Candice: Visual Exploration of Open Government Data using Multidimensional Projections and Interactive Visualization Techniques

Tatiana F. Pereira
Dept. of Computer Science
Universidade de Brasília
Brasília, DF, Brazil
tatiana.franco@aluno.unb.br

Thiago de P. Faleiros
Dept. of Computer Science
Universidade de Brasília
Brasília, DF, Brazil
thiagodepaulo@unb.br

Vinícius R. P. Borges
Dept. of Computer Science
Universidade de Brasília
Brasília, DF, Brazil
viniciusrpb@unb.br

***Abstract*—Brazilian citizens' right to access public data is guaranteed by the Federal Constitution of 1988. In addition, government agencies have sought to ensure the availability of that data through the concept of open data. However, considering its enormous quantity, providing public data in an accessible and intuitive way becomes an issue. Lacking the ability to adequately explore the large amounts being collected, the data may lose its potential usefulness. With that in mind, we present Candice, a tool specialized in conveying the relevant patterns of open government data by means of point placement visualizations. It combines multidimensional projection methods and interactive data visualization techniques, aiming to facilitate the citizens' access to public information. The main goal of Candice is to support specialists and users when interpreting the information on open government data, promoting serendipitous discoveries. Furthermore, its visual exploration process can also benefit tasks that comprise text classification, clustering, entity named recognition and linking, among others.**

***Index Terms*—Multidimensional Projection, Interactive Visualization, Textual Data, Open Data, Information Visualization**

I. Introduction

Brazilian citizens' right to access public data is guaranteed by the Federal Constitution of 1988, in its art, 37. In addition, government agencies have sought to adapt to the concept of open data, which refers to public data represented in digital media and made available under an open license that allows free applications. For instance, the Brazilian Federal Senate has the Open Data Plan [1], through which it asserts its intention of making available all data generated by it or under its custody, except for confidential information, in order to allow its handling by the population. This document also addresses the issue of providing public data in an accessible and intuitive way, considering its enormous quantity.

Lacking the ability to adequately explore the large amounts being collected, the data may lose its potential usefulness. This problem can be solved through visual data exploration and information visualization techniques [1]. With that in mind, we present Candice, a tool specialized in conveying the relevant patterns of open data from the Brazilian government by means of point placement visualizations. It combines multidimensional projection methods and interactive data visualization techniques, aiming to facilitate the citizens' access to public information.

For demonstration purposes, we considered a collection of speeches from parliamentarians of the Brazilian Federal Senate, obtained from their official website [2]. Term Frequency-Inverse Document Frequency (TF-IDF) is employed to generate structured representations from the raw texts. Afterwards, Candice incorporates visualizations based on t-Stochastic Distributed Neighbor Embedding (t-SNE) [2] and Uniform Manifold Approximation and Projection (UMAP) [3], which perform multidimensional projection on the TF-IDF representation to a two-dimensional space.

The main goal of Candice is to support specialists and users when interpreting the information on the open data provided by government agencies, promoting serendipitous discoveries. Furthermore, its visual exploration process can also benefit tasks that comprise text classification, clustering, entity named recognition and linking, among others.

II. Literature Review

Word-count vectors used to represent textual data, such as TF-IDF, usually have thousands of dimensions. In order to generate low-dimensional layouts from them, one may use multidimensional projection. Laurens van der Maaten et al. [2] introduced t-SNE, which is currently one of the most commonly used techniques for such tasks and a variation of Stochastic Neighbour Embedding (SNE) [4].

Although SNE constructs reasonably good visualizations, it is undermined by a cost function that is difficult to optimize and by a problem the authors refer to as the "crowding problem". t-SNE aims to alleviate these issues by using a symmetrized version of the SNE cost function with simpler gradients and a Student-t distribution rather than a Gaussian to compute the similarity between two points. The results

[1]https://www12.senado.leg.br/dados-abertos/pdf/plano-2020-2021-de-dados-abertos-do-senado-federal

[2]https://www12.senado.leg.br/dados-abertos

reveal the strong performance of t-SNE compared to the other techniques. However, t-SNE suffers from limitations such as slow computation time and inability to meaningfully represent very large datasets.

Alternatively, Leland McInnes et al. [3] recently presented UMAP, a new technique for dimension reduction that aims to address the issue of uniform data distributions on manifolds through a combination of Riemannian geometry and the work of David Spivak [5] in category theoretic approaches to geometric realization of fuzzy simplicial sets. It is claimed to preserve more of the global structure than t-SNE, with a shorter run time.

One of the core assumptions of the UMAP algorithm is that there exists manifold structure in the data. Therefore, UMAP can tend to find manifold structure within the noise of a dataset. As more data is sampled, the amount of structure evident from noise will tend to decrease and the algorithm becomes more robust. Nevertheless, care must be taken with small sample sizes of noisy data or data with only large scale manifold structure.

Etienne Becht et al. [6] compares both multidimensional projection techniques mentioned before. Their work highlights the use of UMAP for improved visualization and interpretation of single-cell data and, as evidence, the authors present a qualitative comparison of UMAP with t-SNE. The conclusion was similar to that of UMAP developers [3] and, on the basis of its ease of use and results of their research, they anticipate that UMAP will be a valuable tool for the single-cell analysis community. Our hypothesis is that UMAP could have similar results when applied to textual data.

The interactive visualization techniques implemented in Candice were inspired by works in the style of ATR-Vis [7], a user-driven visual approach for the retrieval of Twitter content. The authors evaluate their approach in scenarios in which the task is to retrieve tweets related to multiple parliamentary debates. Experiments were conducted on two parliamentary data sets, and one of them refers to five mainstream debates being held in the Brazilian Federal Senate. The main goal of their proposed method is, given a set of target debates, to retrieve tweets relevant to each debate.

In order for the framework to be accessible and usable by non-technical persons, the retrieval process and its embedded strategies have been integrated into a visual interface, the ATR-Vis, which is a web-based application. They conclude that the interactive interface gives nontechnical users the tools for obtaining a reliable Twitter collection responding to their information needs before carrying out any data analyses.

The suggestion that serendipity can be facilitated through visualization was previously made by Alice Thudt et al. [8]. In order to demonstrate that, they introduce the Bohemian Bookshelf, which aims to support serendipitous discoveries in the context of digital book collections. It consists of five interlinked visualizations, each offering a different overview of the collection, and represents a first exploration into the use of information visualization to support serendipity.

A deployment at a library revealed that visitors embraced this approach of utilizing visualization to support open-ended explorations and serendipitous discoveries. Through this experiment, it was noted that library visitors made use of some visualizations in a rather targeted way, and some of them asked for possibilities to filter and specify the books displayed to certain topics of interest. This highlights the possible improvement of integrating a textual query interface into the Bohemian Bookshelf visualization.

III. Proposed Method

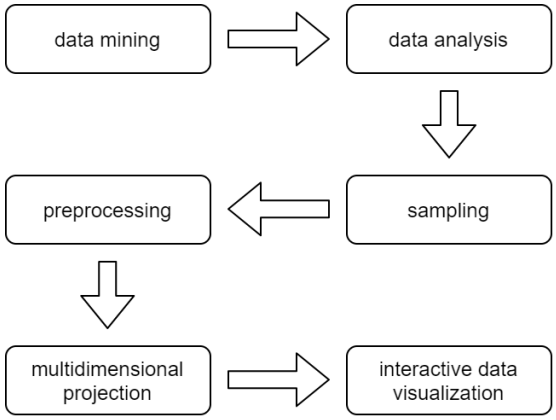

Fig. 1. Proposed method's flowchart.

A. Data Mining

For demonstration purposes, we considered a collection of speeches from parliamentarians of the Brazilian Federal Senate, obtained from their official website [3]. They were extracted by Prof. Dr. Thiago Faleiros research group and organized in a Pickle file.

B. Data Analysis

The original data set present in the pickle file consists of 74981 speeches by Brazilian parliamentarians and has metadata for both the parliamentarian who spoke and the referenced statement. There was a selection of the metadata that could be used as attribute or label for the speeches. Then, the information was reorganized and saved in a CSV file.

C. Sampling

In order to reduce the original data set to a manageable volume, we selected the speeches that took place during the term of the last four Brazilian presidents, specifically, from 2003 to 2020. Such information was also added as an attribute of the speeches. Afterwards, all instances with missing values that would later be used as labels were removed, resulting in a data set with 11222 instances.

[3]https://www12.senado.leg.br/dados-abertos

D. Preprocessing

During the preprocessing, the text was first normalized, removing accents, special characters, additional spaces and non-text characters. Then, the stop-words were removed, and finally, Term Frequency-Inverse Document Frequency (TF-IDF) [9] was employed to generate structured representations from the textual data.

E. Multidimensional Projection

Multidimensional projection, also know as dimensionality reduction, aims at representing multidimensional data in low-dimensional spaces, while preserving most of its relevant structure [10]. Up to now, Candice incorporates visualizations based on t-Stochastic Distributed Neighbor Embedding (t-SNE) [2] and Uniform Manifold Approximation and Projection (UMAP) [3], which perform multidimensional projection on the TF-IDF representation to a two-dimensional space, allowing the generation of 2D-layouts.

F. Interactive Data Visualization

The use of python libraries, such as Dash and Plotly.express, made it possible for users to interact with the 2D layouts implemented by Candice, allowing them to participate more actively in the knowledge discovery process. For instance, the points in the graphical representation are color-coded according to the label selected by the user. The users can browse through the available multidimensional projection techniques and labels, zoom in on a specific set of points and obtain more details on the attribute values of a selected point. Additionally, the user can also filter which cluster of points they wish to highlight by clicking on the cluster name present on the legend.

IV. Experiments and Results

A. Parameters Selection

Fig. 2. Layouts generated after running t-SNE with different parameters.

Figure 2 presents the layouts obtained from the experiments that occurred during the selection of the t-SNE "perplexity" and "learning_rate" parameters. The perplexity value influences the number of neighbors that is taken into account during the training. Larger data sets usually require a larger perplexity. We started with the default value, which is 30, then tested the values 50 and 100.

The learning_rate value determines how much weight the updates given by the algorithm at each step will have. If the learning rate is too high, the data may look like a 'ball' with any point approximately equidistant from its nearest neighbours. If the learning rate is too low, most points may look compressed in a dense cloud with few outliers. The experiments started with the default value of 200, and then the value 100 was tested.

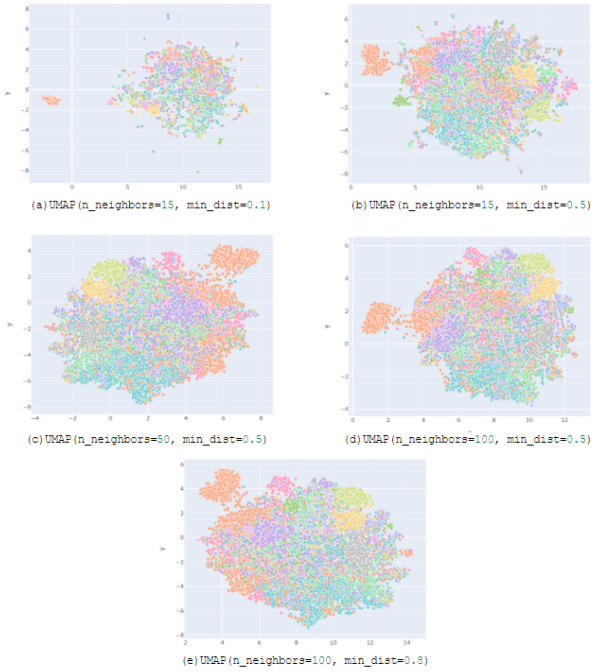

Fig. 3. Layouts generated after running UMAP with different parameters.

Figure 3 presents the layouts obtained from the experiments that occurred during the selection of the UMAP "n_neighbors" and "min_dist" parameters. The n_neighbors parameter controls how UMAP balances local versus global structure in the data. Low values force UMAP to concentrate on very local structure, while large values will push UMAP to look at larger neighborhoods of each point when estimating the manifold structure of the data. We started with the default value, which is 15, then tested the values 50 and 100.

The min_dist parameter controls how tightly UMAP is allowed to pack points together. Low values of min_dist result in clumpier embeddings. Larger values will prevent UMAP from packing points together and will focus on the preservation of the broad topological structure instead. The experiments started with the default value of 0.1, and then the values 0.5 and 0.8 were tested.

B. Qualitative Comparison of UMAP and t-SNE

We ran UMAP and t-SNE techniques on the data set of parliamentarians' speeches. Previously, they were labeled according to five labels determined from the data set's metadata: president-in-office, parliamentary party, federative unit, parliamentarians gender and parliamentarians name. Figure 4 shows some of the layouts generated from these labels. The results demonstrated that the UMAP technique is better at preserving the global structure of the data set when compared to t-SNE. However, both of them had satisfactory results when used to assist knowledge discovery tasks.

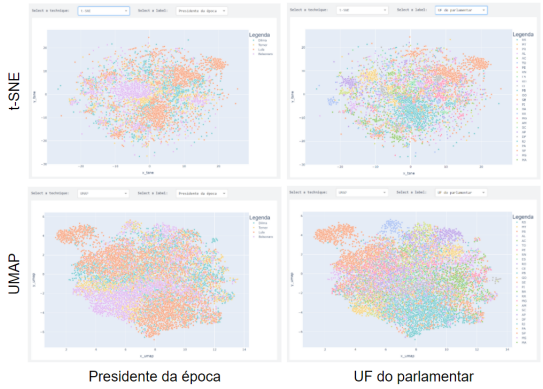

Fig. 4. Layouts generated after running t-SNE (top) and UMAP (bottom). On the left they are color-coded according to the president-in-office and on the right they are color-coded according to the parliamentarians federative unit.

Figure 5 consists of four layouts, which represent the speeches made during the term of each of the last four Brazilian presidents. It can be noted that the points in the layout of ex-president Dilma, are arranged similarly to the points in the layout of ex-president Lula. However, the layout of ex-president Temer is more dispersed, while the layout of the current president, Bolsonaro, seems to have a certain disparity when compared to the layouts of ex-president Dilma and ex-president Lula.

Fig. 5. Layouts generated after running t-SNE and color-coded according to the last four Brazilian president: Lula (upper left), Dilma (upper right), Temer (bottom left) and Bolsonaro (bottom right).

Figure 6 also consists of four layouts, one for each president. However, even though the positioning of the points is different when compared to the layouts generated after running t-SNE, here it is also possible to identify the pattern noted earlier. The layouts of ex-president Dilma and ex-president Lula are similar, the two are divergent when compared to the layout of the current president, and the layout of ex-president Temer has similar degrees of similarity when compared to the other three layouts.

Fig. 6. Layouts generated after running UMAP and color-coded according to the last four Brazilian president: Lula (upper left), Dilma (upper right), Temer (bottom left) and Bolsonaro (bottom right).

C. Interactive Visualizations and Serendipity

Serendipity is the occurrence or development of events by chance in a happy or beneficial way and has been considered to form an integral part of the creative process in the sciences [11]. As mentioned earlier, the main goal of Candice is to support specialists and users when interpreting the information on the open data provided by government agencies, promoting serendipitous discoveries.

By integrating concepts of interactive data visualization with multidimensional projection techniques, Candice is closer to achieving its goal. Since clusters of points share similar patterns, the user can make use of the relative position of points in the layouts to analyse the data set.

For instance, figure 7 presents layouts in which the points are color-coded according to the parliamentarians names. After navigating through the graphical representation, it is possible to note the constant recurrence of two parliamentarians: Paulo Paim and Álvaro Dias. Both made several pronouncements and appear to have positions that rarely coincide, which makes sense given that one is from a left-wing party and the other is from a center-right party.

It is worth mentioning that both analyzes performed previously still need the opinion of a specialist in matters related to government affairs or politics in order to obtain more concrete conclusions from them. However, it was possible to verify that integrating the t-SNE and UMAP methods with interactive

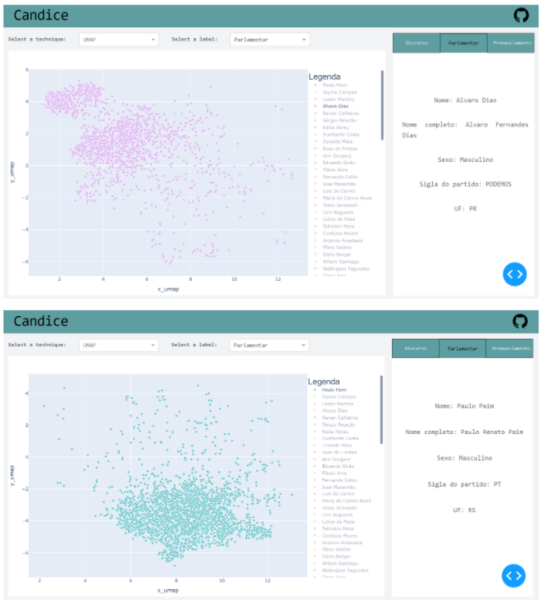

Fig. 7. Layouts generated after running UMAP representing speeches made by parliamentarians Álvaro Dias (top) and Paulo Paim (bottom).

visualization techniques facilitated the search for hypotheses related to the data set.

V. Conclusion

Candice is a simple and intuitive tool for the visual exploration of open data from the Brazilian government. It incorporates concepts of interactive text visualization, in which users participate more actively in knowledge discovery processes, therefore, promoting serendipitous discoveries. The source code is available at a Github's repository [4].

The results of the experiments demonstrated that the UMAP technique is better at preserving the global structure of the data set when compared to t-SNE. However, both had satisfactory results when used to assist knowledge discovery tasks in text mining processes. Furthermore, integrating such methods with interactive visualization techniques facilitated the analysis of hypotheses with the support of the clusters present in the 2D layouts.

There are possible avenues for extending this work. With some simple modifications and the inclusion of new information visualization techniques, Candice could also be a visualization-based tool developed for annotating textual data. Its visual exploration process can also benefit tasks that comprise text classification, clustering, entity named recognition and linking, among others. In addition, performing an evaluation of Candice with more users may help in identifying possible improvements that could be made. Finally, another interesting aspect is to compare t-SNE and UMAP with other multidimensional projection methods and evaluate them in the context of textual data.

[4]Link to the source code: https://github.com/Tatianafp/Candice

References

[1] D. A. Keim, "Visual Exploration of Large Data Sets," Communications of the ACM. 44(8), pp. 38-44, 2001.

[2] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, 9:2579–2605, 2008.

[3] L. McInnes, J. Healy and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction" arXiv:1802.03426v3, 2020.

[4] G.E. Hinton and S.T. Roweis, "Stochastic Neighbor Embedding," In Advances in Neural Information Processing Systems, volume 15, pages 833–840, Cambridge, MA, USA, 2002. The MIT Press.

[5] D. I. Spivak, "Metric realization of fuzzy simplicial sets," Self published notes, 2012.

[6] E. Becht, L. McInnes, J. Healy, CA. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," Nature Biotechnology, 2018.

[7] R. Makki, E. Carvalho, A. J. Soto, S. Brooks, M. C. F. de Oliveira, E. Milios and R. Minghim, "ATR-Vis: Visual and Interactive Information Retrieval for Parliamentary Discussions in Twitter," ACM Transactions on Knowledge Discovery from Data, Vol. 11, No. 4, Article A, 2017.

[8] A. Thudt, U. Hinrichs and S. Carpendale, "The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization," In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12). Association for Computing Machinery, New York, NY, USA, 1461–1470, 2012.

[9] Li-Ping Jing, Hou-Kuan Huang and Hong-Bo Shi, "Improved feature selection approach TFIDF in text mining," Proceedings. International Conference on Machine Learning and Cybernetics, Beijing, China, vol.2, pp. 944-946, 2002.

[10] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 241–250, 2017.

[11] A. Foster and N. Ford, "Serendipity and information seeking: An empirical study," Journal of Documentation, 59, pp. 321-340, 2003.