Bag of words with Bayesian Network to detect languages

Santiago E. Bocel
Universidad Rafael Landívar
Guatemala City, Guatemala
santiagobocel10@gmail.com

Roberto Solares
Universidad Rafael Landívar
Guatemala City, Guatemala
betosolaresgar@gmail.com

Brenner Hernandez
Universidad Rafael Landívar
Guatemala City, Guatemala
velasquezbrenner@gmail.com

Pablo Muralles
Universidad Rafael Landívar
Guatemala City, Guatemala
pablomuralles28@gmail.com

ABSTRACT

Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.

There are different types of text classifiers, for example, language detection, process automation, virtual legislation, sentiment detection, etc. That is why the classification of texts is becoming an increasingly important tool, since it allows us to obtain information from data and make use of them quickly, something that is very important in the information age.

On the other hand, machine learning in its most basic form is the practice of using algorithms to analyze data, learn from it, and then make a determination or prediction about something in the world. Where there are endless techniques for learning, representation and optimization.

It is for these reasons that the combination of machine learning with text classification is a very powerful but at the same time very complex tool and a field in which there is still much to explore.

1 INTRODUCTION

Text classification is one of the applications of Machine Learning and consists of cataloging the texts based on their content, that is, performing an analysis of the words to decide what type of text is being identified.

This work is ideal for a machine as they are ideal for processing large amounts of information. However, since the machine does not initially know how to catalog a text based on any criteria, it requires a learning process in advance.

In this research, the Bag of Words (BOW) model will be used, which is a method used in language processing to classify words according to the tag.

Many language processing tasks involve a classification, in which different machine learning methods can be used such as Maximum Entropy (ME), Support Vector Machines (SVE), Naive Bayes (NB) and many more, that is why that in this work the Naive Bayes algorithm was used, more specifically

the Gaussian Naive Bayes algorithm. The Gaussian Naive Bayes algorithm is of great help as it proposes a solution to the categorization of text by assigning a label or a category to an entire text or document.

All these methods and tools can be applied in different types of classifiers such as the classification of feelings, which is the process of automating or identifying opinions in the text and labeling them as positive, negative or neutral, based on the emotions or labels that it possesses. each one, the spam classification of some text, the automatic generation of subtitles, among others. In the case of this research, we focus on knowing the language in which a text is written, language recognition.

This paper is structured as follows. The Fundamentals section describes the prior knowledge that the reader is recommended to possess in order to have a better understanding of the research. In the section The problem, the problem to be solved with this investigation is detailed as well as its requirements. The Similar Implementations section describes which experiments and projects have solved the same problem and how they have done it. In the section Our Solution, it is explained in detail how the solution was carried out, as well as why certain decisions were made and what problems were encountered when carrying out the same. The Results section describes what information could be obtained both in the training phase as well as in the experimentation and testing phase. Finally, the Conclusions section expresses the thoughts and observations of the authors after conducting the research.

2 FUNDAMENTALS

3 PROBLEM

4 SIMILAR IMPLEMENTATIONS

5 OUR SOLUTION

6 RESULTS

7 CONCLUSIONS

REFERENCES

[1] AshishSingh Bhatia and Bostjan Kaluza. 2018. *Machine Learning in Java.* Packt, Birmingham, Uk.

[2] Tony F. Chan, Gene H. Golub, and Randall J. LeVeque. 1979. Updating the formula and a pairwise algorithm for computing sample variaces. Technical report. Stanford University.

[3] Michael Copeland. [n. d.] What's the difference between artificial intelligence, machine learning and deep learning? https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial - intelligence - machine - learning - deep-learning-ai/.

[4] Praveen Dubey. [n. d.] An introduction to bag of words and how to code it in python for nlp. https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/.

[5] Srinivas Gurrala. [n. d.] Implementation of bag of words using python. https://www.excelr.com/blog/data-science/natural-language-processing/implementation-of-bag-of-words-using-python.

[6] Michael Lanham. 2020. *Practical AI on the Google Cloud Platform.* Jonathan Hassell, editor. (1st. edition). O'Reilly Media, California.

[7] Monkey Learn. [n. d.] What is text classification? https://monkeylearn.com/what-is-text-classification/.

[8] Rajat Mehta. 2017. *Big data analytics with Java.* Packt, Birmingham, Uk.

[9] @timleathart. [n. d.] How to handle a zero factor in naive bayes classifier calculation? https://datascience.stackexchange.com/questions/15526/how-to-handle-a-zero-factor-in-naive-bayes-classifier-calculation.