Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

doi.org/10.51891/rease.v12i9.29935

SEPARATING DOCUMENTS BY LEXICAL REPETITION VIA YULE'S K AND K-MEANS: VALIDATION THROUGH MONTE CARLO SIMULATION

SEPARAÇÃO DE DOCUMENTOS POR REPETIÇÃO LEXICAL VIA K DE YULE E K-MEANS: VALIDAÇÃO POR SIMULAÇÃO DE MONTE CARLO

SEPARACIÓN DE DOCUMENTOS POR REPETICIÓN LÉXICA MEDIANTE LA K DE YULE Y K-MEANS: VALIDACIÓN MEDIANTE SIMULACIÓN DE MONTE CARLO

Raquel Romes Linhares[1]
Regis Nunes Vargas[2]

ABSTRACT: This work proposes and evaluates, through Monte Carlo simulation, a simple methodology for separating textual documents into two groups, high lexical repetition and low lexical repetition, based on Yule's K, a classic measure of vocabulary concentration. The methodology consists of computing Yule's K for each document in a corpus and applying k-means (k = 2) to these values to partition the corpus into the two groups. For each of nine fixed values of the true proportion p (from 0.1 to 0.9), we ran 500 Monte Carlo simulations in which the degree of lexical separation between the two classes is drawn randomly at each repetition, totaling 4,500 simulated corpora. The results show that the method recovers the true proportion with good precision across the whole range of p (mean absolute error of 0.008 and mean classification accuracy of 98.9%), but the error and the variability of the estimate increase systematically at the extreme values of p (close to 0.1 or 0.9), where the minority class is small and more sensitive to scenarios of low separation between the K distributions.

Keywords: Yule's K. K-means. Monte Carlo simulation.

RESUMO: Este trabalho propõe e avalia, por meio de simulação de Monte Carlo, uma metodologia simples para separar documentos textuais em dois grupos, alta repetição lexical e baixa repetição lexical, a partir do K de Yule, uma medida clássica de concentração vocabular. A metodologia consiste em calcular o K de Yule para cada documento de um corpus e aplicar k-means (k = 2) sobre esses valores para particionar o corpus nos dois grupos. Para cada um de nove valores fixos da proporção verdadeira p (de 0,1 a 0,9), foram geradas 500 simulações de Monte Carlo nas quais o grau de separação lexical entre as duas classes é sorteado aleatoriamente a cada repetição, totalizando 4.500 corpora simulados. Os resultados mostram que o método recupera a proporção verdadeira com boa precisão em toda a faixa de p (erro absoluto médio de 0,008 e acurácia média de classificação de 98,9%), mas o erro e a variabilidade da estimativa aumentam sistematicamente nos valores extremos de p (perto de 0,1 ou de 0,9), onde a classe minoritária é pequena e mais sensível a cenários de baixa separação entre as distribuições de K.

Palavras-chave: K de Yule. K-means. Simulação de Monte Carlo.

[1]Professora Doutora, Instituto de Matemática e Estatística - Universidade Federal de Uberlândia (IME-UFU).
[2]Doutor em Engenharia Elétrica. Faculdade de Engenharia Elétrica - Universidade Federal de Uberlândia.

RESUMEN: Este trabajo propone y evalúa, mediante simulación de Monte Carlo, una metodología simple para separar documentos textuales en dos grupos, alta repetición léxica y baja repetición léxica, a partir de la K de Yule, una medida clásica de concentración de vocabulario. La metodología consiste en calcular la K de Yule de cada documento de un corpus y aplicar k-means (k = 2) sobre esos valores para particionar el corpus en los dos grupos. Para cada uno de nueve valores fijos de la proporción verdadera p(de 0,1 a 0,9), se generaron 500 simulaciones de Monte Carlo en las que el grado de separación léxica entre las dos clases se sortea aleatoriamente en cada repetición, totalizando 4.500 corpus simulados. Los resultados muestran que el método recupera la proporción verdadera con buena precisión en todo el rango de p (error absoluto medio de 0,008 y exactitud media de clasificación del 98,9%), pero el error y la variabilidad de la estimación aumentan sistemáticamente en los valores extremos de p (cerca de 0,1 o de 0,9), donde la clase minoritaria es pequeña y más sensible a escenarios de baja separación entre las distribuciones de K.

Palabras clave: K de Yule. K-means. Simulación de Monte Carlo.

INTRODUCTION

Repetition and lexical richness are quantitative properties of text that have historically been explored in several areas of computational linguistics, including authorship attribution, textual complexity assessment, second-language acquisition, and, more recently, detection of text generated by language models. Texts with vocabulary strongly concentrated on a few words tend to be perceived as more repetitive and lexically less sophisticated, whereas texts with more dispersed vocabulary tend to be perceived as richer.

A recurring problem in corpus analysis is separating documents into groups of “high” and “low” lexical repetition without relying on pre-existing labels, either because such labels do not exist or because the goal is precisely to discover this structure in an exploratory way. Addressing this problem requires, first, a way to quantify lexical repetition, and second, a way to partition documents based on that quantity.

Yule's K (YULE GU, 1944) is widely used as a measure of lexical richness/repetition in studies of authorship attribution and textual complexity. K-means, in turn, is a fundamental unsupervised machine learning algorithm designed to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean (centroid), thereby minimizing intra-cluster variance (MACQUEEN J, 1967). Many works in the literature use k-means as an unsupervised clustering technique to separate datasets into distinct groups based on their similarity characteristics (LINHARES RR, VARGAS RN, 2025; LINHARES RR, VARGAS RN, 2026). Here, we propose a method that combines Yule's K and k-means to separate corpora into high and low lexical repetition groups.

A direct practical motivation for the proposed method lies in the curation and cleaning of corpora used to train language models. Corpora scraped from the web frequently contain "degenerate" documents, pages with repetitive text, search-engine-optimization spam, duplicated content, or content generated from automated templates. This concern is not limited to scraped web data: recent work has shown that LLM-based rewriting itself systematically shifts vocabulary richness and other stylometric markers relative to human-written text (VAN NUENEN T, 2026), suggesting that corpora increasingly mixing human and machine-produced text may exhibit drifting lexical-repetition profiles over time. Since such documents are computationally cheap to detect but costly to find manually at scale, an unsupervised method that requires no prior labels, capable of automatically separating high lexical repetition documents from the rest of the corpus, can function as an automatic quality filter. Documents identified as high lexical repetition can then be discarded or flagged for manual review before being used in pre-training or fine-tuning stages, reducing the risk of the model learning and reproducing unwanted repetitive patterns.

This work aims to validate, through Monte Carlo simulation, a method that combines Yule's K and k-means to separate corpora into high and low lexical repetition groups.

METHODOLOGY

We operationally define a document as having high lexical repetition when its word frequency distribution is strongly concentrated on a few types, that is, a small fraction of the vocabulary accounts for a large share of the word occurrences (tokens) in the text. In contrast, a low lexical repetition document displays a more dispersed frequency distribution, in which occurrences are spread across a larger and more balanced set of types.

Yule's K. Yule's K (YULE GU, 1944) is a classic measure of vocabulary concentration, defined from the frequency spectrum of a text:

$$K = 10^4 \cdot \frac{(\Sigma {f_i}^2 - N)}{N^2}$$

where N is the total number of tokens in the document and f_i is the occurrence frequency of the i-th word type in the document. The higher the value of K, the more concentrated the vocabulary (high repetition); the lower the value, the more dispersed the vocabulary (low repetition). Unlike simpler measures such as the type-token ratio (TTR), Yule's K was designed to be relatively independent of text length.

K-means. K-means (MACQUEEN J, 1967; LLOYD S, 1982) is an unsupervised clustering algorithm that partitions a set of observations into k groups, assigning each observation to the group whose centroid is closest, and iteratively recalculating the centroids until convergence. When applied to a single variable (the one-dimensional case, such as Yule's K per document), k-means with k = 2 is equivalent to finding the cut-off point that best separates the distribution into two groups, minimizing the variance within each group.

Proposed methodology. Given a corpus of textual documents, the proposed methodology consists of four steps: (1) compute Yule's K for each document, from its word frequency distribution; (2) apply k-means with k = 2 to the one-dimensional vector of K values, obtaining two clusters; (3) label the cluster with the higher mean K as "high lexical repetition" and the other as "low lexical repetition"; (4) estimate the proportion of high lexical repetition documents as the fraction of documents assigned to the "high repetition" cluster.

Data source: synthetic corpus generation. Since real corpora rarely carry an objective label indicating the true proportion of repetitive documents, validation was performed on synthetic data with a known composition. Each synthetic document is generated by a model based on Zipf's law. A fixed vocabulary of $V = 1000$ word types is defined, each receiving a rank r from 1 (most frequent) to V (least frequent). Each type of rank r is assigned an occurrence probability $P(r)$, proportional to r raised to the power s:

$$P(r) = \frac{\frac{1}{r^s}}{\sum_{j=1}^{V} \frac{1}{j^s}} \quad , r = 1,2,\ldots,V$$

where $s > 1$controls the degree of concentration: values close to 1 produce dispersed vocabulary (low repetition), and larger values concentrate occurrences on a few low-rank types (high repetition). Document length N is drawn around a base value $N_0 = 400$ tokens, with a random variation of up to ±15% (N = round($N_0 \cdot$ (1 + U(−0.15, 0.15)))). Each document is then built by independently drawing N tokens according to $P(r)$,and the frequencies f_i of the resulting types are used to compute Yule's K.

Sampling and simulation design. No human or animal subjects were involved; the study relies entirely on computer-generated synthetic data, so no research ethics committee approval was required. Each synthetic corpus contains 300 documents, mixing a fraction p (true proportion, known by construction) generated with exponent s_{high} (high repetition class) and

a fraction $(1 - p)$ generated with exponent s_{low} (low repetition class); true labels are randomly shuffled among corpus positions. The Monte Carlo design fixed p at nine equally spaced values (0.1, 0.2, ..., 0.9) and, for each value, generated 500 independent corpora in which s_{low} and s_{high} are randomly drawn at each repetition, s_{low} from a uniform distribution between 1.00 and 1.15, and s_{high} from a uniform distribution between 1.15 and 2.00, totaling 4,500 simulated corpora. Drawing the class separation at each repetition, rather than fixing it, is what characterizes this simulation as Monte Carlo in the strict sense.

Analytical procedures. All simulations and calculations were performed in Python, using NumPy, scikit-learn, and Matplotlib. For each corpus, we computed Yule's K per document and applied the proposed method (k-means, k = 2), obtaining the estimated proportion, the document-by-document classification accuracy (after aligning cluster labels to the true labels via the correspondence that maximizes agreement), and the Adjusted Rand Index (ARI). For each value of p, the 500 results were aggregated into: mean and standard deviation of the estimated proportion; mean bias and mean absolute error (MAE); root mean squared error (RMSE); mean accuracy; mean ARI; and the percentage of simulations with absolute error below 5 percentage points.

RESULTS

Table 1 and Figures 1-2 summarize the results of the 4,500 simulated corpora. Across the whole range of proportions tested, the estimated proportion closely tracked the true proportion, with MAE ranging between 0.002 and 0.028 and classification accuracy always above 97% (Table 1).

Table 1- Monte Carlo simulation results by true proportion p_{true} (500 simulations per row; $s_{low} \sim U(1.00,1.15)$ and $s_{high} \sim U(1.15,2.00)$ drawn at each simulation). Source: the authors (2026).

p_{true}	$\hat{p}$ (mean)	SD	Bias	MAE	RMSE	Accuracy	ARI	% error<5pp
0.1	0.110	0.050	0.010	0.011	0.051	0.988	0.961	95.6%
0.2	0.203	0.021	0.003	0.004	0.021	0.994	0.978	97.4%
0.3	0.301	0.012	0.001	0.002	0.012	0.994	0.980	99.0%
0.4	0.400	0.007	-0.000	0.002	0.007	0.992	0.973	99.6%
0.5	0.498	0.010	-0.002	0.003	0.011	0.993	0.976	98.8%
0.6	0.596	0.017	-0.004	0.004	0.018	0.994	0.978	97.4%
0.7	0.691	0.037	-0.009	0.009	0.038	0.988	0.963	94.8%

p_{true}	$\hat{p}$ (mean)	SD	Bias	MAE	RMSE	Accuracy	ARI	% error<5pp
0.8	0.786	0.057	-0.014	0.014	0.058	0.984	0.952	93.4%
0.9	0.872	0.092	-0.028	0.028	0.096	0.972	0.920	90.6%

The estimated proportion in each of the 500 simulations per value of p_{true} is shown in Figure 1, together with the mean and the ± 1 standard deviation band.

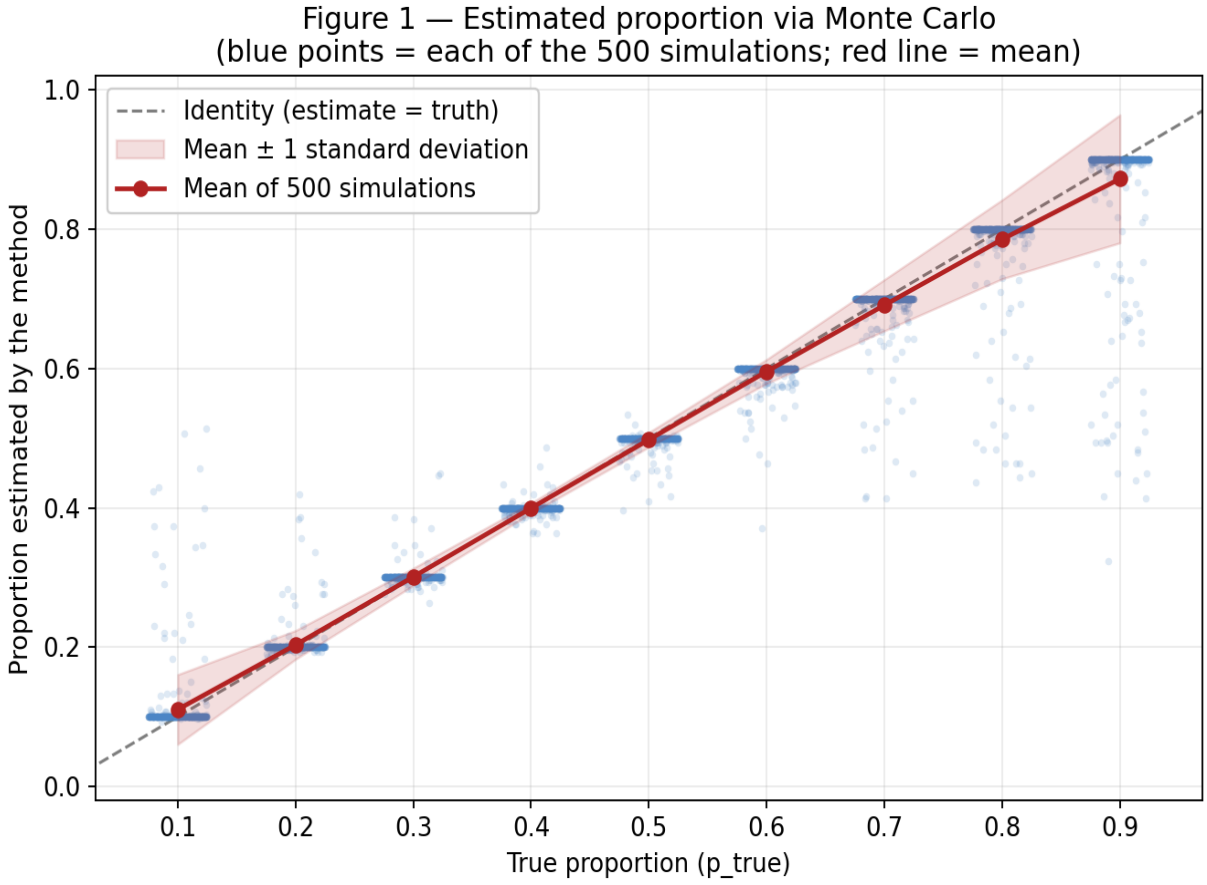

Figure 1 - Estimated proportion via Monte Carlo (blue points = each of the 500 simulations; red line = mean; shaded band = ± 1 SD). Source: the authors (2026).

The mean absolute error (MAE) and the mean classification accuracy as a function of p_{true} are shown in Figure 2, revealing a "U"-shaped error pattern, with minimum error near p_{true} = 0.3-0.4 and maximum error at the extremes (p_{true} = 0.1 and, especially, p_{true} = 0.9).

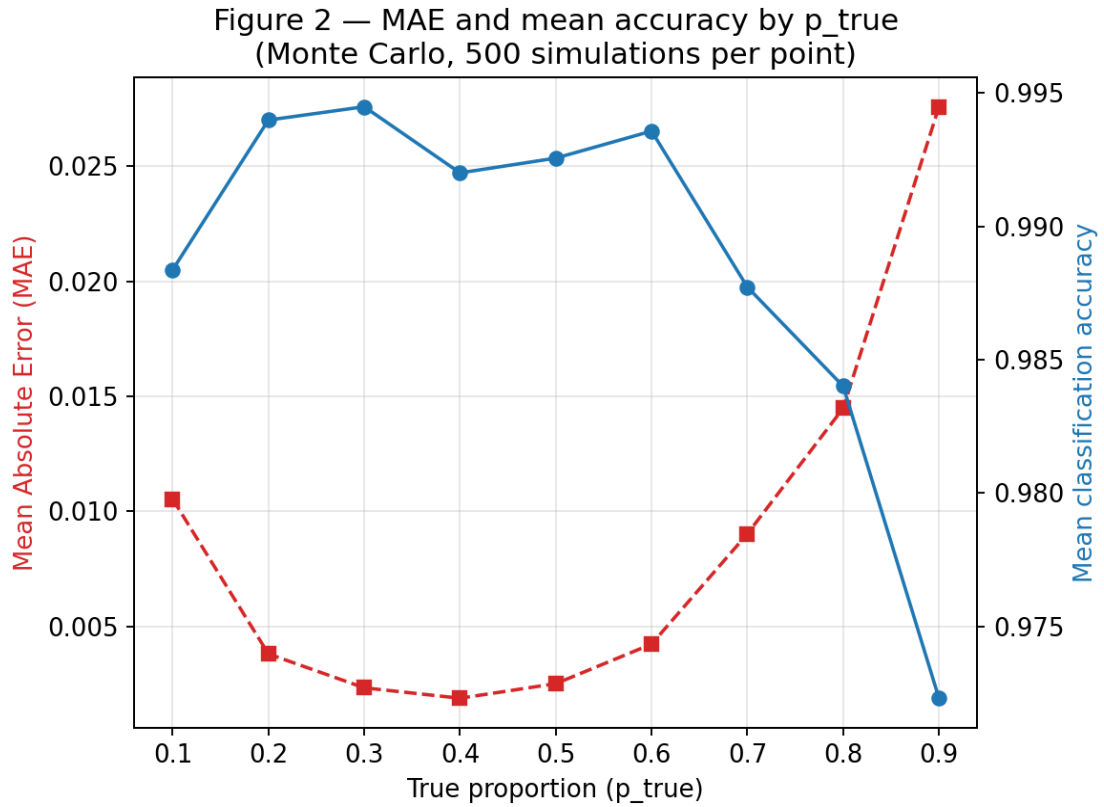

Figure 2 - Mean absolute error (MAE, left) and mean classification accuracy (right) as a function of p_{true}.
Source: the authors (2026).

Across the whole range of proportions tested, the method recovered the true proportion with good precision, confirming that, when integrating over a plausible range of separation scenarios between the classes, the Yule's K + k-means pipeline does not present a gross bias in estimating the proportion of high lexical repetition documents.

The most informative pattern in the simulation is the relationship between the error and the position of p_{true} within the (0, 1) interval: the MAE follows a "U" shape, with minimum values close to p_{true} = 0.3 and 0.4 (MAE ≈ 0.002) and maximum values at the extremes, especially at p_{true} = 0.9 (MAE ≈ 0.028, almost 15 times larger than the minimum). The standard deviation of the estimated proportion follows the same pattern, growing from 0.010 at p_{true} = 0.4 to 0.092 at p_{true} = 0.9. Moreover, the mean bias becomes systematically negative and grows in magnitude at high values of p (bias of −0.028 at p_{true} = 0.9), indicating a tendency of the method to underestimate the proportion when the high repetition class is an overwhelming majority of the corpus.

It is important to highlight the limitations of the synthetic model used. Generating documents via a Zipfian distribution controlled by a single parameter (the exponent s) is a considerably simpler simplification than the actual structure of a natural-language text, which involves collocations, syntax, text genre, and other sources of lexical variation not captured by the model. The validation presented here establishes that the method is internally consistent and reasonably accurate under controlled conditions, but it does not replace an empirical validation on real corpora with some known proxy label for lexical repetition, for example, corpora of second-language learners (in which repetition tends to be higher than in native speakers) or corpora of language-model-generated text versus human text.

CONCLUSION

This work proposed a simple, non-parametric methodology for separating textual documents into high and low lexical repetition groups, combining Yule's K, a classic measure that is relatively independent of text lengt, with one-dimensional k-means (k = 2) as an exploratory clustering method. A direct practical application of this methodology lies in the curation of corpora for training language models, as an automatic filter to identify and flag degenerate or excessively repetitive documents.

Validation through Monte Carlo simulation, 4,500 synthetic corpora, covering nine true proportions between 0.1 and 0.9, each evaluated under 500 randomly drawn class-separation scenarios, demonstrated that the method recovers the true proportion with good precision across the entire range tested (mean MAE of 0.008; mean accuracy of 98.9%). At the same time, the simulation revealed a relevant systematic pattern: both the error and the variability of the estimate increase at the extreme values of p_{true}, where the minority class is small and more vulnerable to unfavorable separation scenarios between the K distributions.

REFERENCES

LLOYD S. Least squares quantization in PCM. IEEE Transactions on Information Theory, 1982; 28(2): 129-137.

LINHARES RR, VARGAS RN. DFA-WK: an integrated approach to estimate the Hurst exponent using DFA, K-means, and Wavelet Transform. Cadernos Cajuína, 2025; 10(5): e1374.

LINHARES RR, VARGAS RN. Non-parametric density estimation based on Kernel, Wavelets, and K-Means. Revista DCS, 2026; 23(92): e6014.

MACQUEEN J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967; 1: 281-297.

VAN NUENEN T. Voice under revision: large language models and the normalization of personal narrative. arXiv preprint arXiv:2604.22142, 2026.

YULE GU. The statistical study of literary vocabulary. Cambridge: Cambridge University Press, 1944.