**УДК 519.765**

# АНАЛИЗ ЛИТЕРАТУРНЫХ ПРОИЗВЕДЕНИЙ НА СОДЕРЖАНИЕ МНЕМОНИЧЕСКИХ ЦИТАТ ДЛЯ НОМЕРОВ

**Забайкин А.В.[1], Идрисов Р.И.[2]**

*[1]ФГБУН «Институт вычислительных технологий Сибирского отделения Российской академии наук», Новосибирск, Россия (630090, Новосибирск, пр. Академика Лаврентьева, 6), e-mail: ict@ict.nsc.ru*
*[2]ФГБУН «Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук», Новосибирск, Россия (630090, Новосибирск, пр. Академика Лаврентьева, 6), e-mail: iis@iis.nsk.su*

**В работе исследуется возможность автоматического сопоставления цифро-буквенных последовательностей к отрывку литературного произведения. Производится численный эксперимент методом Монте-Карло, сравниваются результаты подбора таких отрывков при помощи двух различных способов кодирования чисел (кодирование по первым буквам и кодирование по длинам). Для генерации мнемонических цитат используется приёмы мнемотехники — совокупности специальных приёмов и способов, облегчающих запоминание нужной информации и увеличивающих объём памяти путём образования ассоциаций[1]. В результате была показана логарифмическая зависимость между объёмом исследуемого текста и количеством автоматически сгенерированных цитат, разработано программное средство, реализующее предложенный алгоритм на цифро-буквенной последовательности длиной до 7 символов. Данное приложение может быть полезно для прикладных программных средств, помогающих запоминать номера телефонов, автомобильные номера, химические элементы, пароли, и.т.д.**

Ключевые слова: обработка естественного языка, анализ текста, мнемонические цитаты, мнемоника

# LUCUBRATIONS ANALYSIS FOR CONTAINING MNEMONIC QUOTES

**Zabaykin A.V.[1] , Idrisov R.I.[2]**

*[1] Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia (630090, Novosibirsk, 6 Acad. Lavrentjev avenue) e-mail: ict@ict.nsc.ru*
*[2] A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia (630090, Novosibirsk, 6 Acad. Lavrentjev avenue), iis@iis.nsk.su*

**We investigate the possibility of automatic matching alphanumeric sequences to the passage of a literary work . Numerical experiment is performed using the Monte Carlo method , compares the results of selection of such passages using two different methods of encoding numbers (encoding the first letters of coding and run-length) . To generate quotes mnemonic techniques used mnemonics - the collection of special techniques and methods that facilitate the memorization of the right information and increase the amount of memory by forming associations [1]. The result was shown a logarithmic relationship between the text of the test and the number of automatically generated quotations, developed a software tool that implements the algorithm for alphanumeric string of up to 7 characters. This application may be useful for applications that help to memorize phone numbers, license plates , chemicals , passwords , etc.**

Key words: natural language processing, text analysis, mnemonic quotes, mnemonics

Проблема поиска наиболее подходящего мнемонического[1] образа для запоминания широко рассмотрена в учебнике мнемотехники [1]. Автор в [1] рассматривает различные приёмы запоминания, эффективные как для связной информации, так и для несвязной, которой являются цепочки слов, чисел, карт, буквосочетаний. Тем не менее подбор образа

[1] Под мнемоникой или мнемотехникой мы будем понимать классической определение: мнемоника — совокупность специальных приёмов и способов, облегчающих запоминание нужной информации и увеличивающих объём памяти путём образования ассоциаций (связей). Замена абстрактных объектов и фактов на понятия и представления, имеющие визуальное, аудиальное или кинестетическое представление, связывание объектов с уже имеющейся информацией в памяти различных типов для упрощения запоминания.

предлагается произвести человеку "вручную". Мы же постарались использовать возможности подбора текста, которые довольно затруднительно использовать без помощи вычислительной системы.

Таким образом, сопоставление абстрактного набора цифр к отрывку текста наполняет его образами и позволяет упростить его запоминание. Приведем пример: если кодировать цифры согласными буквами, то каждое слово или предложение соответствует целому числу. Обычно выбирают следующий способ кодирования 1-р, 2-д, 3-т, 4-ч, 5-п, 6-ш, 7-с, 8-в, 9-м (потому что 9 это – “много”). Тогда слова “добрый мой приятель” соответствуют числу 219513. Но это несколько неудобно, поскольку без специальной подготовки не получается быстро выкинуть ненужные буквы, тем не менее, “добрый мой приятель” забыть довольно сложно, что всегда позволит вам находясь в спокойной обстановке вспомнить число 219513. И это весьма заманчиво, поскольку само по себе число это является весьма абстрактным и может запросто перепутаться с другими такими же абстрактными числами. Эти вещи хорошо известны и более подробное описание приёма можно найти в учебнике [1].

**Подход, основанный на кодирование по первым буквам слов.**

Сперва была опробована наиболее распространенная [4] классическая методика, основанная на кодировании по первым буквам. Первоначальная идея заключалась в том, чтобы попробовать подобрать цитату из стихотворной части школьной программы, которая бы соответствовала заданному автомобильному номеру, то есть некоторой случайной последовательности буква-три_цифры-две_буквы (обычный номер без кода региона). При этом предполагалось, что первая буква даёт начало первому слову, каждая из трёх цифр кодируется согласной, каждая из которых также даёт начало слову и последние две буквы – ещё два слова. Притом в российском автомобильном номере не могут содержаться любые буквы, используются только 12 из них это: а, в, с, е, н, т, м, о, к, р, у, х. Было взято несколько крупных стихотворений :Евгений Онегин, Полтава, Руслан и Людмила, Ромео и Джульета, басни Крылова. В процессе анализа генерируется 1000 случайных номеров, для которых подбирается цитата из этих произведений

Таблица 1

Результаты кодирования по первой букве

| Название | Среднее | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Объём |
|---|---|---|---|---|---|---|---|---|---|
| Басни Крылова | 2.434 | 0 | 35 | 530 | 403 | 30 | 2 | 0 | 83Кб |
| Евгений Онегин | 3.237 | 0 | 0 | 120 | 549 | 306 | 24 | 1 | 1.1Мб |
| Полтава | 2.507 | 0 | 17 | 510 | 424 | 47 | 2 | 0 | 85Кб |
| Ромео и Джульетта | 2.821 | 0 | 36 | 239 | 598 | 122 | 5 | 0 | 219Кб |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Руслан и Людмила | 2.617 | 0 | 68 | 359 | 469 | 97 | 6 | 1 | 138Кб |

Можно сказать, что эти результаты не обрадовали, получается, что только для двух номеров из 1000 получилось подобрать цитату. Посмотрим на эти две цитаты: м052рк – “мой. Они, пристрастною душой Ревнуя к”; о817вс – “От воспалённого Руслана Сокрылись вдруг среди”. Некоторая логика в этих фразах конечно присутствует, но незаконченность и обрывочность делает их запоминание не слишком простым делом. Тем не менее, тесты позволили сказать, что даже на основе этих текстов в большинстве случаев получается генерировать последовательность для трёх букв.

Конечно же мы заинтересовались: что происходит когда тексты становятся больше? Быть может появляется большее количество фрагментов, из которых уже можно выбрать. Для следующего теста мы выбрали из библиотеки Мошкова: два завета библии в синодальном переводе, все крупные стихи Бодлера, все романы Достоевского, “Хоббит, или Туда и обратно”, все романы Пушкина в стихах, всего Шекспира, все романы Толстого. Были получены следующие результаты, таблица 2:

Таблица 2

| Название | Среднее | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Объём |
|---|---|---|---|---|---|---|---|---|---|
| Ветхий завет | 3.09 | 0 | 0 | 188 | 557 | 235 | 17 | 3 | 1.1Мб |
| Новый завет | 3.126 | 0 | 13 | 180 | 498 | 289 | 17 | 3 | 1.5Мб |
| Достоевский | 4.053 | 0 | 0 | 10 | 206 | 526 | 237 | 21 | 15Мб |
| Толкиен | 3.12 | 0 | 10 | 148 | 581 | 234 | 27 | 0 | 807Кб |
| Пушкин | 3.234 | 0 | 0 | 114 | 565 | 295 | 25 | 1 | 1.2Мб |
| Бодлер | 2.943 | 0 | 4 | 231 | 594 | 160 | 11 | 0 | 461Кб |
| Шекспир | 4.489 | 0 | 0 | 0 | 49 | 474 | 416 | 61 | 64Мб |
| Толстой | 3.96 | 0 | 0 | 8 | 236 | 555 | 190 | 11 | 11Мб |

Самих цитат получилось довольно много, приведем несколько примеров, отметив, что их характер остался прежним, какой-то незаконченно-загадочный: в488ом – “Вот что, Ваня верь одному: Маслобоев”, т380тт – “три тысячи, вскричал он, три тысячи”, м081не – “мне от вас рабство наслаждение. Есть”. Более наглядно зависимость средней длины от объёма текста представлена на рисунке 1:

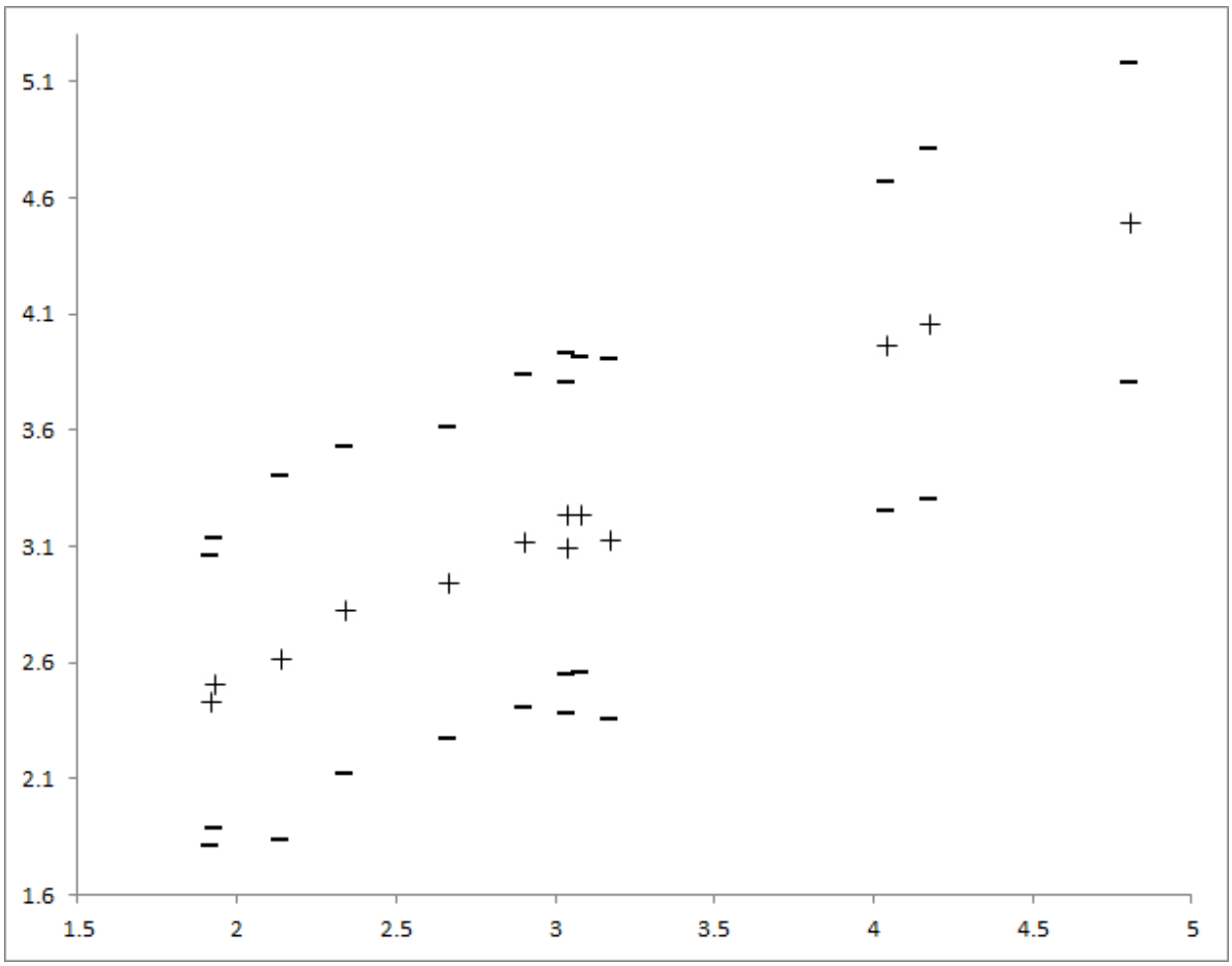


Рис. 1. – График зависимости среднего от десятичного логарифма объёма (по Y – среднее, плюс-минус стандартное отклонение, а по X – логарифм объёма используемого текста)

Как видно из графика, средняя длина последовательности растёт пропорционально логарифму объёма, немного выбиваются только два тома Библии.

**Подход, основанный на кодирование по длинам слов.**

К недостаткам способа кодирования по первым буквам можно отнести то, что слишком много текста остаётся вне рассмотрения, то есть для цифр берутся только 10 букв и для букв 12, остальные слова только разрывают цепочки. Конечно, можно придумать другие способы кодирования, которые используют все буквы или хотя бы только согласные. Эти способы описаны в литературе, но наша идея заключалась в том, чтобы сделать простой в использовании инструмент, такой чтобы пользователь не ломал голову над тем как же соответствует фраза этому номеру, какие нужно выбрасывать буквы, а какие учитывать, в противном случае можно было бы применять кодирование до первой значимой буквы, оно даёт однозначность, но не удобно для человека. Таким образом, решено было реализовать вариант кодирования цифры количеством букв в слове. При таком кодировании возникает проблема представления ноля, но пока не будем на этом останавливаться. Для того, чтобы

сравнить результаты, была проведена серия тестов на том же наборе и по тем же правилам. Результаты работы алгоритма отображены в таблице 3:

Таблица 3

| Название | Среднее | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Объём |
|---|---|---|---|---|---|---|---|---|---|
| Басни Крылова | 2.831 | 0 | 96 | 203 | 490 | 197 | 13 | 1 | 83Кб |
| Онегин | 3.497 | 0 | 96 | 85 | 166 | 536 | 113 | 4 | 1.1Мб |
| Полтава | 2.808 | 0 | 96 | 231 | 459 | 199 | 13 | 2 | 85Кб |
| Ромео и Джульетта | 3.149 | 0 | 96 | 116 | 377 | 371 | 34 | 6 | 219Кб |
| Руслан и Людмила | 2.94 | 0 | 96 | 178 | 443 | 258 | 23 | 2 | 138Кб |

Можно сказать, что ситуация гораздо лучше, только сразу же настораживает одинаковое число 96 в столбике “1”, здесь посчитаны номера, для которых нашлось слово на первую букву, но не нашлось на первую цифру. Это, естественно, номера начинающиеся на ноль. Около 100 таких номеров ещё в столбцах 2 и 3, как можно заметить, их не больше 85. Пример получившейся цитаты: в325нм – “вам: рад бы... право не могу”. Обрывочность в случае со стихами можно компенсировать тем, что приводить цитату от начала строки, пользователю потребуется дополнительно запомнить где в стихотворении начинается номер, например, приведённая цитата должна выдаваться как: “Клянусь вам: рад бы... право не могу.” или даже вместе с предыдущей строкой: “Ах, милостивый рыцарь, Клянусь вам: рад бы... право не могу”. Но так уже перестаёт быть явным начало фразы. Если запоминать начало фразы, то можно и запомнить положение нолей отдельно, тогда получаются следующие результаты, таблица 4:

Таблица 4

| Название | Среднее | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Объём |
|---|---|---|---|---|---|---|---|---|---|
| Басни Крылова | 3.444 | 0 | 0 | 114 | 426 | 367 | 88 | 5 | 83Кб |
| Онегин | 4.26 | 0 | 0 | 2 | 93 | 596 | 261 | 48 | 1.1Мб |
| Полтава | 3.413 | 0 | 0 | 132 | 414 | 367 | 83 | 4 | 85Кб |
| Ромео и Джульетта | 3.791 | 0 | 0 | 39 | 298 | 508 | 143 | 12 | 219Кб |
| Руслан и Людмила | 3.611 | 0 | 0 | 83 | 356 | 445 | 99 | 17 | 138Кб |
| Ветхий завет | 4.189 | 0 | 0 | 5 | 126 | 585 | 243 | 41 | 1.1Мб |
| Новый завет | 4.292 | 0 | 0 | 3 | 83 | 581 | 285 | 48 | 1.5Мб |
| Достоевский | 5.019 | 0 | 0 | 0 | 0 | 208 | 565 | 227 | 15Мб |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Толкиен (Хоббит) | 4.123 | 0 | 0 | 2 | 155 | 587 | 230 | 26 | 807Кб |
| Пушкин | 4.251 | 0 | 0 | 3 | 94 | 593 | 269 | 41 | 1.2Мб |
| Бодлер | 3.946 | 0 | 0 | 18 | 214 | 591 | 158 | 19 | 461Кб |

По этим данным можно предположить, что для случайного номера с вероятностью 22% можно подобрать соответствующую цитату из Достоевского. Цитаты конечно получаются весьма многозначительными, как и в прошлом случае: в725вр – “весело смотрит за нашей весёлой работой”, м582то – “мои слова казалось её тронули, она”, м385нс – “между тем каким-то чудом необыкновенное сходство”, к514нт – “кровавой битве и день настал толпы”.

Теперь предположим, что на вход поступает только последовательность из цифр, тогда с результатами работы программы можно ознакомиться в таблице 5.

Таблица 5

| Название | Среднее | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Басни Крылова | 4.145 | 0 | 23 | 270 | 387 | 214 | 71 | 35 |
| Онегин | 5.248 | 0 | 0 | 9 | 260 | 347 | 242 | 142 |
| Полтава | 4.131 | 0 | 16 | 293 | 385 | 193 | 76 | 37 |
| Ромео и Джульетта | 4.608 | 0 | 0 | 138 | 374 | 286 | 146 | 56 |
| Руслан и Людмила | 4.349 | 0 | 11 | 212 | 381 | 256 | 93 | 47 |

Эти результаты позволяют предположить, что для шестизначного случайного номера с вероятностью почти 40% найдётся соответствующая цитата из произведения “Евгений Онегин”, а это ведь стихи, которые гораздо приятней к запоминанию (не для всех скорее всего, но для большинства всё же).

**Заключение**

Какие ещё возможности остались за кадром: генерация текстов, а именно генерация соответствующих слов или предложений определённой структуры (с нужным количеством букв или ещё как) в принципе это уже давно сделано и без компьютеров. Указанный в литературе учебник [1] предлагает для каждого числа от 0 до 1000 какое-то слово, которое уже подобрано автором, но, к сожалению, такой способ не даёт возможности запоминать большие числа, поскольку образы нельзя соединять, это по словам автора приводит к их стиранию. Оно и понятно, всё начинает наслаиваться и так далее. Вот например простой способ: можно закодировать цифры распространёнными ассоциациями – 3 (от 03) – врач, 5 – отличаться и например 5 – пятница. В этом случае подобрав для каждой цифры по три образа (для каждого положения) все трёхзначные числа можно закодировать очень яркими историями вроде “355 – врач отличился в пятницу”, но так можно запомнить только очень

небольшое количество чисел, потому что всё начнёт мешаться, нужно проводить дополнительные параллели, чтобы запомнить когда именно зарубили врача в этот раз.

Описанные механизмы генерации реализованы кроме того, подбор цитат на шестизначные числа реализован на сайте приложения: YaZapomnil [6], там можно посмотреть какая цитата из классики соответствует дате рождения, телефону, пин-коду карточки или любому другому числу, которое следует запомнить.

***Работа выполнена при частичной поддержке РФФИ, проект 11-07-00561***

## Список литературы

1. Козаренко В. А. Учебник мнемотехники – Москва 2002 – 85с.

2. Марков А.А. Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь. // Известия Имп.Акад.наук, серия VI, Т.X, N3, – 1913 – с.153.

3. Preczewski, S. C., & Fisher, D.L. The selection of alphanumeric code sequences. // Proceedings of the Human Factors Society 34th Annual Meeting – Santa Monica, CA: HFS. – 1990 – P. 224-228.

4. Vu, K.-P. L., Cook, J., Bhargav, A., & Proctor, R. W. (2006, April). Short-term and long-term retention of passwords generated by first-letter and entire-word mnemonic methods. In Proceedings of the fifth annual security conference.

5. Vu, K.-P. L., Tai, B.-L. (B.), Bhargav, A., Schultz, E. E., & Proctor, R. W. Promoting memorability and security of passwords through sentence generation. // Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting. – Santa Monica, CA: HFES. – 2004 – P. 1478-1482

6. YaZapomnil Подбор цитат [Электронный ресурс]. URL: http://yazapomnil.ru/n/ (дата обращения 20.11.2013)

**Рецензенты:**

Барахнин В.Б., д.т.н., доцент, старший научных сотрудник Института вычислительных технологий СО РАН, г.Новосибирск.

Никульчев Е.В., д.т.н., профессор, проректор НОУ ВПО Московский технологический институт "ВТУ", г. Москва.