Identification of Writing Preferences in Wikipedia

Jean-Baptiste Chaudron, Jean-Philippe Magué, Denis Vigier

Identification of Writing preferences in Wikipedia

Jean-Baptiste Chaudron[1], Jean-Philippe Magué[2], and Denis Vigier[1]

[1] Université Lumière Lyon2, Laboratoire ICAR, 69676 BRON Cedex,
jean-baptiste.chaudron@ens-lyon.fr
[2] École Normale Supérieure Lyon, Laboratoire ICAR, 69342 LYON cedex 07

Abstract. In this paper, we investigate whether there is a standardized writing composition for articles in Wikipedia and, if so, what it entails. By employing a Neural Gas approximation to the topology of our dataset, we generate a graph that represents various prevalent textual compositions adopted by the texts in our dataset. Subsequently, we examine significantly attractive regions within our graph by tracking the evolution of articles over time. Our observations reveal the coexistence of different stable compositions and the emergence and disappearance of certain unstable compositions over time.

Keywords: Textual Genre, Dynamic Graph, Temporal Graph, Computational Linguistics, Wikipedia

1 Introduction

1.1 Writing preferences in Wikipedia

In Wikipedia, writers iteratively modify texts, starting from a stub and progressing towards a fully formed article. The motivations driving these modifications can be manifold, resulting in a broad spectrum of textual variations. This can range from simple typo corrections to the complete deletion of a paragraph or even the entire article. It includes the addition of new sections, relocating sections to other articles, and more. Despite being composed of diverse materials and written by various authors over several years, covering a wide array of subjects, articles maintain a well-structured format. They are now considered reliable sources of information.

Even though some guidelines for correct writing exist, authors are free to deviate from them, and in some ways, this deviation might even be encouraged, as stated in Wikipedia's fifth pillar: 'Wikipedia has no firm rules' [16]. Research has also demonstrated that there is a degree of homogeneity in Wikipedia's writing, both in English when compared to other encyclopedias [3], and in languages such as Japanese [13]. We've also observed this trend in our own French corpus (which might result in a future publication).

Consequently, it's worth exploring the mechanisms that govern the creation and composition of articles, especially given the fact that the process is purely

auto-organisational. One hypothesis, explaining the surprising homogeneity of the articles, could be that, even though they are unspoken, there is rules or forces, acting upon the writing process, and driving it. Alternatively, one could consider that the introduction of purely random variations in terms of writing composition would lead to a series of articles with similar structures.

We will postulate that these forces take the form of an ideal version, in the mind of the writers, of the shape an article should have. These forms could either be shared among all members of the writing community or could represent a compromise between the ideals of different communities of writers. To delve deeper into this question, we will introduce some concepts from textual linguistics, specifically from textual genre theory. This theory focuses on the study of textual forms and how they are categorized into types of texts that share similar forms.

1.2 Genre & Prototype theory

Multiple definitions of textual genres have been proposed, each stemming from different perspectives applied to this subject. While it is widely accepted that genres are largely dependent on the social context and the communicative purpose of the production [1], the textual composition and the presence or absence of specific features are also major areas of research in the study of textual genres, especially in computational approaches [10]. These features are predominantly considered to be formal in nature, whether about the layout of the texts [2, 12] or about more textual features on which we will discuss longer.

For the later, features are often more linguistically oriented [8, 14], particularly those related to what we will refer to as textual composition. These features describe a text in terms of part of speech (POS), such as the frequency of nouns, verbs, etc., syntactical dependencies (e.g., frequency of subordinates, conjunctions, etc.), verb morphology (e.g., frequency of past tense, future tense, passive verbs, etc.), lexical diversity, mean word length, and so on.

The literature concerning these features is somewhat mixed because there is a consensus that there isn't a strong theoretical basis justifying their direct relation to genres, but rather to register and style [1]. Nevertheless, despite these reservations, empirical literature has found great success in utilizing them [6–8, 11, 14]. While some features appear to be more relevant than others for distinguishing texts based on their genre, the question of what constitutes a genre remains open. One approach to defining genre is through the lens of *Prototype* theory [17]. This theory posits that genres are classes of texts that bear resemblance to an idealized version of themselves, known as their *Prototype*.

For instance, envisioning the *Prototype* of a genre like *Poetry* would involve specific features such as the presence of verses and rhyming patterns. However, unlike a precise and fixed definition of what constitutes *Poetry*, the *Prototype* theory offers an idealized version from which actual poems can deviate. Consequently, works like Rimbaud's *A Season in Hell*, which lacks traditional verse and rhymes, can still exhibit poetic qualities that place it within the category of *Poetry*.

Because the *Prototype* theory doesn't necessitate a rigid set of features or strict boundaries, it possesses the strength to elucidate the process of fuzzy categorization.

1.3 Prototype & Writing preferences

The *Prototype* of a genre possesses two crucial features. Firstly, it acts as a sort of center of mass for its genre, representing the average or typical instance of that genre. For instance, in the context of *Poetry*, the *Prototype* would resemble an average poem. Secondly, it's believed that the *Prototype* exerts a centripetal force, influencing texts to align with its characteristics.

If we aim to operationalize this concept to model the modification of texts, influencing the generation process of articles in Wikipedia, whether random or directed, we can consider the presence or absence of a *Prototype*. In the case of directed modifications, taking into account the center of mass property, we would anticipate a higher density of texts clustered around a specific point in the space of the *Prototype*. This suggests that these texts share a similar textual composition with the *Prototype*. In terms of the centripetal features of the *Prototype*, the outcome would differ, as each modification would gradually steer texts towards the attractive point, which is the *Prototype*. Linguistically, this can be interpreted as texts converging gradually, in terms of linguistic composition, towards the ideal version.

Conversely, in a context of pure randomness, we would anticipate either the absence of a higher density point or the absence of regions with greater attraction in our phase space. It's worth noting that in the scenario of a single *Prototype* acting as the attracting force, we expect the scaling of the features performed during modifications could tend to diminish the differences between texts, as they already share the same *Prototype*.

Our research question aimed to explore how writing forms within Wikipedia can exhibit homogeneity and harmony across the project despite collaboration among diverse entities with differing goals. We suggested that if there's a shared structure, there should be a driving force behind modifications; otherwise, textual variations might be better modelled as a random process. By introducing the concept of the *Prototype* theory from genre theory, we put forth a strong contender for describing the type of force at play in this context.

To further clarify, let's restate our hypotheses. Firstly, we propose that if genres are organized around *Prototypes*, we should observe either multiple points of higher density in the phase space or regions with significant attractive influence. Secondly, if Wikipedia exhibits homogeneity, there should be just one such *Prototypical* form. Last, if no specific forces drive the process, we should not be able to observe neither specifically high density point, nor highly attractive textual forms.

2 Method

2.1 Dataset

Our corpus consists of 4800 articles sourced from the French Wikipedia, obtained in March 2022 using the PyWikibot Python library. During extraction, these articles were considered as either "good" or "featured", indicating that the community deemed them well-structured and largely comprehensive in terms of content.

For each of these articles, we extracted all available versions, encompassing their entire revision history from the initial version added to Wikipedia up to the present. Each version of an article is linked to its corresponding revision timestamp and the author responsible for the changes. This compilation resulted in a dataset of over 2 million texts, representing every iteration of each article over time.

2.2 Preprocessing of the dataset

To conduct our desired analysis, we needed to transform our texts into vectors. While common methods like text embedding (e.g., using techniques like *Doc2Vec*) could have been employed, our objective was to create vectors that would enable us to differentiate texts based on their genre-related features. As a result, we extracted a set of features that are typically associated with genre characteristics, as described in Table 1.

We draw attention to the fact that we've added Wikipedia-specific features. These features are Wikimarkup elements specific to the platform. They allow users to add metatextual information or structure the text. Examples include links between articles or to other pages, images, citations, dates, and so on. They are relevant in the context of genre as they provide information about the formal structure of the text, thus helping to better isolate features specific to the text's form

Table 1. Linguistic Features for Textual Analysis

Feature	**Description**
Part of Speech (POS)	Frequency of different parts of speech, such as nouns, verbs, adjectives, adverbs, etc.
Syntactical Dependencies (DEP)	Frequency of syntactical constructs like subordinates, conjunctions, and other dependencies.
Verb Morphology	Frequency of different verb forms, such as past tense, present tense, passive voice, etc.
Lexical Diversity	Measurement of vocabulary richness, including the number of unique words used.
Other morphological properties	Average length of words in the text, number of sentences, number of tokens etc.
Wikipedia's features	Frequency of Tag, Text, Templates, Images etc.

Except for Wikipedia's specific features, the features were extracted using SpaCy's proposed model of tagging [4]. In this case, we used the model named 'fr_core_news_lg,' *i.e.*, a generalized purpose model for French, trained on a news dataset. Subsequently, after extracting these features from the texts, we calculated their averages, taking into account the number of tokens present in each text. This process allowed us to create vectors representing each text's genre-related features, which we then used for further analysis.

We were left with slightly over 2 million vectors, each representing a text, with over 120 dimensions. These vectors underwent normalization, resulting in features with a mean of 0 and a standard deviation of 1. Subsequently, we employed Truncated Singular Value Decomposition (tSVD) to reduce the dimensions from 120 to 75. This selection was based on the explained variance ratio score, where we aimed for the minimal number of dimensions that accounted for more than 99% of the dataset's variation. The motivation for such transformation is purely to reduce the correlation between features, which might bias the intended analysis. Another concern might be the potential loss of interpretability of the new features; we address this in Section *2.5* regarding the analysis of the prototypes.

The outcome was vectors representing the genre-related features of all versions of our texts. These vectors exhibited no correlation between features (the maximal correlation between two dimensions being $1e^{-14}$ and had undergone scaling. This set the stage for us to conduct distance analyses, exploring relationships and patterns within the data.

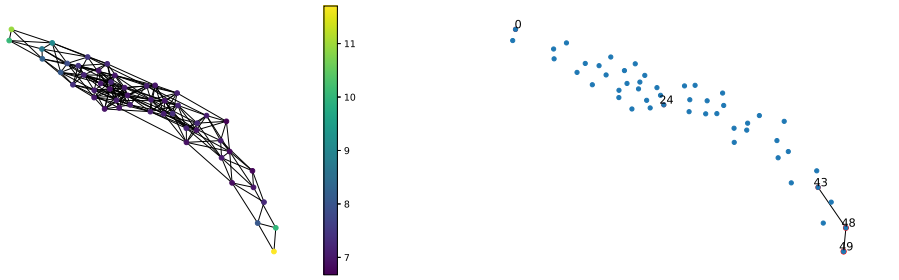
2.3 Prototype Identification

Topology extraction through Neural Gas : A major challenge with our approach is that identifying attractive regions solely from the data isn't straightforward, particularly in the presence of noise. Normally, such a process is executed with knowledge of the vector field associated with the system, rather than in a data-driven manner.

To address this issue, one approach is to discretize the phase space to work with a simpler representation of the data. In our case, we've discretized our dataset into a graph where each region is represented by a node connected to its neighboring regions. To perform such discretization we've used a Neural Gas (NG) [9] enabling this type of reduction. In the fast growing research in machine learning and topological data analysis, a large number of newer and alternative approach exists, however, we've considered that this one is robust, efficient and fast, making it a pragmatic choice for our purpose.

Through the NG discretization, the *Prototypes* becomes regions of the phase space, represented by a specific textual composition. It achieves this by symbolizing regions through their centers, as the nodes in the NG-generated graph are linked to the central vector of the region they model, the result of this process is shown in Fig.1(a). In this article, we've used a 100 node graph to represent the distribution of our data, resulting from the NG learning of the topology. The nodes were related to regions with high density of texts and links to the

adjacency of the nodes *i.e.* the presence of a more or less high density of data points in-between the two nodes.



(a) *Representation of the graph of the topology of the dataset.*

(b) *Subgraph including only significantly attractive regions.*

Fig. 1. Representation of our dataset as a graph. (a) 2D projection of our graph, where nodes represent high density regions and edges represent the existence of more or less high density of point in-between the nodes. Colors represent the log of the number of articles associated with a node. (b) Visualisation of the attractor on the graph (in red), over the whole graph without its edges (in pale blue), 5 disconnected components can be seen.

Attraction Metric : Furthermore, we can now interpret the process of attraction as a region's capacity, over time, to accumulate new articles while preventing them from leaving that region. This phenomenon can be understood within a region as an income rate of texts exceeding the rate at which texts depart.

We can characterize a region's behavior as follows:

- An attractive region (or sink) is one where the income rate is greater than the outcome rate.
- A transitional region exhibits an equal rate of income and outcome.
- A source region experiences a higher outcome rate than income rate.

This metric, when applied to the discrete topology of a graph, is akin to the concept of "divergence" in dynamical systems and vector analysis. Therefore, even though we use terms like attractive, transitional, and source nodes in this paper, readers familiar with the field may recognize the terminology of sink and source nodes.

This interpretation provides a framework for understanding the dynamics of regions within the graph and how they accumulate and retain texts over time.

Statistical significance : The final question to address pertains to determining when the outcome surpasses the income (or vice versa) in significance. For instance, if a node witnesses 200 texts departing while 210 have entered, can we categorize it as an attractive node ?

To address this, we employ a straightforward Binomial test. This test revolves around a random distribution of binary outcomes akin to a coin flip. In our scenario, we consider this random distribution to mimic a fair coin toss, implying that an equivalent amount of income and outcome is expected (akin to heads and tails on a coin). Consequently, we adopt a binomial distribution with a probability of success set at 0.5.

Subsequently, we gauge the probability that this process, repeated N times, would yield M or fewer observed successes. In the case of 210 outcomes and 200 income, the likelihood of such a process producing over 210 outcomes among 410 events is 0.33. Consequently, we would not categorize such a region as a definitive source; rather, it seems to be more of a transitional region.

In our analysis, we choose a significance level of $p < 0.01$ to consider significant the process. Additionally, we apply a Bonferroni correction to the p-value due to the calculation being performed for 50 nodes. This correction brings down the threshold to 0.01/50, or $p < 2\text{e}10^{-4}$.

This approach ensures a methodical determination of whether a region's outcome surpasses its income, shedding light on the regions' behavior and significance within the system.

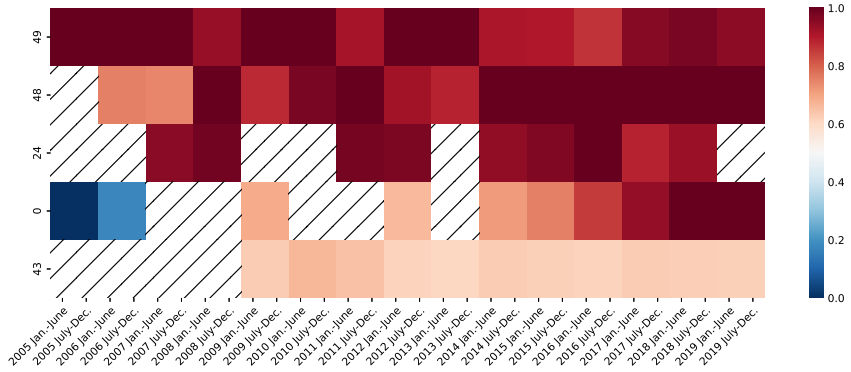


Fig. 2. Time serie of the attractiveness of nodes in the topological graph, ordered by the most active node to the less active. The score indicate the ratio of texts attracted to a node over the number of texts moving into or out of the node, 1 meaning that texts only entered the node region and 0 that texts only went out of this region. The hached part signal that these exchanges were not significant.

2.4 The whole procedure

Putting everything together, we want:

1. Infer the Topology of the data, using an NG

2. Compute the number of texts entering each node and the number of texts leaving each node, for a given time span
3. Compute the p-value that each node is either a sink or a source during this time span

If there is only one prototype, we expect to have only one region with the attractiveness property, or a set of connected regions. If there are several prototypes, we would expect to observe several unconnected regions that appear as attractive at different dates. If there is no prototypical region, we would expect to see no attracting nodes.

2.5 Prototype analysis

If attractive regions are to be identified, two situations can arise. Firstly, attractors may be disconnected in the graph, indicating that they do not account for adjacent regions or, in other words, they do not exhibit a similar textual composition. In this situation, we would consider them as distinct attractive forms of texts. Conversely, attractive regions could be adjacent, implying that the corpora they represent are spread over these regions. In this case, we would consider them to be one attractor.

In either case, we can perform an analysis of the corpora associated with these attractors. Since the NG-graph is produced after dimensionality reduction, interpreting the dimensions might be challenging. To address this, we propose using the Relief algorithm [5], which allows us to contrast features of texts from different clusters to identify those most specific to each cluster. This way, we can understand the procedure as a two-step approach: first using topological approximation of the data to identify prototypical regions, and then analyzing the texts inside these regions to understand what specifically ties them together.

However, as the purpose of this paper is not to conduct an extensive linguistic analysis of these attractors, we will limit ourselves to the first three features and provide a broad overview of their meaning.

2.6 Implementation details

We list here some implementation details. As we've stated, we performed a tSVD on the entire dataset, reducing the dimensions to 75, with the goal of retaining 99% of the explained variance.

Regarding the Neural Gas (NG) implementation, we followed the details outlined in the paper [9] and utilized the following parameters: Firstly, we trained it with 50 nodes over 50,000 iterations (i.e., the number of samples seen). Secondly, we set the values of the parameters λ, η, and the age threshold for edge removal to 10, 0.5, and 15 as initial values, and 2, 0.05, and 50 as final values, respectively. The value of each parameter evolves according to the function $p_i \frac{p_f}{p_i}^{(n/t_{max})}$, where p_i and p_f are the initial and final values of the parameter, n is the iteration number, and t_{max} is the total number of iterations, which in this case is 50,000.

For computing the attraction scores, we filtered our dataset into six-month periods, starting from January-June 2004 to July-December 2022. We also considered the latest version, if available, of texts before these periods and incorporated them into the filtered dataset. Subsequently, we identified the region to which they belong by finding the node center to which they are closest individually. Following this, we examined, article by article, the node transitions they underwent, if any. Finally, we calculated, node by node, how often an article transitioned into and out of each node to compute the p-value for the node's attractiveness.

At last, we've also made the choice to perform what we'll call a trajectory smoothing. Given that modifications in Wikipedia can vary greatly, some being substantial while others minimal, the evolution of our texts aren't continuous and could be likened to the concept of jumping into hyperspace as seen in science fiction where a spaceship vanishes from one point only to reappear at another. If we assume that only *Prototype* forces guide the evolution of texts, then we can posit that the trajectories of similar texts, understood as the time series of the vectors we've created, are similar, even though the degree of variation might differ. Hence, to facilitate smoothing and enable more meaningful comparisons between evolutions of texts, we've adjusted the clusters' transitions to represent the shortest path between the two clusters. If the two nodes are already linked, the textual variation is presented in the same manner, as a transition between the same two nodes. If not, every node that forms the shortest path between the initial two nodes is included in the transitions. This inclusion doesn't substantially alter the result, as the difference between input and output remains the same, but it may serve to reduce the significance of this difference. Indeed, without this procedure, nodes reached by a small fraction of texts can be considered as significantly attractive, when the smoothing allow to make them appear more as a transitive regions.

3 Results

The initial research question we aimed to address concerned the possible existence of attractors within Wikipedia. In Fig. **??**, we present the outcomes of the analysis concerning the nodes' attractiveness within the graph. As is evident, multiple nodes exhibit attracting characteristics, indicating potential *Prototypes*. Another interesting aspect of these results is the dynamics of these attractors, which appear at different times during the process. For instance, we can see node 43 starting to be significantly attractive only in 2009, or node 0 roughly around 2014. Also, node 0 starts as a source and then becomes a sink, indicating a curious dynamic. We can also observe that the first two regions (49 and 48) are almost continuously attractive, while node 24 is intermittently attractive.

Another aspect we considered was the presence of multiple distinct attractors. Although Fig.**??** already displays the existence of several attractors, we postulated that if two attractive regions were connected, they could potentially be treated as a single attractor. The connectivity of the considered regions is

showcased in Fig.1(b). Notably, three disconnected areas arise, allowing us to treat them as distinctive attractive regions as a whole.

Furthermore, the Relief analysis, presented in Tab. 2, allowed us to better understand which features were associated with each *Prototype*. This table shows, for each group of attractive nodes, the most specific features of the texts associated with these nodes, in contrast with the other texts not belonging to these nodes.

Without delving too much into the details, we can already observe a few interesting trends. For instance, the first group related to node 0, which was both a source and a sink, seems to have a strong relationship with sentence construction. Indeed, the *markers* and *open clausal complement* indicate to us that this group is specific in its utilization of subordinates, potentially resulting in long and complex sentences.

On the contrary, the group 24 seems to be specific only by its frequency of Wikipedia-specific features, such as *wikilinks* or the presence or absence of the *Featured article* template.

Lastly, we can observe that the group 48, 49, 43 is specific in its use of modifiers of nominals, indicating a distinct sentence structure, potentially simpler than the one in the group 0.

Table 2. Linguistic Features for Textual Analysis

Nodes	Feature 1	Feature 2	Feature 3
0	DEP : Marker	POS : Pronoun	DEP : open clausal complement
48, 49, 43	DEP : Modifier of nominal	POS : Numeral	Template : Sister project links
24	# Wikilink	Template : Featured article	# Text

4 Discussion

From these results, we draw the following conclusions. First, there are standard ways of writing in Wikipedia, and in fact, there appear to be several. Second, some of these writing styles have remained quite consistent throughout the encyclopedia's history, while others are more sporadic, appearing and disappearing within six-month or one-year spans. Third, based on our observations, it seems that these specific ways of writing strongly revolve around Wikipedia's specific markups or syntactical constructions. However, our present investigations are not sufficient to clearly determine the linguistic and functional interpretation of them.

Through these inquiries, we sought to understand if Wikipedia indeed comprises articles with similar genres and whether various types of texts exist within it. We believe we've demonstrated that there are indeed preferred ways of composing text, associated with genres, but that multiple coexisting styles are present. Furthermore, we've highlighted that these writing preferences can emerge or fade over time, with some remaining stable. Lastly, we've gained a deeper understanding of the preferred writing composition of authors, enabling a more thorough exploration of writer preferences.

While we believe these results provide valuable insights into the writing process of Wikipedia contributors, we also acknowledge that our dataset's limitation to *featured* and *good* articles might lead to the identification of specific standard Wikipedia articles. It's possible that other preferred writing styles exist but aren't represented in these more standardized articles. Moreover, we recognize that the linguistic composition of the text doesn't entirely uncover the complexities that determine textual genre. Additional structural features should be explored to gain a deeper understanding of writers' preference dynamics regarding textual structure.

Finally, we've annotated our data using available algorithms from spaCy, which are known to make mistakes, especially on potentially noisy datasets like ours. Older versions of Wikipedia texts are often in formats that introduce some noise into the data. These errors could have impacted the annotation process and subsequently, the text vectorization, potentially leading to analysis errors.

Considering these aspects, we believe future directions for this work could involve removing noisy features through qualitative analysis, expanding the corpus to include more articles, and incorporating more structural features—either from the discourse structure or page layout. Such enhancements could enrich textual linguistic and psychological studies. We contend that these results directly unveil writers' preferences.

References

1. Biber, D., Conrad, S.: Register, Genre, and Style. Cambridge University Press (2019)
2. Chen, N., Blostein, D.: A survey of document image classification: problem statement, classifier architecture and performance evaluation. International Journal on Document Analysis and Recognition 10, 1?16 (2007). doi:10.1007/s10032-006-0020-2
3. Emigh, W., Herring, S.C: Collaborative Authoring on the Web A Genre Analysis of Online Encyclopedias. In : Proceedings of the Annual Hawaii International Conference on System Sciences 5, pp.99?99. (2005). doi:10.1109/hicss.2005.149
4. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 7(1), pp.411?420. (2017).
5. Kira, Kenji & Larry A. Rendell.: A practical approach to feature selection. Machine learning proceedings 1992. Morgan Kaufmann, pp.249?256. (1992). doi.org/10.1016/B978-1-55860-247-2.50037-1

6. Lagutina, K.V., Lagutina, N.S., Boychuk, E.I.: Text Classification by Genres Based on Rhythmic Characteristics. Automatic Control and Computer Sciences 56, 735?743 (2022). doi:10.3103/S0146411622070136
7. Lee, Y.B., Myaeng, S.H.:Text genre classification with genre-revealing and subject-revealing features. In : Proceedings of the 25tg annual international ACM SIGIR conference on Research and development in information retrieval, pp. 145?150. (2002). doi:10.1145/564376.564403
8. Lieungnapar, A., Todd, R.W., Trakulkasemsuk, W.: Genre induction from a linguistic approach. Indonesian Journal of Applied Linguistics 6, 319?329. (2017). doi:10.17509/ijal.v6i2.4917
9. Martinetz, T., Schulten, K.: A" neural-gas" network learns topologies. (1991)
10. Mirończuk, M. M., Protasiewicz, J.:A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications 106, 36?54. (2018).doi:10.1016/j.eswa.2018.03.058
11. Santini, M.: A Shallow Approach To Syntactic Feature Extraction For Genre Classification. In : Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, pp. 6?7. Birmingham, UK (2004).
12. Shin, C., Doermann, D., Rosenfeld, A.: Classification of document pages using structure-based features. International Journal on Document Analysis and Recognition 3, 232?247. (2001). doi:10.1007/PL00013566
13. Skevik, K.A.: Language Homogeneity in the Japanese Wikipedia. In : Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, pp. 527?534. (2010).
14. Vicente, M., Maestre, M.M., Lloret, E., Cueto, A.S.: Leveraging Machine Learning to Explain the Nature of Written Genre. IEEE Access 9, 24705?24726. (2021). doi:10.1109/ACCESS.2021.3056927
15. Wan, M., Fang, A. C., Huang, C. R. : The discriminativeness of internal syntactic representations in automatic genre classification. Journal of Quantitative Linguistics 28, 138?171. (2021). doi:10.1080/09296174.2019.1663655
16. Wikipedia:Five pillars. https://en.wikipedia.org/wiki/Wikipedia:Five.pillars
17. Wołowski, W.: La sémantique du prototype et les genres (littéraires). Studia Romanica Posnaniensia 33, 65?83. (2006). doi:0.14746/strop.2006.33.005