|  |
| --- |
| **Identifying Authors by Their Writings** |
|  |
| **Inês Diogo, Lara Neves, Susana Paço 20190301, 20190867, 20190821** |
|  |
|  |
|  |
|  |

Introduction

The goal of the project is given a training set develop a model capable of learning the author of a text and later, receiving a test set, be able to identify the author.

The training set is composed by a set of several labeled books written by six different authors, Almada Negreiros, Camilo Castelo Branco, Eça De Queirós, José Rodrigues Santos, José Saramago and Luísa Marques Silva. This is not a homogeneous training set, since the number of books of each author is different and the number of words in each book is also diverse.

The test set has unlabeled text excerpts with 500 or 1000 words, which are the ones the model will associate an author to each.

Method/Approach

Before training a model, a visual analysis was made where metadata was detected on the beginning of the texts, which can lead to a higher bias in the models. It was decided to remove it from the corpora and test if its presence made a significant difference in the final result. Most of the metadata was constituted by the name of the author and by references to the author’s work, like name, edition number, year and place.

In addition, it was noticed the texts were written in different time periods and that means they are written in different eras of the Portuguese language, which has changed a lot over the years.

First, it was created a standardized form to identify each file ‘.txt’ in the training set, where the names of the files were changed to the names of theirs authors plus a number, to distinguish the texts of the same author.

To clean the metadata, it was decided to remove the author’s name and the book’s name.

It was of interest to understand the importance of the presence of punctuation because one of the authors, José Saramago, is known for abusing its use and this could be a factor in identifying his texts.

Therefore, three functions for pre-processing the text were created, two to be able to distinguish the difference of keeping or removing the punctuation and a third one to completely clean the texts. The order decided for the preprocessing was to lowercase, remove the punctuation, remove the stop words (with both nltk and spacy packages because they contain different stopwords for Portuguese and so complement each other), to lemmatize and finally to remove the accents. Removing the accents was done at the end because some words without accents would not be considered as stopwords, and therefore would not be removed as well as not be lemmatized.

Word-Cloud was developed to have a visual analysis of the most frequent words written in each books.

Knowing the training set is unbalanced in favor of Camilo Castelo Branco, F1 score was chosen to be our evaluation metric, as it integrates Precision and Recall. Although, accuracy was also chosen.

To test the importance of the punctuation, previous referred, a dummy classifier was used on two texts from Saramago and two texts from José Rodrigues Santos. It was run a number of times and if the classifier is better with punctuation a positive difference should be observed, meaning that the classifier that runs on the text with punctuation should have a better accuracy score than the one without. A hypothesis test was also involved to confirm its importance.

Since the number of books is not a big, it was decided to divide the texts into chunks of 500 words, increasing the training set.

Even though, there is already a test set defined, there is a value in having an evaluation/development set to determine the model performance as it’s being trained and to be capable of adjusting its parameters. Thus k-fold cross-validation method was selected.

Then, it was of interest to create a naíve baseline to compare the model. Two different baselines were executed, a Dummy Classifier and a Bag-Of-Words. Neither involved any pre-processing of data or removal of metadata and both using the method of k-fold cross-validation. Using the Bag-Of-Words baseline it was possible to test the different levels of preprocessing. In addition, the evaluation metrics chosen, as missioned before, were accuracy and f1-score.

Finally, two models were chosen to be developed, Continuous Bag-Of-Words and Long Short-Term Memory.

Continuous Bag-Of-Words was preferred over Skip-Gram because Portuguese is a language that

fastText

Long Short-Term Memory

Results and Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Results

An example equation is shown below:

(1)

Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Acknowledgments

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

References

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

1. Appendices

Appendices are material that can be read, and include figures and tables that are not critical to the reading and understanding of the report.