Segmentation of Legal Documents: unsupervised approach with BERT embeddings

Author: Fabio Collado

Advisors: Andre Assumpção, Roberto Lotufo e Rodrigo Nogueira

December 2021

Abstract

In obedience to the principle of publicity, all administrative acts of interest to the citizens of a municipality must be published in official gazettes. These documents are not standardized, which makes access by citizens and inspection programs difficult. The first step in properly structuring this data is to identify the document sections. We propose an unsupervised algorithm to divide official journals into sections.

1 Introduction

Municipal official gazettes are essential for enabling the implementation of the principle of publicity and, at the same time, guaranteeing the transparency of the Public Administration. Through these documents, all administrative acts of a municipality can be known and inspected. Typically, journals can be divided into the following sections:

- Normative Acts: Obligations or rights to citizens resulting from the activity of the municipal executive (or legislative) power.
- Personnel Acts: Changes in the activities of municipal employees, contemplating entry into a public career, dismissal, granting temporary leave, etc.
- Public Accounts: Debt obligations of municipalities and budget requests in respect of the fiscal responsibility law.
- Public Hearings: Informs the population about public hearings regarding activities performed by the municipal government and the implementation of local public policies.
- Bids: Reports on the main instruments for contracting private goods and services in Brazil by the public sector.

- Disciplinary Processes: Reports of investigative and disciplinary proceedings on the conduct of public servants or contractors of the municipal executive power.

- Tax Processes: Reports of judgments between the municipal tax authorities and taxpayers in tax administrative proceedings that affect the municipality's collection.

- Municipal Councils: Reports on the inspection of public policies by civil society.

Finding the section where each piece of information is located can help the citizen's navigation or serve as input for other more complex tasks. However, these journals are made available in an unstructured and non-standard format. This makes it difficult to compare multiple documents and access information by optimization algorithms. Therefore, splitting these documents into sections by machine learning can improve their usefulness.

In this work, we look at multiple ways to accomplish this through unsupervised training.

2 Methodology

The focus of the project is to use machine learning to find the limits of each section of a gazette. Once these limits are known, a rules-based approach will suffice to classify each delimited group of paragraphs. Therefore, we will need to implement the following steps:

1. Download the PDF files containing the gazettes.

2. Extract them and convert to text format. Two libraries were tested: PyPDF2 and pdfminer. PyPDF2 provided the best results.

3. Preprocess the text and split the paragraphs. Preprocessing was performed by a set of regular expressions carefully tested on the available data.

4. Convert each paragraph into a vector using the first embedding of a BERT [1] encoder. Since our dataset is in portuguese, we are going to use Bertimbau [3].

5. Compute a distance matrix from the vectors. Before compute the distance, PCA was used to avoid the curse of dimensionality [2]. Multiple distance metrics were tested.

6. Create an algorithm to, using the distance matrix, find the best point to divide a group of paragraphs, in order to minimize the mean of the distances in each group.

7. Use the algorithm above to split every paragraph in a gazette. This is done by repeatedly dividing the block with the highest average distance into two. The stopping criteria is a cut quantity hyperparameter.
8. Classify a delimited group of paragraphs as a section based on rules, such as key word frequency.
9. Create a validation dataset to adjust the hyper-parameters.
10. Create a test dataset to report results.

3 Data set

The dataset is formed by texts extracted from official journals of the municipality. For validation and testing purposes, a task of extracting the purchase section was considered. For evaluation, 10 gazettes were manually annotated. Each dataset consists of a list of entities with the shape: ('filename', ['begin', 'end']), considering:

- 'filename': name of the pdf file.
- 'begin': index of the paragraph where purchase section begins.
- 'end': index of the paragraph where purchase section ends.

4 Experiments

Due to the nature of the task and since there was no previous annotated data, most experiments were qualitative. After annotating the validation dataset, to reduce the error, the following experiments were performed:

- preprocessing: improving pre-processing, cleaning data, and changing paragraph splitting rules.
- clustering algorithms: using brute force and changing the criteria to stop or choose the block to split.
- PCA: reducing dimensionality improved performance; 15 components worked better.
- distance metric: cosine worked better than Euclidean distance.
- number of blocks: how many blocks the gazette should be divided into.
- rules for purchase section: which keywords should appear and how often.

4.1 Validation

After the experimentation the best results are presented in the following table:

Validation Dataset					
File Name	Label	Label	Pred	Pred	Diff
	Begin	End	Begin	End	
523daa184adca0.pdf	462	564	460	629	67
d5dc11f00ce57b.pdf	172	243	166	245	8
7c5907de55fda5.pdf	179	217	176	255	41
cf06176a1dd49e.pdf	429	486	424	479	12
caad3e57f26751.pdf	345	436	342	455	22

4.2 Test

After choosing the hyperparameters, five extra gazettes were annotated, providing the following results:

Test Dataset					
File Name	Label	Label	Pred	Pred	Diff
	Begin	End	Begin	End	
3e971d7f662fda.pdf	209	233	209	239	6
9348175ad335ae.pdf	266	291	271	288	8
d5ef2743759abd.pdf	189	198	145	203	38
bbac1f01a469bd.pdf	271	383	266	386	8
33d9690c3a7edd.pdf	366	425	335	484	90

5 Conclusion

Tackling a real problem presents different challenges than when working with clean, ready-made data. The biggest challenges we faced were lack of annotated data, PDF extraction errors, data noise and lack of standardization of journals. This explains the difficulty the algorithm experienced in finding the boundaries of each section with greater precision.

The unsupervised algorithm's difficulty in approaching the real division might also mean that clustering by similarity is insufficient for the task. In this sense, a supervised training could find patterns that are beyond the possibility of solving the task by clustering paragraphs.

6 Future Work

In this work, we use a bag of paragraphs, disregarding the order in which they were presented. However, order is indeed important. That said, if supervised training remains out of the question due to lack of annotated data, including paragraph order might be helpful.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] R Indhumathi and S Sathiyabama. Reducing and clustering high dimensional data through principal component analysis. *International Journal of Computer Applications*, 11(8):1–4, 2010.

[3] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: Pretrained bert models for brazilian portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, pages 403–417, Cham, 2020. Springer International Publishing.