Towards Explaining the Spectrogram of Graph Spectral Clustering in Text Document Domain

Mieczysław A. Kłopotek[0000−0003−4685−7045] and
Sławomir T. Wierzchoń[0000−0001−8860−392X] and
Bartłomiej Starosta[0000−0002−5554−4596] and
Dariusz Czerski[0000−0002−3013−3483] and Piotr Borkowski[0000−0001−9188−5147]

Institute of Computer Science, Polish Academy of Sciences,01-238 Warsaw, ul.Jana Kazimierza 5, Poland, klopotek, stw, barstar, dcz, pborkowski@ipipan.waw.pl
https://ipipan.waw.pl/

Abstract. In previous research, the authors found that the spectrogram of eigenvalues of combinatorial Laplacian of the document similarity matrix is relevant for tasks like graph spectral classification, clustering etc. This paper investigates the hypothesis that this property can be attributed to the specific "style" of writing, that is to the distribution of words in the documents belonging to a given category of documents. The investigation is performed via generating artificial documents from a predefined parameterized word distribution. The document similarity matrices are computed and the spectrum of the corresponding combinatorial Laplacian is interrogated. The parameters are varied to determine their impact. We present the impact of these parameters on the shape of the spectrogram.

Keywords: Explainable AI · Graph Spectral Clustering · Eigenvalue Spectrum of A Laplacian · Artificial Text Generation from Simple Language Model

1 Introduction

The aim of this work is to investigate the causes of specific shapes of Laplacian combinatorial spectrograms. This should give an insight into understanding the results of Graph Spectral Clustering and other spectral clustering methods [13, 7]. Explainable Artificial Intelligence (XAI) is a hot topic for years now [3]. The driving factor for its emergence was the “black-box” nature of many AI methods which was not quite acceptable in business settings. This is especially true of Graph Spectral Analysis (GSA) methods in which the analysis results are expressed in terms of eigenvectors and eigenvalues [13, 14, 21].

Earlier research in GSA concentrated on exploring a few eigenvalues and eigenvectors [13]. However, in our earlier research we discovered that one can exploit also the full eigenvector spectrogram of eigenvalues and ignore the eigenvectors when performing classification [5], incremental clustering [11], investigating hashtag similarity [19], and other. Encouraged by these results, this paper

investigates the hypothesis that the possibility to characterize clusters/class via spectrograms can be attributed to the specific "style" of writing, that is to the distribution of words in the documents belonging to a given category of documents.

The investigation, outlined in detail in Section 3 is performed via generating artificial documents from a predefined parameterized word distribution. The document similarity matrices are computed and the spectrum of the corresponding combinatorial Laplacian is interrogated. The parameters are varied to determine their impact. The experimental results are presented in Section 4. Section 2 overviews related work, while Section 5 presents our conclusions and outlines future research.

2 Related Work

The traditional way to perform graph spectral clustering is based on the relaxation of ratio cut (RCut) graph clustering methods. The k-means algorithm is applied to the rows of the matrix, the columns of which are eigenvectors associated with the k lowest eigenvalues of the corresponding graph Laplacian of a similarity matrix [12]. For a similarity matrix S between pairs of items (e.g. documents), a combinatorial Laplacian L is defined as

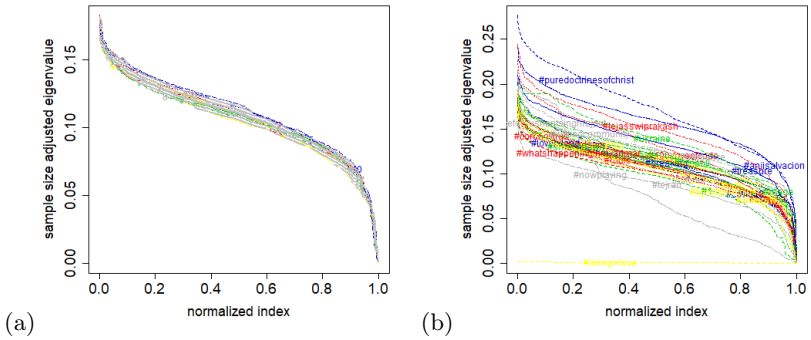
$$L(S) = T(S) - S, \tag{1}$$

where $T(S)$ is the diagonal matrix with $t_{jj} = \sum_{k=1}^{n} s_{jk}$ for each $j \in [n]$. The RCut criterion means finding the partition matrix $P_{RCut} \in \mathbb{R}^{n \times k}$ that minimizes the formula $H'LH$ over the set of all partition matrices $H \in \mathbb{R}^{n \times k}$. As the problem is NP-hard, the relaxation is made assuming that H is a column orthogonal matrix. Then the solution is simple: the columns of P_{RCut} are eigenvectors of L corresponding to the k smallest eigenvalues of L. Further details can be found in e.g. [12] or [20].

Eigenvalues of a combinatorial Laplacian are always non-negative, with the smallest one being equal zero. The figure 1 presents real spectrograms in so-called normalized form (indexes of eigenvalues and eigenvalues themselves were divided by the collection size). The reason for normalization was that the shapes of spectrograms depend on the document collection size [5], so the normalization is necessary if different collections are to be compared. In other figures, we show indexes and eigenvalues "as is", because our experiments with artificial data will use the same fixed size of document samples.

The similarity between textual documents is usually computed as cosine similarity between bag-of-words representations of the documents (see e.g. [20]). Hence, for simulation purposes, it is sufficient to consider models of word distributions. One of the earliest proposals of word distribution functions was so-called Zipf law [23]. The formulation of this law was that the probability of occurrence of a word w_i amounts to

$$Prob(w_i; \alpha) = \frac{\frac{1}{i^\alpha}}{\sum_{\ell=1}^{n_w} \frac{1}{\ell^\alpha}} \tag{2}$$



(a) (b)

Fig. 1. Real spectrograms. (a) (overlapped) spectrograms of samples from the same population, (b) (overlapped) spectrograms of samples from different populations. Spectrograms were normalized (indexes of eigenvalues, X-axis, were divided by the number of documents in the collection; also the eigevalues, Y-axis, were divided by the number of documents; see [19]).

where i ranges from 1 to n_w, where n_w is the number of words in the dictionary, and α is a parameter, usually set to 1. Among its generalizations, let us mention that of Mandelbrot [15], where the distribution is proportional to:

$$Prob(w_i; \alpha, b) = \frac{\frac{1}{(i+b)^\alpha}}{\sum_{\ell=1}^{n_w} \frac{1}{(\ell+b)^\alpha}} \tag{3}$$

where b is a kind of distribution shift parameter, usually $\alpha \approx 1$ and $b \approx 2.7$.

Other word distribution models have been proposed in the past, including extensions of Zipf's law [16], the lognormal model [6] or generalized inverse Gauss-Poisson law [18]. For a comparatory study of some of them see e.g. [1].

One of the most serious problems of cluster analysis in general [2] and of GSA in particular is the explainability of the results, though recent years have brought visible progress in this area. [8] propose an exemplar-based approach to clustering explanation which may be suitable for various embedding types, like auto-encoders or word embeddings. [4] presents a similar idea, but rather based on prototypes. [17] concentrate on explanations via relevant keywords. [9] suggest a quite universal method for text cluster explanation, based on creating an equivalent neural network model for a given clustering of text document. [22] use a (hidden variable) probabilistic model with the detection of hidden topics generating word pairs to perform clustering into topics and then to describe the topics by the distribution of word pairs implied by the topic.

3 Experimental Settings

The investigation was performed to identify a generative model of artificial documents from a predefined parameterized word distribution that would be similar to the real ones. The real shapes of spectrograms are visible in Figure 1 and come from the research described in [19]. The left spectrograms are a superposition of spectrograms coming from samples from the tweets having the very same hashtag in common. The curves are quite close to each other. The right one is a superposition of samples of tweets where each sample stems from a different hashtag. We see that these curves differ even very strongly.

The detailed goal of our experiments was twofold (1) qualitative: find a generative model for documents, yielding eigenvalue spectrum of the combinatorial Laplacian similar in shape to the real spectrograms, (2) quantitative: find generative model parameters impacting the generated spectrogram significantly.

A generator was created, producing artificial documents, consisting of a bag of words sampled from the dictionary according to various parameters of word distribution and other features of documents. For each set of generated documents, the document similarity matrices are computed and the spectra of the corresponding combinatorial Laplacian were interrogated. The parameters are varied from set to set to determine their impact.

We worked with the Zipf-Mandelbrott model (formula (3)) and checked the following parameters:

- n_{ob} - number of documents (fixed for most experiments),
- n_w - dictionary size
- b - Zipf-Mandelbrot distribution parameter,
- α - Zipf distribution parameter,
- $doclen0$ - document basic length
- σ - document length distribution standard deviation (set to zero when dealing with fixed length documents).

In the experiments, one parameter was changed at a time, while the other ones were kept at default level. Default parameters were: $n_{ob} = 1600$, $n_w = 1000$, $b = 0$, $\alpha = 1$, $doclen0 = 60$, $\sigma = 0.3$. Table 1 lists the parameter value ranges used in the experiment.

Parameter	**Value Range**
n_{ob}	{200,400,800,1600}
n_w	{800,1000,1200,1400,1600}
b	$\{0, 0.7, 2.7, 4.7, 6.7\}$
α	{0,0.9,1,1.1, 1.2}
$doclen0$	{ 30,60,120,240,480}
σ	{0, 0.1,0.2,0.3,0.4,0.5,0.7,1, 2}

Table 1. Ranges of parameters used in the experiments. %

Two sets of experiments were performed: one with fixed document length (Section 4.1) and the other with varying document length (Section 4.2).

In the latter set, it is assumed that the distribution of document length in terms of words follows normal distribution with mean $doclen0$ and standard deviation $\sigma * doclen0$ and truncated to the range $(0.5, 1.5) * doclen0$. We tried out two other distributions of document length: fixed document lengths, and uniform length distribution. In the Figures, $\sigma = 0$ emulates fixed length of all documents, and $\sigma = 2$ is an approximation of uniform distribution over an interval. $b = 0$ means the Zipf distribution.

4 Results

4.1 Fixed Length Documents

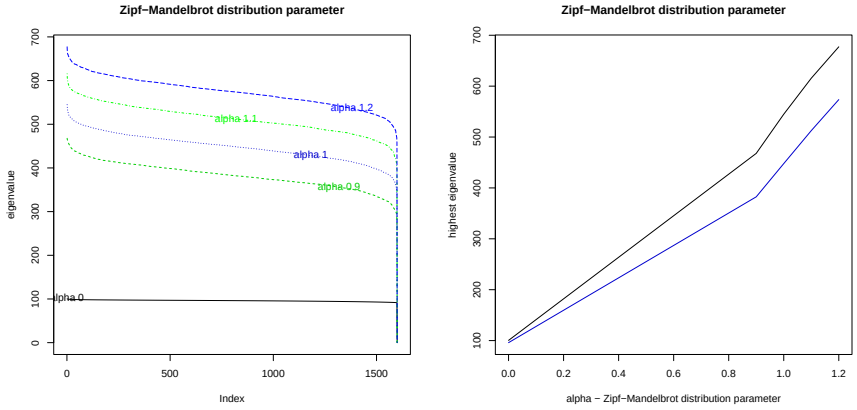


Fig. 2. Spectrogram dependence on α for artificial data generated based on probability of vocabulary $\frac{1}{(n_w+b)^\alpha}$ according to Zipf-Mandelbrot distribution; Zipf distribution is given by $b = 0$ and $\alpha = 1$; uniform is if $\alpha = 0$.

In these experiments, the document length was fixed to 60 words, except for Fig.4, where it was varied. All the experiments, presented in Figs 2, 3, 4 and 5 were performed 5 times, and each time the spectrograms were (nearly) identical. Therefore the results from repetitions were not presented. In the mentioned figures, the left figure contains the median spectrograms from these 5 runs. The right figure shows the dependence of the highest eigenvalue from the left one (black line) and of the average eigenvalue (blue line) on the inspected parameter value.

From these presentations we see that when varying α (Fig. 2) the increase of α moves the spectrograms upwards. When varying b (Fig. 3) we see that the

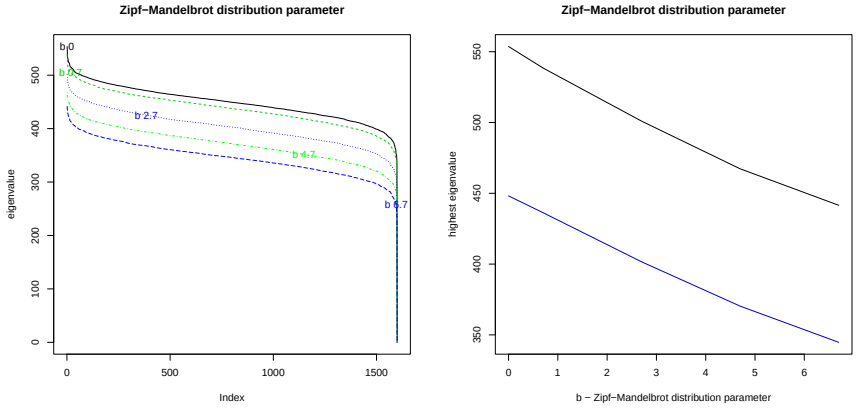


Fig. 3. Spectrogram dependence on b for artificial data generated based on probability of vocabulary $\frac{1}{(n_w+b)^\alpha}$ according to Zipf-Mandelbrot distribution.

increase of b moves the spectrograms downwards. Varying $doclen0$ (Fig. 4) we

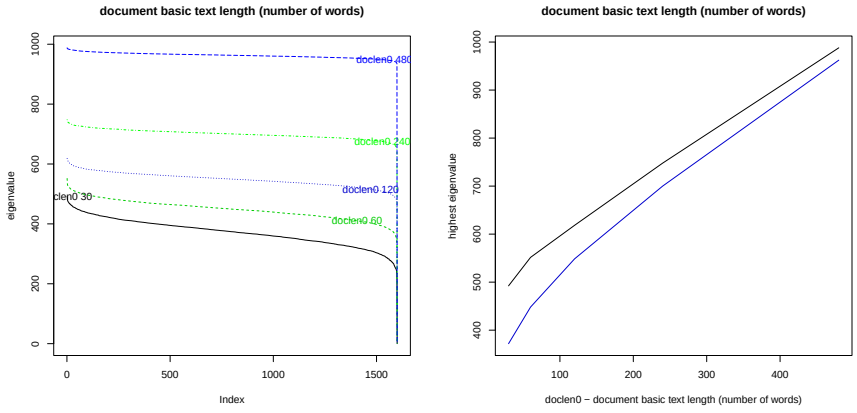


Fig. 4. Spectrogram dependence on document length for artificial data generated.

see that the increase of $doclen0$ moves the spectrograms upwards. Lastly, we observe that the increase of n_w moves the spectrograms downwards.

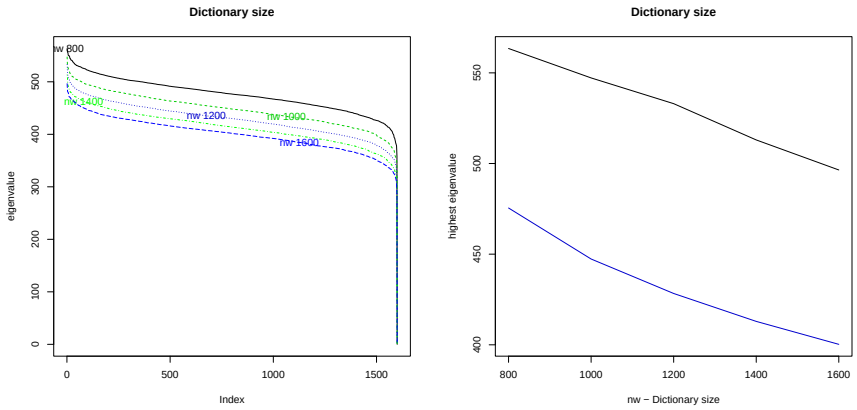


Fig. 5. Spectrogram dependence on number of words in the dictionary for artificial data generated

4.2 Varying Length Documents

In these experiments, the document length was varied according to the normal distribution follows normal distribution with mean $doclen0 = 60$ and standard deviation $\sigma * doclen0$, where $\sigma = 0.3$ and truncated to the range $(0.5, 1.5) * doclen0$, except for Fig.9, where $doclen0$ was varied and Fig.11, where σ was varied. All the experiments, starting with those presented in Fig.6, were performed 15 times and the left figures in Figs 6, 8, 9, 11 and 13 contain the median spectrograms, and the right ones present the highest eigenvalue (black line) and mean eigenvalue (blue line) for each parameter value.

Fig. 6 presents the experimental results – the median spectrograms. The right figure in Fig. 6 shows the dependence of the highest/mean eigenvalue from the left one depending on the α parameter value. For α we see that there are differences in spectrograms on changing α. The increase of α moves the spectrograms upwards. This is the very same pattern as with fixed-length documents. Note that this time, with varying document length, the individual spectrograms differ. Detailed spectrograms of the 15 runs for each parameter value α are visible in Fig. 7. Dot-lines there are median, 25 and 75 percentiles resp. In analogous way to Fig.6, the Figures 8, 9, 12, 13 and 11 are laid out. Only for parameter $doclen0$, an analogy of Fig.7, that is Fig. 10 is presented, because the other parameters do not have the impact that we are interested in.

The experiments for the parameter b of the Zipf-Mandelbrot law (see formula (3)) are summarized in Fig.8. This parameter ranged over the set of values $\{0, 0.7, 2.7, 4.7, 6.7\}$. As one sees, the parameter b does not have such an impact as the parameter α. This is different from the case of fixed length documents. The variation of spectrograms due to variation of document length blurs the relationship.

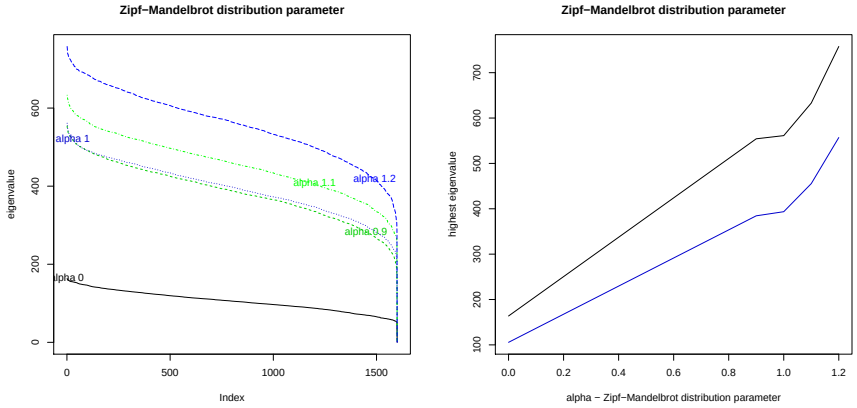


Fig. 6. Spectrogram dependence on α for artificial data generated based on probability of vocabulary $\frac{1}{(n_w+b)^\alpha}$ according to Zipf-Mandelbrot distribution; Zipf distribution is given if $b = 0$ and $\alpha = 1$; uniform is if $\alpha = 0$.

The experiments for the parameter *doclen*0 (average document length) are summarized in Fig.9. its value ranged over the set { 30,60,120,240,480}. As one sees, increase of the *doclen*0 moves the spectrograms upwards. Detailed spectrograms of the 15 runs for each parameter value α are visible in Fig. 10. Dot-lines are median, 25 and 75 percentiles resp. The pattern is the same as with fixed-length documents from previous subsection.

The experiments with varying the number of objects in the sample n_{ob} are shown in Fig.12. Sizes were chosen from the set {200,400,800,1600}. The phenomenon, described extensively in [5], is visible here: samples of various sizes from the same document set are scaled by the sample size.

The experiments with varying dictionary size n_w are shown in Fig.13. The dictionary size ranged over {800,1000,1200,1400,1600}. No clear dependence can be derived from these experiments, contrary to fixed-length documents from previous subsection. Apparently, the variation of document length disturbs the otherwise evident dependence.

5 Conclusions

We have studied the dependence of spectrograms of combinatorial Laplacian on several parameters of document collections generated artificially from widely accepted models of word distributions. We have studied two different settings: fixed length documents and varying-length documents. Under the fixed length setting, we detected that four parameters of generated document sets impact the shape of the spectrograms. However, if we take into account the variability of the document length, only the parameters α and *doclen*0 seem to influence the shape

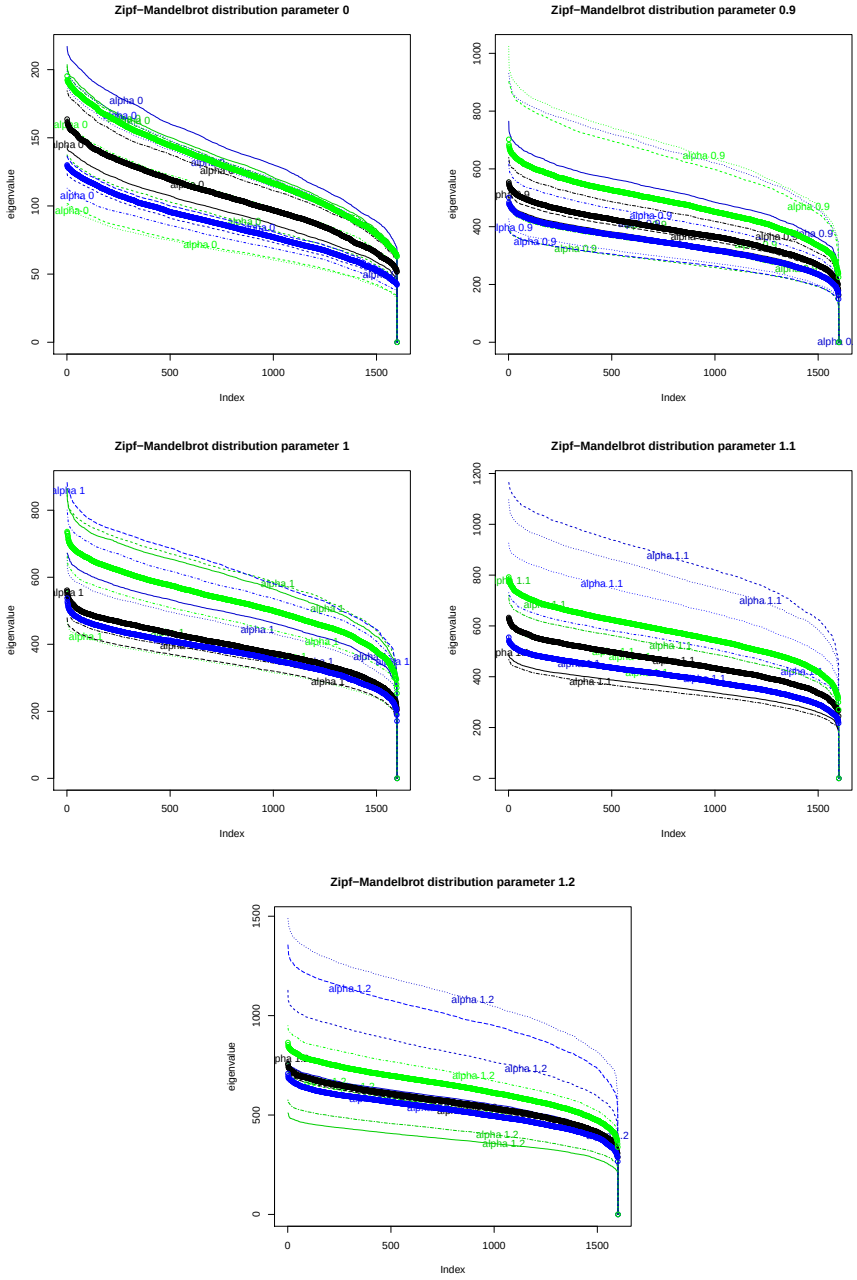


Fig. 7. Spectrogram dependence on α for artificial data generated based on probability of vocabulary $\frac{1}{(n_w+b)^\alpha}$ according to Zipf-Mandelbrot distribution; Zipf distribution is given if $b = 0$ and $\alpha = 1$; uniform is if $\alpha = 0$. Details of 15 runs for each parameter.

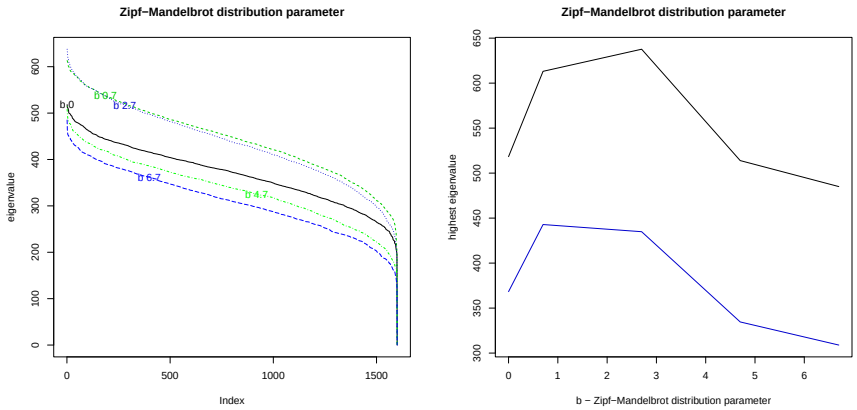


Fig. 8. Spectrogram dependence on b for artificial data generated based on probability of vocabulary $\frac{1}{(n_w+b)^\alpha}$ according to Zipf-Mandelbrot distribution.

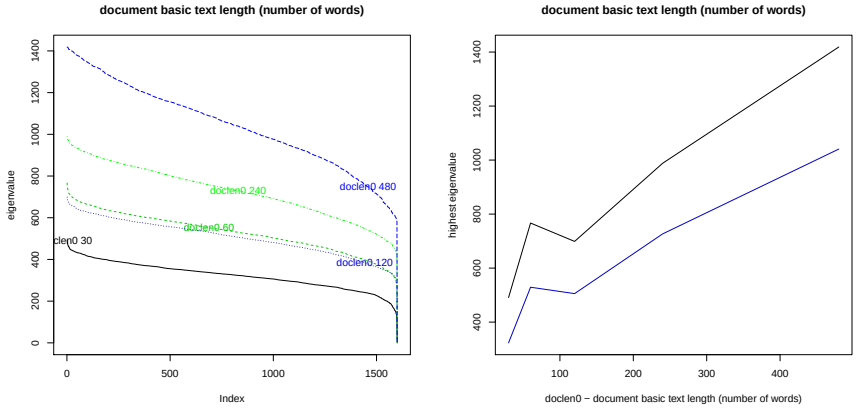


Fig. 9. Spectrogram dependence on average base document length for artificial data generated.

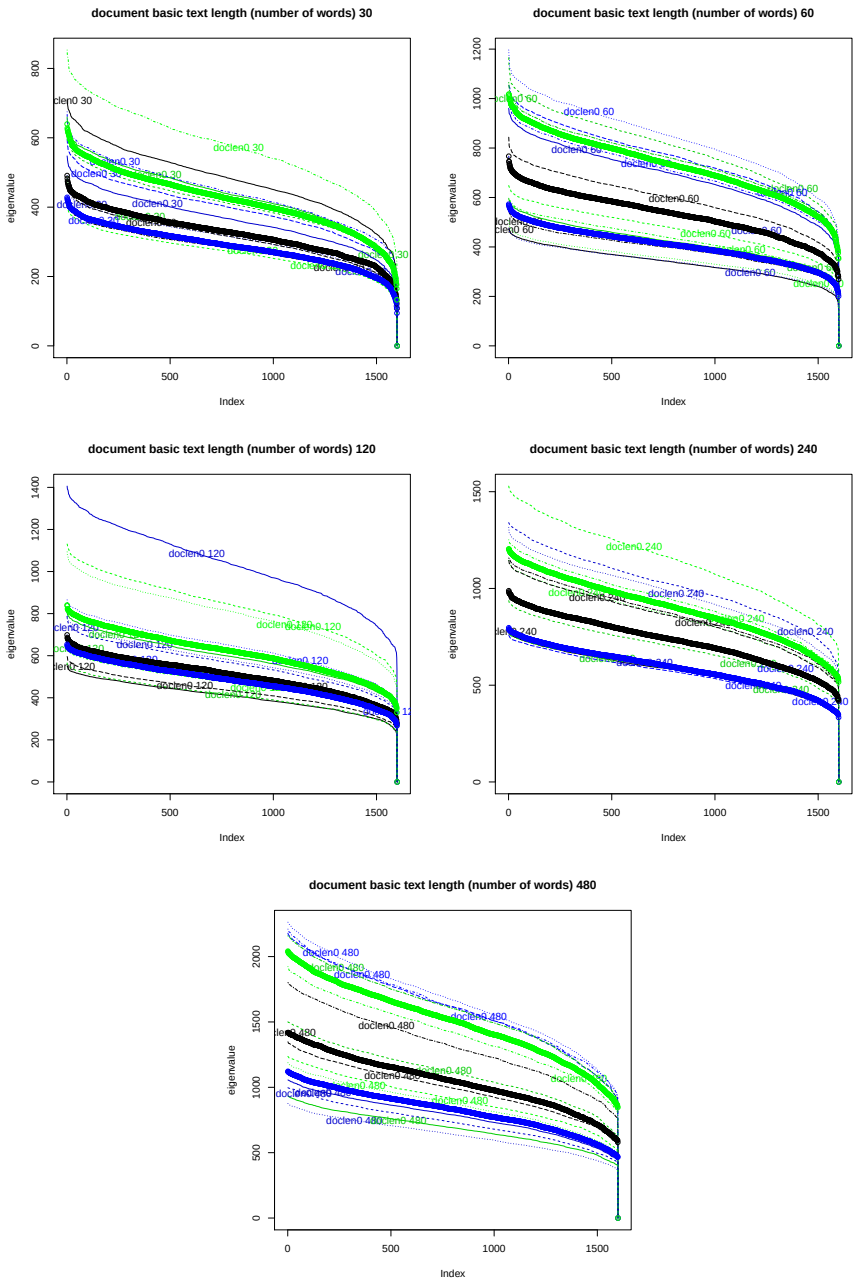


Fig. 10. Spectrogram dependence on average base document length for artificial data generated. Details.

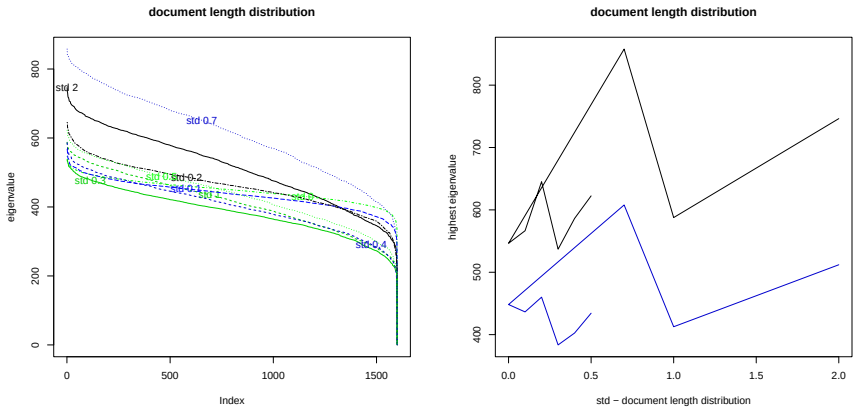


Fig. 11. Spectrogram dependence on varying document length following truncated normal distribution for artificial data generated; with respect to $\mu = 1$ for range $(0.5, 1.5)$; hereby $\sigma = 0$ means fixed document length, $\sigma \to \infty$ uniform distribution.

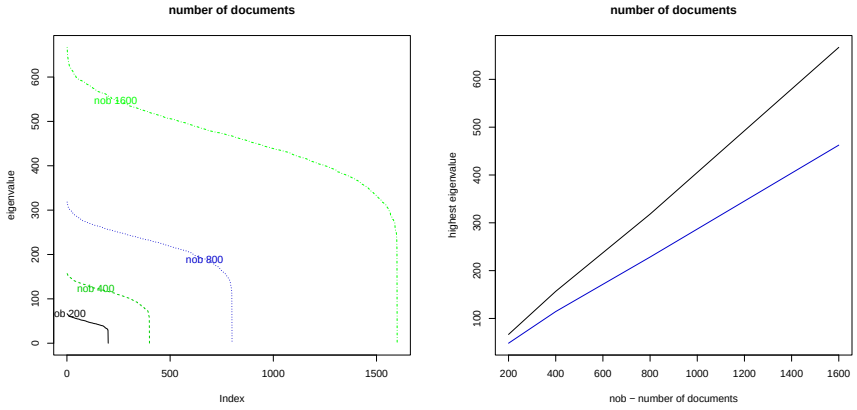


Fig. 12. Spectrogram dependence on number of objects for artificial data generated

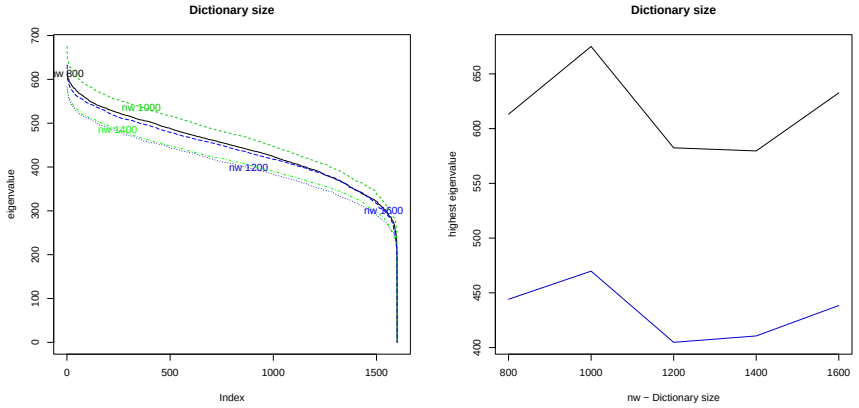


Fig. 13. Spectrogram dependence on number of words in the dictionary for artificial data generated

of the eigenvalue spectrum (and thus the structure of connections) significantly and predictably, going beyond what has been discovered so far in [5]. So it seems to be justified to use fixed length documents for theoretical studies of the GSA.

The results of the presented research can be applied as a foundation for experiments with artificial data on the usefulness of Laplacian eigenvalue spectra for Graph Spectral Analysis based clustering, incremental clustering, and classification of documents as well as research on document group similarity or collective authorship. They are also an attempt to provide additional explanations of the results of traditional spectral clustering. One way of explaining clusters is by seeking more and less similar ones. Based on the spectrogram, one could seek dissimilarity by computing the area between spectrograms. This research points at the possibility of translating this dissimilarity to distribution parameter differences.

This paper focuses on the analysis of the widely used Zipf-Mandelbrot word distribution model. However, in future research, we want to explore alternative theoretical distributions, such as e.g. the lognormal distribution, to deepen the insights gained. Furthermore, we intend to study, to what extend real data from sources like Twitter (now X) follow in fact these theoretical distributions, and with what kind of parameter values. Also, a non-parametric study is envisaged, taking the distribution(s) as-is for various data portions (related to hashtags). We intend also to go beyond the domain of short documents like those on Twitter (both for theoretical distributions and for real datasets) to see to what extent the limitation of document length impacts the spectrogram. A respective paper is under preparation [10].

It seems that there is no simple way to explain the spectrogram shape from the original document texts in the collection. In this study, we hypothesized

that the shape of the spectrogram could be attributed to writing style, while flattening it to word distribution. The results seem to confirm that in fact the spectrogram can be shaped by manipulating word distribution parameters. This hypothesis is driven by the fact that document similarity measures, underpinning the application of GSA, are based on word/term related measures (tf, tfidf etc.) not taking into account the structure of the documents. With this simplification in mind, further studies shall be conducted in order to explain the style sources, which may include: common authorship, common topic (broader than just the hashtag collection), mixture of authors for different hashtags etc. Furthermore, as the embedding models are on the rise that take into account document structural features when computing document similarity measures, also an extension of the study to this tricky area may generate new insights.

Acknowledgments. This study was funded by Polish Ministry of Science.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baayen, R.H.: Statistical models for word frequency distributions: A linguistic evaluation. Comput. Humanit. **26**(5-6), 347–363 (1992). https://doi.org/10.1007/BF00136980, https://doi.org/10.1007/BF00136980
2. Bandyapadhyay, S., Fomin, F.V., Golovach, P.A., Lochet, W., Purohit, N., Simonov, K.: How to find a good explanation for clustering? (2021). https://doi.org/10.48550/ARXIV.2112.06580, https://arxiv.org/abs/2112.06580, https://arxiv.org/abs/2112.06580
3. Barredo Arrieta, A.e.a.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82 – 115 (2020)
4. Bobek, S., Kuk, M., Szeląż, M., Nalepa, G.: Enhancing cluster analysis with explainable ai and multidimensional cluster prototypes. IEEE Access **10**, 101556–101574 (2022)
5. Borkowski, P., Kłopotek, M., Starosta, B., Wierzchoń, S., Sydow, M.: Eigenvalue based spectral classification. PLOS ONE **18**(4), e0283413 (2023), https://doi.org/10.1371/journal.pone.0283413
6. Carroll, J.: On sampling from a lognormal model of word frequency distribution. In: Kurera, H., Francis, W. (eds.) Computational Analysis of Present-Day American English, pp. 406–424. Providence: Brown University Press (1967)
7. Chaudhuri, K., Chung, F., Tsiatas, A.: Spectral clustering of graphs with general degrees in the extended planted partition model. In: Mannor, S., Srebro, N., Williamson, R.C. (eds.) Proceedings of the 25th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 23, pp. 35.1 – 35.23. PMLR, Edinburgh, Scotland (25 - 27 Jun 2012), https://proceedings.mlr.press/v23/chaudhuri12.html
8. Davidson, I., Livanos, M., Gourru, A., Walker, P., Velcin, J., Ravi, S.S.: Explainable clustering via exemplars: Complexity and efficient approximation algorithms. CoRR **2209.09670** (2022)

9. Kauffmann, J.R., Esders, M., Montavon, G., Samek, W., Müller, K.: From clustering to cluster explanations via neural networks. CoRR **abs/1906.07633** (2019), http://arxiv.org/abs/1906.07633
10. Kłopotek, M., Wierzchon, S.T., Starosta, B., Czerski, D., Borkowski, P.: Dependence of spectrogram from graph spectral clustering in text document domain; under preparation (2024)
11. Kłopotek, M.A., Starosta, B., Wierzchoń, S.T.: Eigenvalue-based incremental spectral clustering (2023)
12. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)
13. Luxburg, U.v.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007), http://dx.doi.org/10.1007/s11222-007-9033-z
14. Macgregor, P., Sun, H.: A tighter analysis of spectral clustering, and beyond. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 14717–14742. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/macgregor22a.html
15. Mandelbrot, B.: An informational theory of the statistical structure of languages. In: Jackson, W. (ed.) Communication Theory, pp. 486–502. Academic Press, Princeton (1953)
16. Orlov, J., Chitashvili, R.: On the distribution of frequency spectrum in small samples from populations with a large number of events. Bulletin of the Academy of Sciences, Georgia **108.2**, 297–300 (1982)
17. Penta, A., Pal, A.: What is this cluster about? explaining textual clusters by extracting relevant keywords. Knowledge-Based Systems **229**, 107342 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.107342
18. Sichel, H.: On a distribution law for word frequencies. Journal of the American Statistical Association **70**, 542–547 (1975)
19. Starosta, B., Kłopotek, M., Wierzchoń, S.: Hashtag similarity based on laplacian eigenvalue spectrum. In: Proc. PP-RAI'2023 - 4th Polish Conference on Artificial Intelligence , Progress in Polish Artificial Intelligence Research 4, Łódź, Poland 2023 (2023)
20. Wierzchoń, S., Kłopotek, M.: Modern Clustering Algorithms, Studies in Big Data, vol. 34. Springer Verlag (2018)
21. Xu, Y., Srinivasan, A., Xue, L.: A Selective Overview of Recent Advances in Spectral Clustering and Their Applications, pp. 247–277. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-72437-5_12
22. Zhao, Y., Liang, S., Ren, Z., Ma, J., Yilmaz, E., de Rijke, M.: Explainable user clustering in short text streams. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 155–164. SIGIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2911522, https://doi.org/10.1145/2911451.2911522
23. Zipf, G.: Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, MA (1932)