Text properties

Text

- Information Theory
  - Amount of information is related to the distribution of symbols in the document.
  - Entropy: $E = -\sum_{i=1}^{\sigma} p_i \log_2 p_i$
  - where σ is the number of symbols of the alphabet
  - Definition of entropy depends on the probabilities of each symbol.
  - Text models are used to obtain those probabilites
  - Example - Entropy
    - 001001011011 $E = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$
    - 111111111111 $E = -\left(0\log_2(0) + 1\log_2(1)\right) = 0$

Text

- Modeling Natural Language
  - Symbols: separate words or belong to words
  - Symbols are not uniformly distributed
    - binomial model: each symbol is generated with a certain probability
  - Dependency of previous symbols
    - *k*-order markovian model: the propability of a symbol depends on the previous k symbols
  - Alternatively, we can take words as symbols

Text

- Modeling Natural Language
  - Words distribution inside documents:
    - Zipf´s Law: *i*-th most frequent word appears $1/i^\theta$ times of the most frequent word

$$f_i = \frac{n}{i^\theta H_{|V|}(\theta)}$$

$$H_{|V|}(\theta) = \sum_{j=1}^{|V|} \frac{1}{j^\theta}$$

  - where f_i is the number of times the i-most frequent word appears in a text of n words and V is the vocabulary of the text, i.e. the set of different words in the text
  - The distribution of words is very skewed: a few hundred words take up 50% of the text (stopwords)
  - Real data fits better with θ between 1.5 and 2.0

Text

- Modeling Natural Language
  - Example - word distibution (Zipf's Law)
    - |V|=1000, θ = 2
    - most frequent word: n=300
    - 2nd most frequent: n=76
    - 3rd most frequent: n=33
    - 4th most frequent: n=19

Text

- Modeling Natural Language
  - Distribution of words in the documents
    - negative binomial distribution

$$F(k) = \binom{\alpha + k - 1}{k} p^k (1 + p)^{-\alpha - k}$$

    - where F(k) is the fraction of documents containing the word k and p, a are parameters that depend on the word and the document collection.
  - Number of distinct words (vocabulary) in a text of n words:
    - Heaps' Law: $V = Kn^{\beta}$
  - K is between 10 and 100
  - β is between 0,4 and 0,6
  - example: n=400000, β = 0.5
    - K=25, V=15811
    - K=35, V=22135
  - The set of different words of a language is fixed by a constant, but the limit is too high