# VisNote: a tool for interactive visual exploration and text annotation of government gazettes

Tatiana F. Pereira, Lívia G. C Fonseca, Matheus S. V. de Oliveira, Teofilo E. Campos, Vinícius R. P. Borges
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brazil
tatiana.franco@aluno.unb.br, livia.gomes@aluno.unb.br, matheusstauffer@aluno.unb.br, teodecampos@unb.br, viniciusrpb@unb.br

***Abstract*—In this paper, we present VisNote, a visualization-based tool developed for exploring and annotating textual data from Government Gazettes. VisNote integrates interactive text visualization techniques with machine learning approaches and presents an intuitive interface to enable user's navigation. We describe VisNote's applications related to the annotation and exploratory analysis of text collections from the *Diário Oficial do Distrito Federal (DODF)* Gazettes.**

## I. Introduction

Official publications from government gazettes are important sources of information related to specific acts that takes place in government agencies, such as admissions, resignations, personnel replacement, among others. These text documents are acquired in a unstructured form and displays text blocks related to detailed descriptions of such acts. Text analysis tasks in collections of government gazettes are unfeasible when conducted by humans due to the large number of gazettes, as well as the complexity and the language of the publications. This scenario is appropriate to employ interactive visual exploration for discovering meaningful knowledge on government gazettes. For that purpose, we are developing VisNote, a tool specialized in conveying the relevant patterns of official gazettes by means of point placement visualizations. The visual exploration process incorporated in VisNote can also benefit tasks that comprise text classification, clustering, entity named recognition and linking, among others.

## II. Proposed Method

VisNote has been conceived to visualize text collections related to official publications from Government Gazettes. The goal of VisNote is to support specialists and users when interpreting the information on these documents in decision making process, such audit, fraud detection, entity linking etc. As some of these tasks require the use of labeled text collections, VisNote includes two functionalities: visual text exploration and text annotation.

In this work, we considered a collection of the *Diário Oficial do Distrito Federal* [1] for demonstration purposes. Term Frequency-Inverse Document Frequency (TF-IDF) is employed to generate structured representations from the raw texts, so that it can be used as input to the point placement visualizations based on multidimensional projections.

Up to now, VisNote incorporates visualizations based on Principal Component Analysis and t-Stochastic Distributed Neighbor Embedding (t-SNE), which performs dimensionality reduction on the TF-IDF representation to a two-dimensional space, allowing the generation of 2D-layouts. Users can interact within the graphical representation by interactive tools, such as filtering, zooming and providing details on demand.

Currently, VisNote supports tasks related to the visual exploration and annotation of textual data sets. In the visual exploration, after selecting the text collection and generating a point placement based-visualization, the user can browse through the available text collections, zoom in on a specific set of points and obtain more details on the attribute values of a selected point. In the text annotation process, the user can make use of the relative position of points in the graphical representation, since clusters of points share similar patterns. The interaction tools can be used to select and label single or multiple points at once. Moreover, the labels can be created on demand and the user can download the labeled text collection afterwards.

## III. Conclusion

VisNote is a simple and intuitive tool for the visual exploration of government gazettes, as well as for creating labeled text collections. VisNote also incorporates some concepts of interactive text visualization, in which users participate more actively in knowledge discovery processes.

Future work comprises the implementation of active learning to suggest labels automatically, the integration with text segmentation methods, and the inclusion of new information visualization techniques.

## Acknowledgment

The authors would like to thank *Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF)* for funding the research project "KnEDLe - Knowledge Extraction from Documents of Legal content" (*Convênio 07/2019*).

## References

[1] P. H. Luz de Araujo, T. E. de Campos, and M. Magalhaes Silva de Sousa, "Inferring the source official texts: can SVM beat ULMFiT?" in *International Conference on the Computational Processing of Portuguese (PROPOR)*, ser. Lecture Notes on Computer Science (LNCS). Evora, Portugal: Springer, March 2-4 2020.