Statistical Design Considerations for Lexicon-Based Information Extraction

By Mireya Diaz, PhD

Department of Biomedical Sciences, Division of Epidemiology and Biostatistics

- Large proportion of information extraction (IE) via natural language processing (NLP) is based on lexicons
- Design considerations for a useful lexicon are:
  Lexicon size, document's word length, prevalence of non-stop words, unique tokens (lemmas) distribution, words' length distribution

Probability that sample will contain one or more instances of the token of interest follows a Poisson distribution[1,2]:

$p(x > 0) = 1 - e^{-\lambda}$ Λ: mean rate of token's occurrence

For each token: $\lambda_i = f_i/(N/n)$ f_i: frequency of token i in the corpus

N: size of corpus n: size of sample

Juckett[2] worked further the Poisson arrival of tokens within a sample focusing in the capture probability for each token given its length providing the following formulae:

$$P = \frac{\sum_{j=W_s}^{W_l} b_j P_j}{\sum_{j=W_s}^{W_l} b_j}$$

P: aggregate capture probability across all token sizes
b_j: overall frequency of tokens of size j
W_s, W_l: smallest and largest token sizes chosen

$$P_j = \sum_{i=1}^{m_j} a_{i,j}\,(1 - e^{-\lambda}) \quad \text{with} \quad \sum_{i=1}^{m_j} a_{i,j} = 1 \quad \text{for each value of } j$$

Example: for 1 token/10 documents and a desired capture probability of 0.95 it would be required 30 documents to extract the token

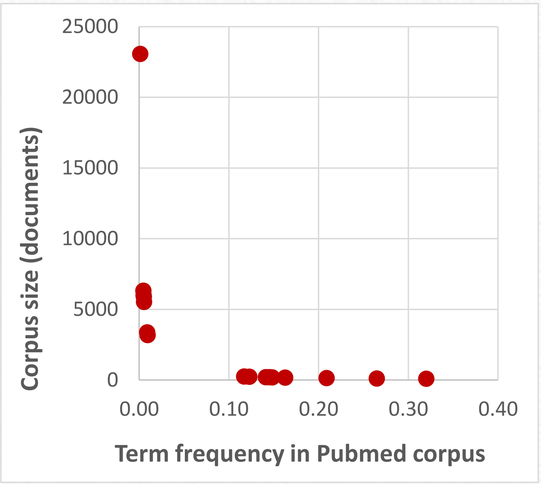
Example: would like to build a corpus for words in the toxic substances and disease registry

Glossary from www.atsdr.cdc.gov contains 166 terms (75 unigrams -20 are acronyms, 52 bigrams, 26 trigrams, 13 n-grams). Frequency of these terms in Pubmed documents were obtained (displayed in table below under terms and average probability columns). For comparison purposes: frequency of words from the Thorndike and Lorge English list as extracted by EllegaRd[1].

Freq level	Terms	Average probability	Words[1]	Average probability[1]
10^1	1	1.8×10^{-1}	-	-
10^2	24	3.8×10^{-2}	12	1.33×10^{-2}
10^3	29	3.5×10^{-3}	109	2.67×10^{-3}
10^4	35	4.0×10^{-4}	969	2.58×10^{-4}
10^5	37	3.9×10^{-5}	4,810	3.08×10^{-5}
10^6	19	2.8×10^{-6}	11,100	3.47×10^{-6}
10^7	11	4.5×10^{-7}	16,000	3.71×10^{-7}
10^8	9	6.0×10^{-8}	20,000	3.95×10^{-8}

- Over 23K documents need to be extracted to build the desired corpus if all 166 terms are sought.
- If 5 terms with the smallest frequency are excluded, then only 6,329 documents would be necessary to obtain the other terms with a frequency of 0.047% or higher, with a 0.95 capture probability.

Corpus size for terms depending on their frequency and word length



References

1. EllegåRd A. Estimating vocabulary size. WORD 1960;16:2:219-244.
2. Juckett D. A method for determining the number of documents needed for a gold standard corpus. J Biomed Informatics 2012;45:460-470.