BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 2, Number 372 (2018), 120 – 125

UDC 004.056.5

A. A. Zhatkanbayev

Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: wildlife.kz@gmail.com

THE APPLIANCE OF Natural Language Processing TOOLS FOR IMPLEMENTATION OF STATISTICAL ANALYZER FOR PAGES WITHIN SECURE web SEARCHING

Abstract. The complex of program based on Natural Language Processing (NLP) tools and statistical algorithms for text analysis are described. Implemented web analyzer for pages can be easily embedded on server side of Internet Server Provider (ISP) for processing query of users, so that in consequence to limit access on inappropriate web-sites. The large number of remote custom parameters allows to increase, or decrease sensitivity of searching process. Security of web space is provision of parental control tools, and web analyzer for secure searching. In order to solve powerful statistical tools was implemented, substring searching algorithms, and NLP toolkit for analysis of input texts.

Key words: Natural Language Processing, Fourier distribution, Dirichlet distribution, Back-off smoothing, Knuth Morris Pratt O(nm)/O(n+m) algorithms (KMP), searching.

Introduction. Information security is not only complex of cryptographic problems solutions for providing confidentiality, integrity of data for their protection from third non-authorized parties, but it also security of Internet space including development of parental control tools, web analyzer for secure search of pages in the network. For solution of that problems it is required implementation of powerful statistical tools, substring searching algorithms, toolkit of Natural Language Processing (NLP) for analysis of input texts. Before implementation of program's complex there are studied the materials on theory of probability were studied, in particular probabilistic models, methodology of constructing probability of word occurrences in the text, Fourier probability (Fourier distribution), Dirichlet distribution, Back-off smoothing, Bernoulli distribution and also substring searching algorithms.

Formulation of problem. The purpose of project is programmatically implement high speed Natural Language Processing toolkit for the analysis of input text (for example, from web-pages) on malicious information contents upon using follows: Fourier Probability (Fourier Distribution), Dirichlet Distribution, Back-off smoothing, Frequency (count of each n-gram occurrences into input text), POS Tagging (Knuth Morris Pratt O(nm)/O(n+m) prefix-function algorithms), Levenshtein Distance for text correction, PorterStemmer (for processing plural forms of words in single form, base of words (with/without prefixes) with using prefix function of Knuth Morris Pratt O(nm)/O(n+m) algorithms). Program complex would allow using as statistical methods and toolkit of Natural Language Processing to reach high quality text analysis on presence of words set from restricted category, additionally following complex of programs would let to achieve high quality implementation of web analyzers (web filters) and function of parental control. Most of antivirus product solutions using in their analyzers information about site with the helps of polls, complaints from users. Such approach does not guarantees filtering of all unwanted for viewing web-pages and even more so could block legitimate resources. Developed program complex with Natural Language Processing toolkit would allow to achieve precise analysis of text and to make work of web-pages analyzer more qualify. Research work in following area is important for information security as high-performance tool of web-pages processing and tools of parental control.

Constructing mathematical model for research object. Fourier probability (Fourier distribution). Fourier probability is used to calculate probability of occurrence n-gram (often two grams, one gram) at predetermined text. Statistical appliance of Fourier probability could be used to measure frequency of n-grams (higher the probability of n-gram than more words occurring or repeated in the text). Fourier probability can be obtained for each n-grams in text or for the defined n-gram [1].

Fourier probability calculating by following formula:

$$P(w_i|w_i-1) = c(w_{i-1}w_i)/Ew_i c(w_i-1w_i)\,,$$

let the $P(w_i|w_i-1)$ be probability of two-gram occurrence than $P(w_i)$ would be probability of appearing cut one gram and $P(w_{i-1})$ probability of appearing wrote in reverse order two gram.

Example: Matt Jarvis headed the Hammers in front as they threatened to extend Arsenal's winless league run to five games. But Podolski leveled with a shot on the turn two minutes later for the impressive FA Cup finalists. Olivier Giroud's classy finish and Podolski's driven second sealed the win as Arsenal moved up to fourth. Impressive goal have been made on the ending of a second time. However the Cup finalists where not, yet determined, but of course the main competitor for Cup semi-final is still be determined in few weeks.

$$P(w_i|w_i-1) = c(w_{i-1}w_i)/Ew_i c(w_i-1w_i)$$
$$P(Cup\ finalists) = p(Cup|.)\ p(finalists\ Cup)\ p(.|finalists) =$$
$$c(.Cup)/Ew_c(.w)c(Cup\ Finalists)/Ew_c(Cup\ w)\ c(finalists\ .)/Ew_c(finalists\ w),$$

where $c(.Cup)$ - count of how many times word 'Cup' meets in the text. $Ew_c(.w)$ - total amount of sentences. $c(Cup\ Finalists)$ - count of how many times strict context 'Cup finalists' occurs in text. 'Cup', 'finalists', 'finalists Cup' are not considered. $Ew_c(Cup\ w)$ - count of how many n-bigrams meets (bigrams, trigrams and so on). In other words, count how many times structure 'Cup' + any word occurs. $c(finalists.)$ - Count how many times meets only the 'finalists'. $Ew_c(finalists\ w)$ - count of amount n-grams finalists + any word occurs. $P(Cup\ finalists)$=1/5*2/3*2/1=0.2*0.67*2=0.268.

It should be noted that in Fourier probability probability of n-gram could be higher than 1 (not as in base probability theory where it must be not higher than 1 or where it must be not negative). High value of Fourier probability for n-gram means that distribution (amount of occurrences in the text) for that particular n gram is high [2].

Dirichlet distribution. In comparison with Fourier distributions which is precise for calculation of n-gram occurrence (two grams or higher) [3]. Dirichlet distribution more applicable for one-grams and requires less computational operations and not requires reverse $P(w_i|w_i-1)$ operations. For example, $c(.Cup)$ - count of how many times word 'Cup' meets in text, $Ew_c(Cup\ w)$ - count of how many times n-bigrams are meet (bigrams, trigrams and so on). By other words it is required to compute how many times structure 'Cup' + any word meets. Ew_c – full probability beginning with counting when first coincidence meets, $E\ P\big(k, B(Betta)\big) = B_k$, where E is sum. As with the case of Fourier probability in that occasion Dirichlet distribution differs from common probabilities where probability can be not higher than one. Note that in Dirichlet distribution probability of n-gram can be higher than 1 (not as in common probability theory where probability can be not higher than 1 or can be not negative). More higher complete probability of Dirichlet distribution for n-gram means that distribution - occurrence in the text for following particular n-gram is high [4,5].

Dirichlet distribution allows to store all events and correlate their outputs (probabilities) to k.

$$P\big(k, B(Betta)\big) = B_k,$$

where P is probability and for all k (amount of all events from 1…….. k), B (probability of each event), B_k (probability of all events from B_1……B_k (for all B_{k0}).

Example: Manchester City's players are the best paid in world sport according to a survey by Sporting intelligence. The City first team receive an average annual wage of 5.337M GBP a year-equivalent of 102.634 GBP a week. That is slightly more than their counterparts at the New York Yankees