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:

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

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.

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

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 and LA Dodgers baseball teams. slightly there is a possibility that Manchester players will be going to next season.

$$P(Manchester) = p(Manchester|.) = \frac{c(.Manc\,hester\,)}{Ew_c(.w)} = 2/4,$$

where $c(.\,Manchester)$ – amount of occurrences 'Manchester', $Ew_c(.\,w)$ - full amount of sentences. In Dirichlet distributions $E\;P\big(k, B(Betta)\big) = E_{p(k,B)} = B_k$, considering E – is sum, $P\big(k, B(Betta)\big)$ – probability of word $B(Betta)$ in k sentence, $E_{p\,(k,B)} = B_k$ - probability of word $B(Betta)$ in k sentence.

Back-off smoothing. Backoff smoothing is the process of adding artificial probability for defined n-gram. If the frequency for particular n-gram is high in text except others n-grams, than its probability significantly less than that word. Than in following cased it is required to apply Bernoulli distributions or Backoff smoothing.

$$\text{For all } \frac{Ew_c(w_i|w_i-1)}{P(w_i|w_i-1)} \text{ и } P\big(k, B(Betta)\big) = B_k$$

Dirichlet distributions a +=~0.90 (or at least 0.00001) or higher values must be added in order to artificially increase probability, but this is conduction in case if probability of word aims to almost 0, for example 0.00001.

Example: Mat Jarvis headed the arsenals cup hammers in front as they threatened finalists to extend Arsenals winless cup finalists cup finalists finalists cup league run to five games. but Podlski.

As it was seen from other examples probability of word 'cup' and 'cup finalists' very high as the frequency of 'cup', 'cup finalists' higher than others n-gram words have. It means occurs more often - appears twice as others n-grams.

$$P(Cup\,)P(Cup\,) = p(Cup|.)\,p(.\,|Cup) = \frac{\dfrac{\dfrac{c(.\,Cup)}{Ew_c(.\,w)c(Cup.\,)}}{Ew_c(.\,w)}c(cup\,.\,)}{Ew_c(cup\;w)} = 0.140625$$

$$P(Finalists\,)P(Finalists\,) = p(Finalists|.)\,p(.\,|Finalists) = \frac{\dfrac{\dfrac{c(.\,Finalists)}{Ew_c(.\,w)c(Finalists.\,)}}{Ew_c(.\,w)}c(Finalists\,.\,)}{Ew_c(Finalists\;w)}$$
$$= 1.30645161290323$$

$$P(Cup\;Finalists\,) = p(Cup|.)\,p(.\,|Finalists\;Cup)P(.\,Finalists) =$$
$$= \frac{\dfrac{\dfrac{c(.\,Cup)}{Ew_c(.\,w)c(Cup\;Finalists.\,)}}{Ew_c(Cup\;w)}c(Finalists\,.\,)}{Ew_c(Finalists\;w)}\;P(Cup\;Finalists)$$
$$= 0.0634920634920635$$

Obviously see probability of $P(Cup\,), P(Finalists\,), P(Cup\;Finalists)$ significantly higher than $P(run)$, by her own, $P(run)$ is almost 0, $P(run)$ =0.0714285714285714.

Therefore, Back-off smoothing must be added in order to make its probability on the level of $P(Cup\,), P(Finalists\,), P(Cup\;Finalists)$.

Results of Back-off smoothing (results of implemented solution) are follows. As obviously see that after appliance of Back-off smoothing $P(run)$ now appear on the level of $P(Cup\,), P(Finalists\,), P(Cup\;Finalists)$.

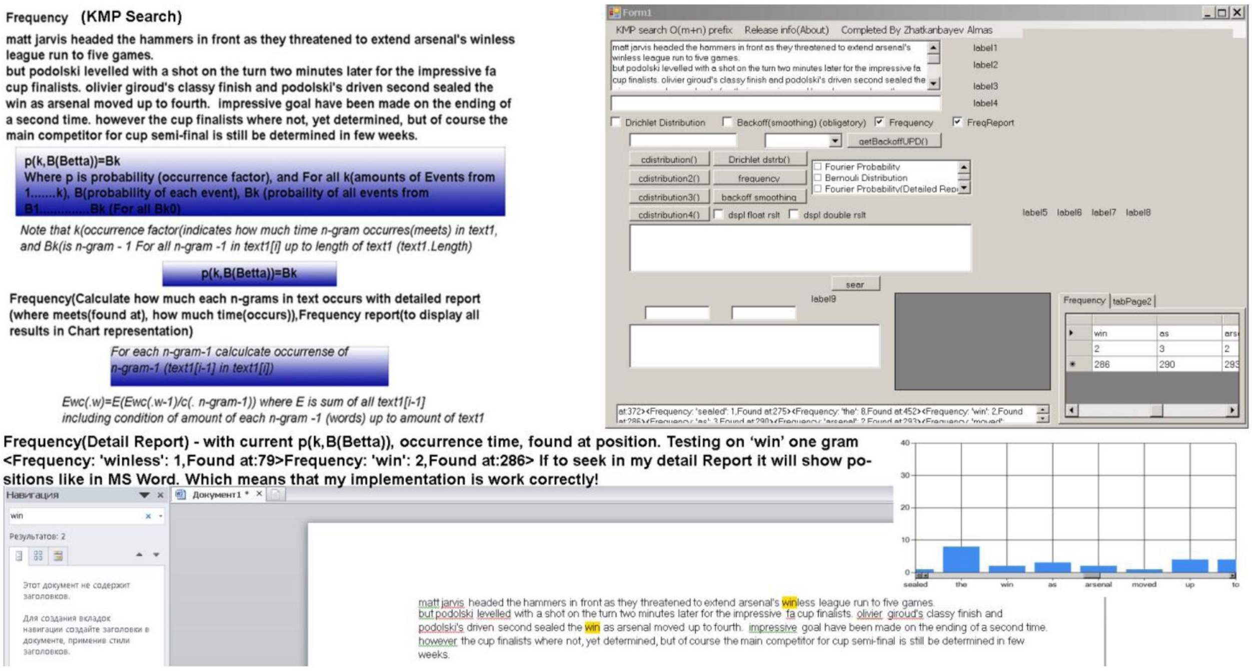
$P(run)$ result without appliance of Back-off smoothing:

$$P(Run) = p(Run|.)\,p(.\,|Run) \;= \frac{\dfrac{\dfrac{c(.\,Run)}{Ew_c(.\,w)c(Run.)}}{Ew_c(.\,w)}c(Run\,.)}{Ew_c(Run\;w)} = 0.0714285714285714$$

$P(run)$ result with appliance of Back-off smoothing:

$$P(Run) = p(Run|.)\,p(.\,|Run) \;= \frac{\dfrac{\dfrac{c(.\,Run)}{Ew_c(.\,w)c(Run.)}}{Ew_c(.\,w)}c(Run\,.)}{Ew_c(Run\;w)} = 0.145161290322581$$

Receiving of theoretical and appliance results with usage of computer technologies.



Frequency (KMP Search)

matt jarvis headed the hammers in front as they threatened to extend arsenal's winless league run to five games.
but podolski levelled 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(k,B(Betta))=Bk
Where p is probability (occurrence factor), and For all k(amounts of Events from 1.......k), B(probability of each event), Bk (probaility of all events from B1..............Bk (For all Bk0)

Note that k(occurrence factor(indicates how much time n-gram occurres(meets) in text1, and Bk(is n-gram - 1 For all n-gram -1 in text1[i] up to length of text1 (text1.Length)

p(k,B(Betta))=Bk

Frequency(Calculate how much each n-grams in text occurs with detailed report (where meets(found at), how much time(occurs)),Frequency report(to display all results in Chart representation)

For each n-gram-1 calculcate occurrense of n-gram-1 (text1[i-1] in text1[i])

Ewc(.w)=E(Ewc(.w-1)/c(. n-gram-1)) where E is sum of all text1[i-1] including condition of amount of each n-gram -1 (words) up to amount of text1

Form1

KMP search O(m+n) prefix Release info(About) Completed By Zhatkanbayev Almas

matt jarvis headed the hammers in front as they threatened to extend arsenal's winless league run to five games.
but podolski levelled 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

□ Drichlet Distribution □ Backoff(smoothing) (obligatory) ☑ Frequency ☑ FreqReport

□ Fourier Probability
□ Bernouli Distribution
□ Fourier Probability(Detailed Rep

□ dspl float rslt □ dspl double rslt

	win	as	ars
	2	3	2
*	286	290	293

at:372><Frequency: 'sealed': 1,Found at:275><Frequency: 'the': 8,Found at:452><Frequency: 'win': 2,Found at:286><Frequency: 'as': 3,Found at:290><Frequency: 'arsenal': 2,Found at:293><Frequency: 'moved':

Frequency(Detail Report) - with current p(k,B(Betta)), occurrence time, found at position. Testing on 'win' one gram <Frequency: 'winless': 1,Found at:79>Frequency: 'win': 2,Found at:286> If to seek in my detail Report it will show positions like in MS Word. Which means that my implementation is work correctly!

Навигация

win

Результатов: 2

Этот документ не содержит заголовков.

Для создания вкладок навигации создайте заголовки в документе, применив стили заголовков.

matt jarvis headed the hammers in front as they threatened to extend arsenal's winless league run to five games.
but podolski levelled 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.

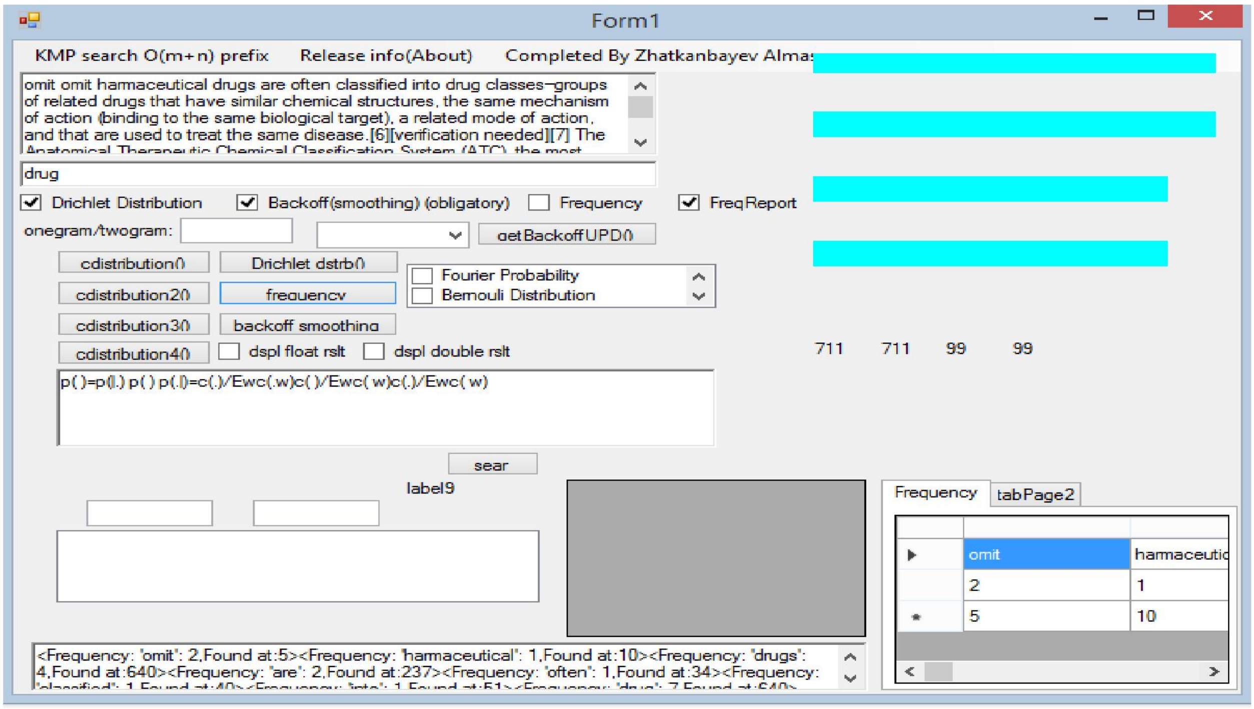
Figure 1 – Statistical program complex for calculation Dirichlet Distribution, Back-off smoothing, frequencies, KMP substring search algorithm

Measurement of time between implementations of KMP O(N+M), KMP O(NM) during prefix-function calculation, KMP substring search algorithm on length of text equals to 192 symbols

№	KMP version	Time processing
1	KMP O(nm)	01M:57C:30MC (~117 c)
2	KMP O(n+m)	0.01MC

Conclusion. Considering that all complex of programs based on native implementation and speed of text processing is sufficiently high, thus following web-analyzer of pages can be easily applied on the server side of Internet service provider for processing requests of users so that to do limit access on inappropriate sites finally. Also, high amount of adjustable parameters would allow to increase, or decrease search threshold sensitivity level. Example of frequencies output of all words in the sentences from all or part of text collected about drugs article is a Wikipedia online encyclopedia. Considering educational context of article, it is seen that frequencies of word drugs not a high.

Example: Pharmaceutical drugs are often classified into drug classes – groups of related drugs that have similar chemical structures, the same mechanism of action (binding to the same biological target), a related mode of action, and that are used to treat the same disease. The Anatomical Therapeutic Chemical Classification System (ATC), the most widely used drug classification system, assigns drugs a unique ATC code, which is an alphanumeric code that assigns it to specific drug classes within the ATC system. Another major classification system is the Bio-pharmaceutical Classification System. This classifies drugs according to their solubility and permeability or absorption properties.



omit omit harmaceutical drugs are often classified into drug classes–groups of related drugs that have similar chemical structures, the same mechanism of action (binding to the same biological target), a related mode of action, and that are used to treat the same disease.[6][verification needed][7] The Anatomical Therapeutic Chemical Classification System (ATC), the most

drug

☑ Drichlet Distribution ☑ Backoff(smoothing) (obligatory) □ Frequency ☑ FreqReport

onegram/twogram:

□ Fourier Probability
□ Bernouli Distribution

□ dspl float rslt □ dspl double rslt

711 711 99 99

```
p( )=p(|.) p( ) p(.|)=c(.)/Ewc(.w)c( )/Ewc( w)c(.)/Ewc( w)
```

▶	omit	harmaceutic
	2	1
*	5	10

```
<Frequency: 'omit': 2,Found at:5><Frequency: 'harmaceutical': 1,Found at:10><Frequency: 'drugs':
4,Found at:640><Frequency: 'are': 2,Found at:237><Frequency: 'often': 1,Found at:34><Frequency:
'classified': 1,Found at:40><Frequency: 'into': 1,Found at:51><Frequency: 'drug': 7,Found at:640>
```

Figure 2 – Form 1 of statistical program complex for calculation Dirichlet Distribution, Back-off smoothing, frequencies, KMP substring search algorithm

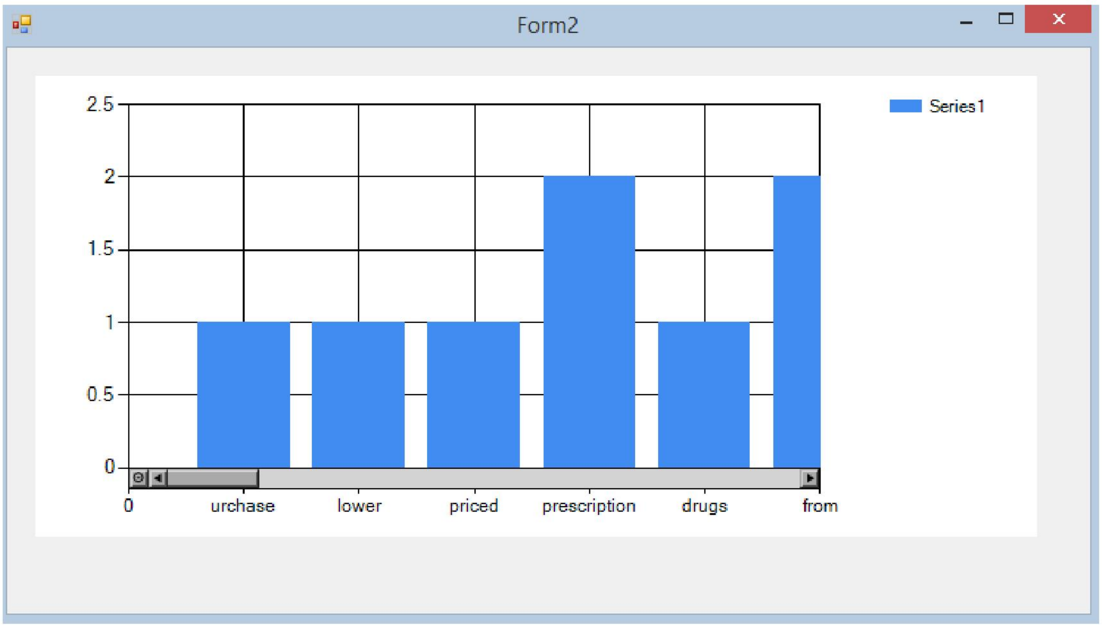


Figure 3 – Form 2 of statistical program complex for calculation Dirichlet Distribution, Back-off smoothing, frequencies, KMP substring search algorithm

REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan, Heinrich Schutze. Introduction to information search. M.: Williams, 2011. 528 pp. (In Russian).

[2] Gavrilova T.A. Knowledge bases of intellectual systems. SPb.: Peter, 2000. 384 p. (In Russian).

[3] Bavrin I.I. Theory of Probability and Mathematical Statistics. M.: Vyssh. shk., 2005. 160 p. (In Russian).

[4] Vilenkin N.Ya. Combinatorics. M.: FIMA, MCNMO, 2006. 400 p. (In Russian).

[5] Gmurman V.E. Theory of Probability and Mathematical Statistics: A Tutorial for Bachelors. M.: Yurayt, 2013. 479 p. (In Russian).

А. А. Жатқанбаев

Казахский национальный университет им. аль-Фараби, Алматы, Казахстан

ПРИМЕНЕНИЕ СРЕДСТВ Natural Language Processing ДЛЯ РЕАЛИЗАЦИИ СТАТИСТИЧЕСКОГО АНАЛИЗАТОРА СТРАНИЦ ДЛЯ БЕЗОПАСНОГО веб-ПОИСКА

Аннотация. В статье описывается комплекс программ, основанный на инструментариях Обработки Естественного Языка (Natural Language Processing, NLP) и статистических алгоритмов для анализа текста. Реализованный веб-анализатор страниц можно легко перенести на серверную часть интернет провайдеров для обработки запросов пользователей, чтобы впоследствии ограничивать доступ на неприемлемые сайты. Большое количество настраиваемых параметров позволит увеличивать, снижать чувствительность поиска. Безопасность интернет пространства – это обеспечение средств родительского контроля, веб-анализатора для безопасного поиска. Для решения реализованы мощные статистические средства, алгоритмы поиска подстрок, инструментарии NLP для анализа входных текстов.

Ключевые слова: обработка Естественного Языка (Natural Language Processing, NLP), распределение Фурье, распределение Дирихле, Back-off сглаживание, Back-off smoothing, Knuth Morris Pratt O(nm)/O(n+m) algorithms (KMP), поиск.

А. А. Жатқанбаев

Әл-Фараби атындағы Қазақ ұлттық университетіб Алматы, Қазақстан

ҚАУІПСІЗ ВЕБ ІЗДЕУ ҮШІН СТАТИСТИКАЛЫҚ БЕТТІ ТАЛДАУ ҚҰРАЛЫН \ІСКЕ АСЫРУ ҮШІН Natural Language Processing ҚҰРАЛДАРЫН ПАЙДАЛАНУ

Аннотация. Мақалада Natural Language Processing (NLP) құралдарына негізделген бағдарламалардың жиынтығы және мәтінді талдау үшін статистикалық алгоритмдер сипатталған. Орындалған веб-парақ анализаторы, кейіннен қолайсыз сайттарға кіруді шектеу үшін пайдаланушы сұрауларын өңдеу үшін интернет-провайдерлердің серверлік бөлігіне оңай ауыса алады. Көптеген бапталатын параметрлер көбейтіледі, іздеудің сезімталдығын төмендетеді. Ғаламтор қауіпсіздігі – ата-ана бақылауы, қауіпсіз іздеу үшін веб-анализатор. Шешім үшін күшті статистикалық құралдар, субстраттар үшін іздеу алгоритмдері, кіріс мәтіндерін талдау үшін NLP құралдары іске асырылады.

Түйін сөздер: Natural Language Processing (NLP), Фурье таралымы, Дирихленің үлестірілуі, Back-off жалтырату, Knuth Morris Pratt O(nm)/O(n+m) алгоритмдер (KMP), іздеу.

Information about author:

Zhatkanbayev Almas Altayuly – bachelor of technique and technology by specialty 5B070400 «Computer systems and software» (Kazakh-British Technical University), master student of 2nd course of specialty 6M100200 «Systems of information security» of Kazakh National University after Al-Farabi. E-mail: wildlife.kz@gmail.com