SPANISH LITERATURE WORK CLASSIFICATION

Sergi Liesegang, Lucas Rodés

{sergi.liesegang, lucas.rodes}@alu-etsetb.upc.edu
BarcelonaTech - Universitat Politcnica de Catalunya (UPC)

ABSTRACT

Classification of literary works according to their author is a field of interest within the text processing area. In particular, this practice becomes relevant for categorizing anonymous works or author identity verification. Thereby, for this cases, the use of large databases of labelled works, together with the appropriate machine learning algorithms, can help determining the author of unknown samples.

Index Terms— Literature, Text classification, Natural Language Processing

1. INTRODUCTION

In this work we provide the results and review the theoretic framework of text classifiers in the context of Spanish literature. In this regard we have built own training and test sets thanks to some online resources [1]. Similar works have been carried out [2], however they only consider English language and use already prepared training sets.

Our work has focussed on preparing suitable training and test corpus in a Linux environment, which required an exhaustive usage of Bash language. Furthermore, the classification algorithmic was implemented using Python.

1.1. Organization

Section 2 gives the sources that have been used as corpus for this project. It further explains how the samples have been distributed among the train and test sets. Section 3 details the classifiers that have been employed and briefly explains some of its theoretic fundamentals. Section 4 shows some of the obtained results and their interpretation. Finally, section 5 finishes this report with some conclusions.

2. CORPUS

The main purpose of this project is to classify text to their corresponding author. To do so, large datasets are required to train properly the machine learning algorithms. Hence, selecting a suitable corpus is fundamental for an optimal performance. Particularly, we opted for the samples provided in [REF], which content is presented in 1 (including name, epoch and genre).

Table 1. Authors Set

Author	Epoch	Genre
Becquer		
Calderon		
Cervantes		
Garcia		
Lope		
Pardo		
Prez		
Quevedo		

It can be appreciated that most of the authors are important and classical figures within the Spanish Literature. Other subjects may be used for this objective.

2.1. Datasets

In our case, two main approaches have been developed to fulfil the aforementioned goal. Actually, given the nature of the database, some processing was needed to adapt it to the software implementation. We divided then each text into blocks of 10 lines, fact that resulted in a vast amount of samples as books are not used entirely but by chunks. Regarding the approaches, once the set is split, we need to separate into training and test. On the one hand, a strategy is to mix all the parts and assign a certain percentage to both sets. Note then, that this would imply samples of all works in both corpus (included indeed). On the other hand, a different methodology is to separate training and test sets in a work basis so that blocks from the same one only appear in one of them, i.e. each corpus is blind to the other one. Therefore, this last point of view may result more useful and adequate for realistic scenarios, where no samples of the target work are available for training.

2.2. Features

3. CLASSIFIERS

3.1. Naive Bayes Classifier

3.2. Support Vector Machine

3.3. Further enhancements

4. RESULTS

5. CONCLUSION AND FUTURE WORK

6. ACKNOWLEDGEMENT

We want to thank Prof. Jose A. Fonollosa for allowing us to have access to the servers of TALP Research Group.

7. REFERENCES

[1] http://www.edu.mec.gub.uy/, "Lectura solidaria," [Online accesed on 19-May-2016].

[2] Caitlin Colgrove, Sheldon Chang, and Phumchanit Yiam Watanaprakornkul, "Literary period classification," 2010.