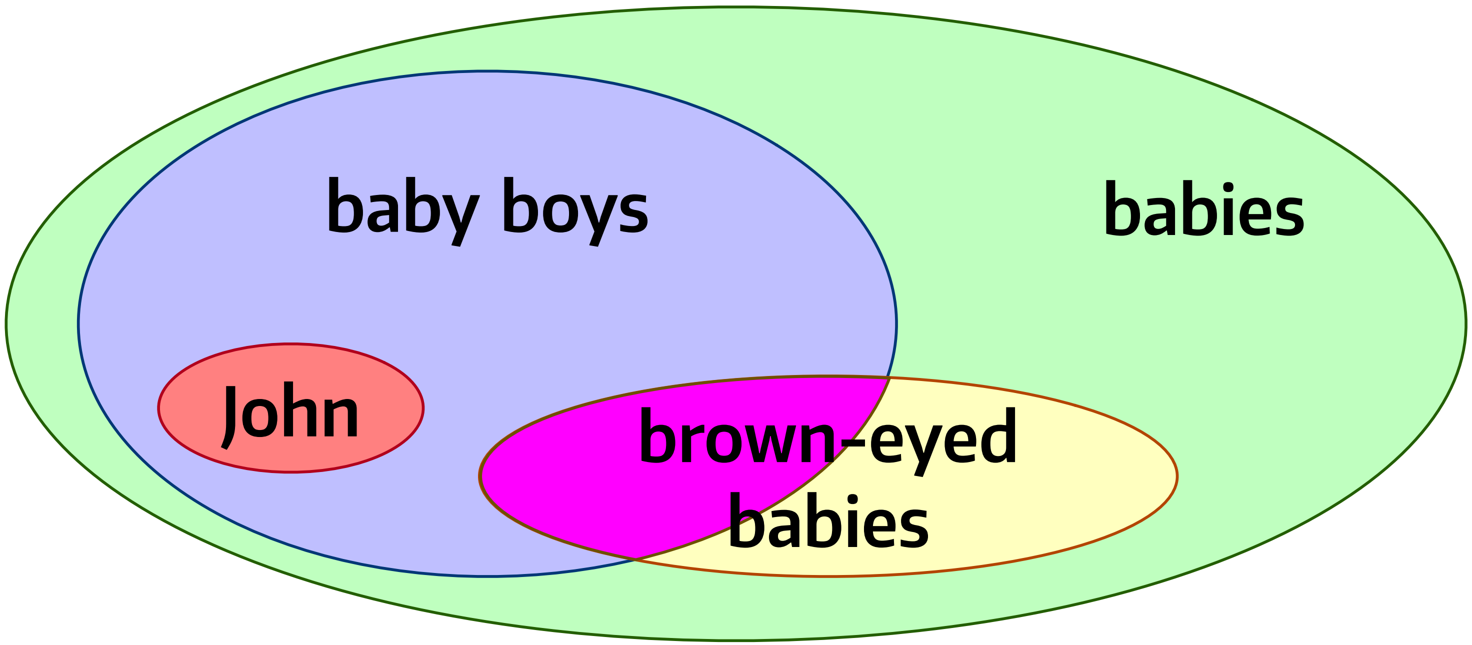
In other words…

- $P(X)$ means "probability that X is true"
  - $P(baby\ is\ a\ boy) \cong 0.5$
    - % of total that are boys
  - $P(baby\ is\ named\ John) \cong 0.001$
    - % of total named John



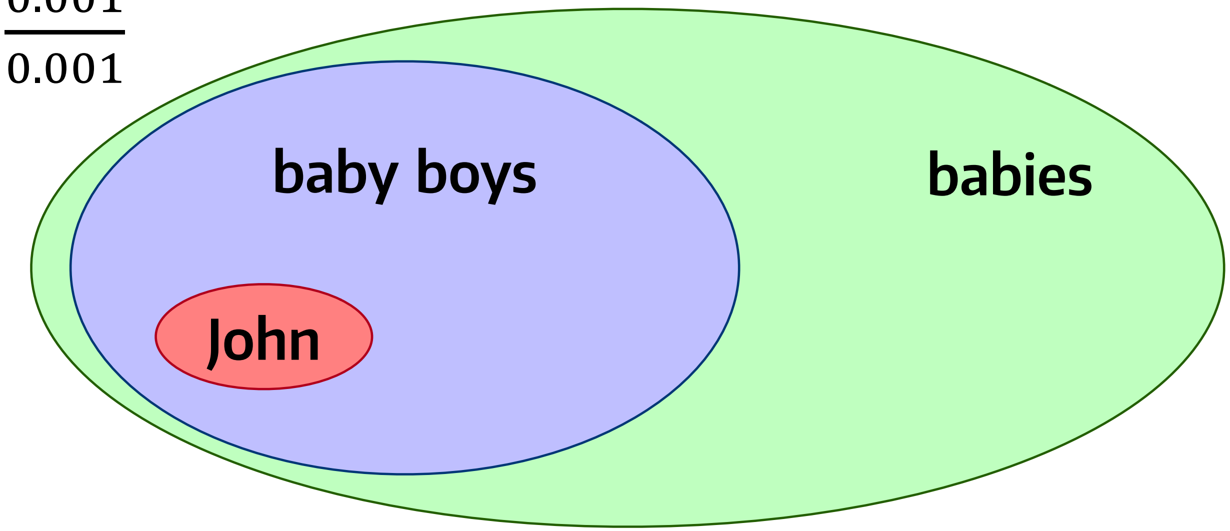
In other words…

- $P(X, Y)$ means "probability that X and Y are both true"
  - Size of $X \cap Y$ relative to Ω
- $P(brown{-}eyed\ baby, baby\ boy)$
  - Size of $brown{-}eyed\ baby \cap baby\ boy$ relative to $babies$



In other words…

- $P(X|Y)$ means "probability that X is true while Y is true"
  - Size of $X \cap Y$ relative to Y
- $P(baby\ is\ named\ John \mid baby\ is\ a\ boy) = 0.002$
  - $\frac{p(john, boy)}{p(boy)} = \frac{0.001}{0.5}$
- $P(baby\ is\ a\ boy \mid baby\ is\ named\ John) = 1$
  - $\frac{p(john, boy)}{p(john)} = \frac{0.001}{0.001}$



Expectation

- Informal definition
  - the expected value of a random variable is intuitively the long-run mean or average value of repetitions of the experiment it represents
- Expectation of a random variables X
  - a set of values $x_1, x_2, \ldots, x_n$
  - a probability $p(x_i), \forall i \in [1..n]$

$$E(X) = \sum_{i=1}^{n} p(x_i) x_i$$

- Example: a dice
  - 6 equiprobable (1/6) resting positions (1, 2, ... 6)
  - $E(dice) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 \frac{1}{+6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5$

Variance

- Variance (Var)
  - the expectation of the squared deviation of a random variable from its mean
  - measures how far a set of (random) numbers are spread out from their average value
  - $Var(X) = E\left((X - E(X))^2\right) = E(X^2) - E(X)^2$
  - For a discrete random variable X: $x_1 \longmapsto p_1, \ldots, x_n \longmapsto p_n$
  - $Var(X) = \sum_{i=1}^{n} p(x_i).(x_i - E(X))^2$
- Standard deviation (σ)
  - quantify the amount of variation or dispersion of a set of data values
  - Low: points close to the mean (expected value)
  - High: points spread out over wider range of values
  - $\sigma^2 = Var(X)$

Variance

- Example with the dice

$$
\begin{aligned}
Var(X) &= \frac{1}{6}(1-3.5)^2 + \frac{1}{6}(2-3.5)^2 + \frac{1}{6}(3-3.5)^2 \\
&\quad + \frac{1}{6}(4-3.5)^2 + \frac{1}{6}(5-3.5)^2 + \frac{1}{6}(6-3.5)^2 \\
&= \frac{1}{6}\left((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2\right) \\
&= \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \\
&= 2.917
\end{aligned}
$$

Distributions

- Uniform
  - All events are equiprobable
  - $p(x) = p(y)$ for all x, y

- Binomial
  - a series of trials with binary output (eg success / failure) with probability p of success
  - the probability of k successes in n trials is given by the probability mass function:
  - $Pr(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$
  - with $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ *(binomial coefficient)*

Bayesian estimation

- Model M, Data D
  - What is the most likely model given the data? => $p(M|D)$
  - $p(M|D) = \frac{p(D|M)\,p(M)}{p(D)}$
  - $argmax_M\, p(M|D) = argmax_M\, p(D|M)\, p(M)$
  - with
    - $p(M)$: a priori probability of the model
    - the estimation of a model $p(w)$ with the frequencies of words corresponds to a Bayesian estimation with a uniform prior probability (estimated by maximum likelihood)

Entropy

- Important concept that measures the "degree of disorder"
  - $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$

- Examples
  - 1 event: $p(a) = 1$
    - $H(X) = 0 = -1\log1$
  - 2 equiprobable events: $p(a) = 0.5, p(b) = 0.5$
    - $H(X) = 1 = -0.5\log0.5 - 0.5\log0.5$
  - 4 equiprobable events:
    - $H(X) = 2$

Entropy

- 4 events with a more likely than other
  - $p(a) = 0.7, p(b) = 0.1, p(c) = 0.1, p(d) = 0.1$

$$
\begin{aligned}
H(\mathrm{X}) \quad &= \quad -0.7\log_2 0{,}7 - 0.1\log_2 0.1 \\
&\qquad -0.1\log_2 0.1 - 0.1\log_2 0.1 \\
&= \quad -0.7\log_2 0.7 - 0.3\log_2 0.1 \\
&= \quad -0.7\times-0.5146 - 0.3\times-3.3219 \\
&= \quad 0.36020 + 0.99658 \\
&= \quad 1.35678
\end{aligned}
$$

Entropy

- Intuition:
  - a good model should have a low entropy ...

- Many probabilistic models in language processing lead to a reduction of entropy

Information theory and entropy

- Suppose we want to encode a sequence of events X
- Each event is encoded by a sequence of bits
- Examples
  - Coin: $a = 0, b = 1$
  - Four equiprobable events: $a = 00, b = 01, c = 10, d = 11$
  - Huffman coding (less bits for more frequent letter)
- The number of bits needed to encode the events of X is greater than or equal to the entropy of X

References

- Manning and Schutze: “Foundations of Statistical Language Processing”, 1999 , MIT Press, available online
- Jurafsky and Martin: “Speech and Language Processing”, 2000, Prentice Hall.
- Rajman M. “Speech and language engineering”, 2007, EPFL Press.