TRANSMUTATION-BASED APPROACH FOR WEB DATA EXTRACTION

Michael Dimitrov

Summary: *The research article presents an in-depth examination of a C# codebase specifically designed for the intricate handling and transformation of textual data. Central to this exploration is the innovative transmutation methodology, which is operationalized through the construction of a "legend". This patterned sequence acts as a blueprint, offering a systematic depiction of the content that has been distilled from the data. The coding framework integrates the use of regular expressions to pinpoint patterns, string operations to alter text, and sophisticated data structures to systematically arrange and purify the information that has been gleaned. In an effort to quantify the efficacy of this transmutation-centric technique, the study outlines the findings from a confidential survey administered to experts within the field. These insights are instrumental in formulating weighted indices that measure key performance indicators such as execution time and accuracy. The ultimate objective is to evaluate the overall effectiveness of the transmutation strategy and to benchmark it against an alternative solution that relies exclusively on REGEX for implementation. This comparative analysis is expected to yield valuable insights into the optimization of code for text processing tasks, especially in the field of security-related topics.*

Key words: *significant part extraction, transmutation, Big Data processing.*

INTRODUCTION

The recent years have witnessed a surge in the capabilities of AI, particularly in the realm of machine learning, which stands at the forefront of this technological renaissance. This paves the way for advanced Big Data processing and ushers in a new era of automated meaning extraction, seamlessly integrated into decision-making systems. Numerous solutions circumvent the challenge of identifying the significant portions within an extracted HTML document. Extracting the substantive content from an HTML document is a formidable task, fraught with the challenge of sifting through tags and extraneous text like advertisements. Yet, paradoxically, when dealing with data on a massive scale, one can often bypass this obstacle and still extract meaningful insights, though the precision of such results may be somewhat compromised. There are many articles that discuss data harvesting techniques but generally, they do not put emphasis on this problem (Khder, 2021; Patnaik, & Babu, 2022).

This paper unveils a versatile approach for pinpointing the crux of content across diverse web domains, eschewing rigid anchor-based methods in favor of a mission-aligned strategy. Our quest is to dissect and evaluate the prowess of this method, juxtaposed against the backdrop of an entirely

REGEX-based paradigm. To achieve this, we embark on a series of investigative endeavors:

1. **Charting the qualitative blueprint** of the transmutation-centric methodology.
2. **Balancing the scales** of speed versus precision.
3. **Computing the effectiveness** of both the transmutation-centric and REGEX-based methodologies.

In pursuit of scientific rigor, our methodology is twofold:

1. Data collection through an **anonymous survey of domain experts**.
2. **Quantifying overall effectiveness** of the approach through a formula adapted for our research goals:

$$Effectiveness = w_1 . \left(\frac{1}{Execution\ time}\right) + w_2 \, . Accuracy \qquad (1)$$

The **research hypothesis** posits that a transmutation-based approach yields superior effectiveness compared to a purely REGEX-centric methodology, particularly in the context of web data extraction from documents containing a substantial Cyrillic component.

The **research constraints** stem from the implementation specifics, which rely on C# and the Puppeteer library for web automation, coupled with the Uglify JavaScript parsing and minification library for the removal of HTML tags. An additional limitation springs from the hands-on nature of our accuracy assessment, confining our analysis to a curated collection of 100 HTML documents.

The **findings of the study** may prove beneficial for enhancing existing software solutions that are geared towards processing information on topics strictly related to security or falling within adjacent scientific fields. In this case, the monitoring of information exchange will be linked to the geopolitical aspect of the concept of "security," as defined by Dimitar Yonchev (2022): "It is more accurately used in a negative sense, as a state of insecurity generated by the activity of a geopolitical player or players, which creates tensions beyond the control of other participants" (p. 10).

A direct positive effect can also be achieved in terms of tracking processes and phenomena affecting the balance at a strategic level. This is particularly true when considering the conclusions of Yordan Bakalov (2023): "An analysis of the Russian nuclear doctrines adopted over the years can lead to the conclusion that Russia has adopted the strategy of 'escalation for de-escalation' and itself 'mistakenly assesses that the threat of nuclear escalation or the actual first use of nuclear weapons will serve to 'de-escalate' a conflict on terms favorable to Russia'" (p. 205).

Undoubtedly, the automated and accurate extraction of information about political processes holds potential for an adequate response to strategic

challenges in a regional format, as they are defined for Southeast Europe by Marian Ninov (2021). Combining the approach presented here with artificial intelligence methods may prove beneficial and facilitate the conduct of information-analytical work. At the same time, it should be noted that the approach proposed here, along with the software solutions of which it is a part, may find it challenging to be applicable in areas such as anti-money laundering and countering the execution of hawala transactions, particularly considering the characteristics of the same, as defined by Gergana Yordanova (2023). The applicability of software for processing Big Data in the context presented here is closely linked to the presence of an information flow passing through publicly accessible sources.

1. APPLYING TRANSMUTATION-BASED APPROACH FOR SIGNIFICANT PART EXTRACTION

Text processing is an essential component of numerous software applications and is central to the endeavor of figuring the main portion of HTML document. The code under discussion exemplifies a robust method for extracting relevant data through specific patterns from text, altering the text based on these patterns, and preparing the text for further analysis or display. The initial stage of the method involves defining a regular expression to match site names within a larger body of text. This pattern is designed to identify sequences of characters that represent site names, followed by a space and a domain in order to avoid sentence segmentation errors.

```
Regex sites = new Regex("[pattern]");
```

The matched patterns are then processed to extract unique site names using LINQ operations, ensuring that each site name is represented only once.

```
var sitesInText = sites.Matches(input)
    .OfType<Match>()
    .Select(m => m.Groups[0].Value)
    .Distinct()
    .ToList();
```

Central to the code's functionality is the creation of a "legend", a symbolic representation of the text that categorizes each character according to its role. This legend serves as a guide to the structure and content of the text, facilitating efficient processing and manipulation. It is specifically aimed at highlighting the Cyrillic content and its interaction with the other parts of the document.

```
char[] legend = new char[input.Length];
for (int i = 0; i < input.Length; i++)
{
    // Code to populate the legend array
}
```

Furthermore, the position of the respective token relative to the entire HTML document is utilized to determine specific predominant requirements for the presence of characters that do not correspond to the main part of the text. Both the actual string of the text and the created "legend" are maintained and manipulated simultaneously. An example of the appearance of the "legend" string is provided in Figure 1.

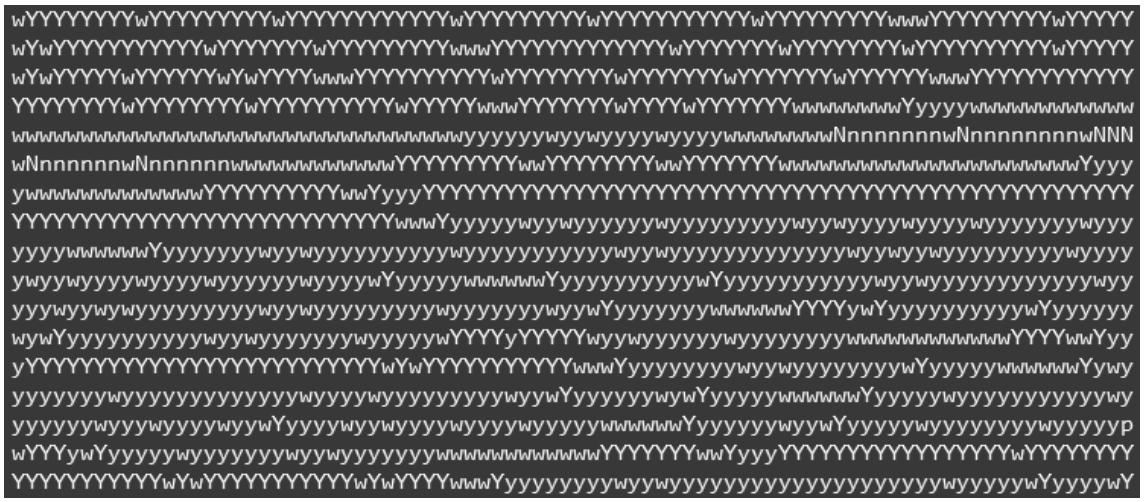

Figure 1. Resultant patterned sequence (Image generated by the author in Microsoft Visual Studio 2022)

The legend abstracts the text into a form that is both informative and malleable, using symbols to represent different character types. This abstraction simplifies the subsequent analysis and manipulation tasks. Following the creation of the legend, the code replaces identified site names with placeholders, crucial for restoring the original text after other manipulations which remove symbol patterns that otherwise might distort correct determination of the actual content of the document.

```
for (int i = 0; i < sitesInText.Count; i++)
{
    // Code to replace site names with placeholders and store them
}
```

The code meticulously handles abbreviations and specific phrases, replacing them with unique markers to maintain the integrity of the text's meaning and to avoid unnecessary segmentation.

```
if (input.ToLower().Contains("abbreviation"))
{
    // Code to replace abbreviations with unique markers
}
```

The legend is then utilized to segment the text into logical parts and identify patterns for alteration or removal, leading to a cleaned and processed output. The method applies a set of heuristics to recognize and filter out common advertisement patterns and other non-essential sections. This includes analyzing the size, position, and content of elements to determine their relevance to the main article. Certain colloquial expressions that may interfere with the precise delineation of the core concept are temporarily substituted with direct encoding within the text, both in the actual string and the patterned sequence, and are reinstated at a subsequent phase. The algorithm traverses in reverse order through the isolated tokens and outputs potential starting points of the main section, which are subsequently subjected to verification according to specific criteria that align with the program's objectives. Once the main content is identified, the method applies a purification process to remove any remaining non-essential text or elements. This process ensures that the final output is a clean and concise representation of the original document's primary content.

The method has been tested on a diverse set of HTML documents with varying layouts and content types. It should be emphasized that the method is specifically adapted to a predefined set of sites (but without an implementation of site-specific strategies). The results demonstrate a high degree of accuracy in extracting the main content while effectively discarding advertisements and other data (links to other content). The method proves to be quite robust and as such, it constitutes a foundational phase in the preprocessing of data for machine learning algorithms, which is designed to preclude the incorporation of extraneous information, thereby enhancing the precision of automated web data surveillance systems.

2. COMPARISON WITH REGEX APPROACH

The preliminary phase of the comparative analysis involves establishing the relative importance of the two pivotal factors: execution time (w_1) and accuracy (w_2). To achieve this, a survey is administered to an expert panel comprising professionals in the fields of security and information technology. Participants have been instructed to assign a value to each factor on a scale from 1, denoting 'not important at all,' to 10, indicating 'extremely important'. The survey (conducted from 18th to 20th March 2024) design incorporates a stratification mechanism where respondents are categorized into three experience-based cohorts, each assigned a corresponding weight: 'Weight 1' for less than 5 years of experience, 'Weight 2' for 5-10 years, and 'Weight 3' for over 10 years.

To compute the weighted averages, each factor's average importance rating, as determined by the respondents, is multiplied first by the number of respondents in their respective group and then by the group's weight. These products are then aggregated, and the sum is divided by the total of all weights, yielding the weighted average for each factor. Mathematically, the weighted average is calculated using the formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i . x_i}{\sum_{i=1}^{n} w_i} \quad (2)$$

The resultant weighted averages are subsequently utilized to ascertain normalized weights within the overarching effectiveness formula. This methodological approach ensures that the insights of more seasoned practitioners exert a commensurately greater impact on the final weight determinations. The implementation of these steps results in weighted average for execution time 5.095238095 and weighted average for accuracy 9.404761905. Thus, the normalized values for $\mathbf{w_1}$ and $\mathbf{w_2}$ are respectively **0.351395731** and **0.648604269**.

Before we proceed with the comparison of the effectiveness of the two approaches under consideration, let us examine in more detail the results of the conducted survey. The distribution of the respondents by weight groups is presented in Figure 2.

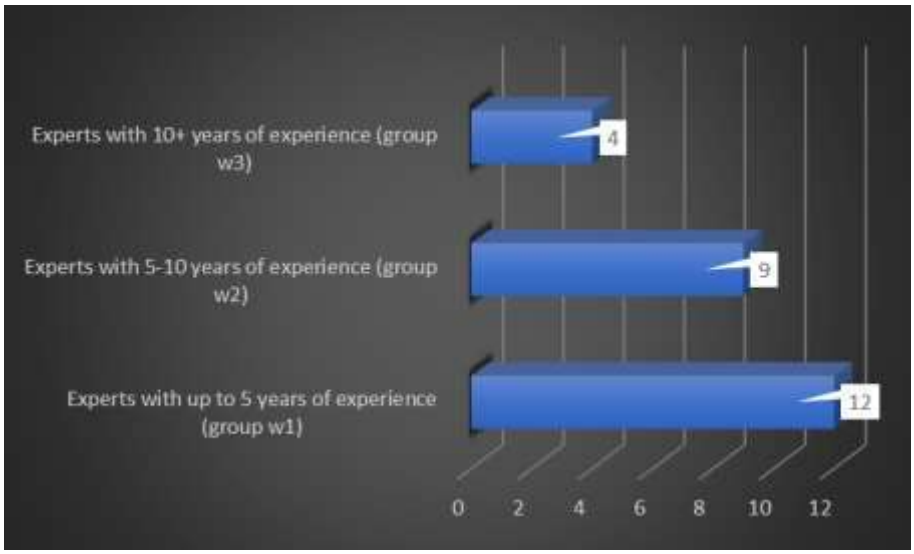

Figure 2. Distribution of respondents by weight groups (Image generated by the author in Excel)

Without compromising the anonymity of the survey, it should be noted that respondents with experience strictly in the security sector place a distinct emphasis on the importance of accuracy. Conversely, individuals associated

with cybersecurity and information technology, while acknowledging productivity as a factor, provide more balanced responses, although for them, accuracy also remains paramount over the speed of execution. In order to achieve an in-depth analysis of the results obtained, a table with the average values for the two factors of interest, for each of the three weight categories, is provided below.

Table 1. Table of average values for execution time and accuracy across weight categories (Table generated by the author in Word)

Weight category	Average value for perceived importance of execution time	Average value for perceived importance of accuracy
W3	4	9.833333
W2	4.222222	8.888889
W1	7.5	9.75

Respondents with over 10 years of experience (W3) attribute high importance to accuracy (approximately 9.83), with less emphasis on execution time. This suggests seasoned professionals prioritize precise and correct outcomes. The intermediate experience group (W2) places a slightly higher importance on execution time (approximately 4.22) and less on accuracy (approximately 8.89), indicating a growing appreciation for efficiency and a more pragmatic approach in relation to the programmatic tools utilized. The least experienced users (W1) rate execution time quite highly (7.5) while also valuing accuracy (approximately 9.75), which may reflect an initial focus on quick results, but with a strong understanding of the importance of accurate outputs.

These conclusions indicate that as users become more familiar with the role of software designed for processing Big Data from the web, they may develop a more nuanced understanding of the trade-off between speed and accuracy, ultimately valuing accurate results even if it requires more time. This observed pattern may play a crucial role in shaping the design of software systems that manage extensive web-based datasets, ensuring they align with the diverse requirements and preferences of users at varying expertise levels.

We shall now proceed to examine the actual values achieved by the two approaches – the transmutation-based and the one relying entirely on regular expressions – in terms of the two indicators by which they will be compared: execution time and accuracy. For this purpose, they have been applied to a set of 100 HTML documents. The transmutation approach achieves an average speed of **3268.12 milliseconds**, while the REGEX approach achieves an average speed of **1527.08 milliseconds** (these values will be

used in seconds in the formula for assessing overall effectiveness). A significant advantage of regular expressions is observed in terms of this indicator.

Regular expressions are known for their efficiency and speed, particularly when it comes to parsing and matching patterns within strings of text. This efficiency is due in part to the ability of REGEX to perform complex searches and manipulations with minimal code. According to Microsoft's documentation, REGEX is designed to match patterns "in input text" using "one or more character literals, operators, or constructs," which allows for a concise and powerful way to process text (.NET, 2022).

For example, the use of character classes and quantifiers enables REGEX to match multiple instances of a pattern efficiently. The REGEX engine is optimized to perform these operations quickly, which is why it can often outperform manual string operations, especially when dealing with large or complex datasets. In summary, the speed of REGEX can be attributed to its compact syntax, which allows for the execution of sophisticated string operations with relatively little overhead, making it a preferred method for text processing tasks.

When we turn our attention to accuracy, the transmutation approach achieves significantly better results than its competitor – 88.36% (**0.88361896**) versus 67.16% (**0.67158255**). Taking into account the complexity of the processed HTML documents, the REGEX-based approach does not yield poor results, but it is incomparable to the approach presented here. Considering the values derived, we can proceed with presenting the results for the overall effectiveness of the transmutation-based approach (formula No. 3) and the REGEX-centric solution (formula No. 4).

$$\mathbf{0.680641185} = 0.351396 \,.\left(\frac{1}{3.26812}\right) + 0.648604 \,.\, 0.883619 \qquad (3)$$

$$\mathbf{0.665701172} = 0.351396 \,.\left(\frac{1}{1.52708}\right) + 0.648604 \,.\, 0.671583 \qquad (4)$$

The effectiveness scores suggest that the transmutation-based approach has a slightly higher overall effectiveness score (0.680641185) compared to its entirely REGEX counterpart (0.665701172). This indicates that, according to the weighted factors of execution time and accuracy, the transmutation-based approach is marginally more effective for the users' needs.

The transmutation approach, despite being slower in execution time, compensates with its higher accuracy score, resulting in a greater overall effectiveness score. On the other hand, the REGEX-centric solution, while

faster, has a lower accuracy score, which slightly reduces its overall effectiveness.

All of this highlights the trade-off between execution time and accuracy in determining the effectiveness of programmatic solutions for Big Data processing from the perspective of the users. It suggests that for tasks where accuracy is highly valued and weighted more heavily than speed, a method that may take longer but yields more accurate results could be considered more effective overall. This is particularly true in the interaction between deep learning models and expert systems with schematic knowledge, for which Svetoslav Velkov (2023) notes that: 'Human experts supplement the base of expert knowledge, refine the knowledge schema, and enrich the ontology; they do not concern themselves with the quality of the statistical data entering the deep self-learning system, nor do they monitor the quality of this system's output. On the contrary, it is the expert system that corrects the deep learning process at all stages.'(p. 207).

CONCLUSION

The empirical evidence from the study **supports the research hypothesis** that a transmutation-based approach to web data extraction is more effective than a purely REGEX-centric method, especially when dealing with documents containing significant Cyrillic content. The findings underscore the importance of accuracy over speed in certain contexts of Big Data processing. While the transmutation method may require more time to execute, its superior accuracy score suggests that it is a more suitable approach for scenarios where precision is paramount. This research contributes to the ongoing discourse on optimizing data extraction techniques and highlights the necessity of balancing execution time and accuracy to meet user requirements effectively. **Future studies** could further explore this balance, potentially leading to the development of hybrid methodologies that harness the strengths of both transmutation and REGEX to offer optimized solutions for diverse data processing needs.

Of particular interest is the synergistic integration of the transmutation approach with the application of regular expressions targeting specific "anchors" pertinent to websites of interest, thereby enhancing overall efficiency. Another avenue of work is tethered to the utilization of computer vision to forge a unified methodology for discerning the substantive segment of HTML documents. In the latter case, the comprehensive architecture of the software will not be predicated on Puppeteer, but rather on a specially devised algorithm for automating the process of content access.

These future studies would seek to amalgamate the transformative potential of the transmutation technique with other existing tools. In this context, it would be beneficial to compare the computational resource requirements for each implemented and planned-for-development approach.

This comparative analysis could provide valuable insights into the efficiency and scalability of different methodologies in data processing applications.

BIBLIOGRAPHY:

Бакалов, Й. (2023). Руски ядрени доктрини, ядрени сили, модернизация. *Сигурност и отбрана*, (1), 191-211. https://institute.nvu.bg/sites/default/files/inline-files/2023-1-14-bakalov.pdf // Bakalov, Y. (2023). Ruski yadreni doktrini, yadreni sili, modernizatsiya. *Sigurnost i otbrana*, (1), 191-211. https://institute.nvu.bg/sites/default/files/inline-files/2023-1-14-bakalov.pdf

Велков, С. (2023). Методи за контрол на изкуствения интелект. *Сигурност и отбрана,* (2), 199-211. https://institute.nvu.bg/sites/default/files/inline-files/2023-2-15-velkov.pdf // Velkov, S. (2023). Metodi za kontrol na izkustveniya intelekt. *Sigurnost i otbrana*, (2), 199-211. https://institute.nvu.bg/sites/default/files/inline-files/2023-2-15-velkov.pdf

Йончев, Д. (2022). Сигурна ли е сигурността? *Сигурност и отбрана*, (1), 7-14. https://institute.nvu.bg/sites/default/files/inline-files/2022-1-01-yonchev.pdf // Yonchev, D. (2022). Sigurna li e sigurnostta? *Sigurnost i otbrana*, (1), 7-14. https://institute.nvu.bg/sites/default/files/inline-files/2022-1-01-yonchev.pdf

Йорданова, Г. (2023). Разграничение между „черни“ и „бели“ Хауала транзакции в контекста на противодействието на изпирането на пари. *Сигурност и отбрана,* (1), 129-143. https://institute.nvu.bg/sites/default/files/inline-files/2023-1-10-yordanova.pdf // Yordanova, G. (June 2023 r.). Razgranichenie mezhdu „cherni“ i „beli“ Hauala tranzaktsii v konteksta na protivodeystvieto na izpiraneto na pari. Sigurnost i otbrana, (1), 129-143. https://institute.nvu.bg/sites/default/files/inline-files/2023-1-10-yordanova.pdf

Нинов, М. (2021). Стратегическото предизвикателство за сигурността в Югоизточна Европа. В: Дойков, Н. (съст. и науч. ред). *Сборник с научни трудове от Международна научна конференция „Кризи и сигурност – корелации и предизвикателства“, 14 май 2021 г., Т. 1. Кризи и сигурност на международно и регионално ниво* (с. 69-78). Научно-технически съюз по машиностроене „Индустрия 4.0“. ISBN 978-619-7383-22-5. // Ninov, M. (2021). Strategicheskoto predizvikatelstvo za sigurnostta v Yugoiztochna Evropa. V: Doykov, N. (sast. i nauch. red). *Sbornik s nauchni trudove ot Mezhdunarodna nauchna konferentsiya „Krizi i sigurnost – korelatsii i*

predizvikatelstva“, 14 may 2021 g., T. 1. Krizi i sigurnost na mezhdunarodno i regionalno nivo (s. 69-78). Nauchno-tehnicheski sayuz po mashinostroene „Industriya 4.0“. ISBN 978-619-7383-22-5.

.NET. (2022, June 18). Regular Expression Language - Quick Reference. https://learn.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and Its Applications, 13*(3), 144–168. https://doi.org/10.15849/ijasca.211128.11

Patnaik, S. K., & Babu, C. N. (2022). A web information extraction framework with adaptive and failure prediction feature. *Journal of Data and Information Quality, 14*(2), Article 12. https://doi.org/10.1145/3495008