Universidad de Huelva
Facultad de CC. Experimentales
Grado en Química

Using Machine Learning Techniques in the Detection of Fake News

`Spring Term Project`**, Lisanna Lehes**

PROF. FRANCISCO PÉREZ-BERNAL

DATE: *8th June 2020*

DUE : *30th June 2020*

PROJECT This ambitious assignment implies the following tasks:

Task: You are provided with a dataset of news tagged as reliable or unreliable from the Kaggle website [1]. The final, and ambitious, project goal is to devise a sound fake news detector using *machine learning* techniques. The detailed project goals are:

- Create a project in *GitLab* to accomodate the files of this project and share your work with the professor.
- Get acquanted with machine learning (ML) techniques making use fo the `scikit-learn` set of tools and apply these techniques to build a fake news detector training the system with an existing dataset.
- A LaTeX file (`tex` and `pdf`) summarizing the work of the student in the project. It should include both the developed code and an outline of the results.
- The slides for a 15 minutes seminar in which the student will present the subject and the main results obtained in the project to the rest of the class.

PROBLEM DEFINITION

INTRODUCTION The recent possibility of accessing masive amounts of digital data and the increase of the computing power in processors has allowed for the application of statistical techniques and algorithms for the extraction of patterns. Techniques like neural networks were known long time ago, but the recent publication of papers where an efficient way of training such networks is presented [2] have provoked an unprecedent degree of attention on these techniques. One should distiguish between *supervised*, *unsupervised*, and *reinforcement* learning [3].

In the present project the student should get familiar with the `scikit-learn` library, an open source (BSD licensed) collection of tools written in Python that supports supervised and unsupervised learning [4] with a modern API design [5]. `Scikit-learn` provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.

The nature of the work implies to manage strings intesively and this could be achieved with the `TfidfVectorizer` tool, meant for computing the frequencies of terms or words in documents.

This tool will be applied to a Kaggle dataset which contains over 7000 news articles and each one of them has been classified either real or fake [1] that would be used to train the fake news detector.

COMPUTATION AND RESULTS

Create an account in `GitLab` and a project in this Git provider for the different files used in the present work.

FIRST PART Get acquainted with `scikit-learn` tools, in particular those related to the task at hand.

SECOND PART Download the Kaggle dataset that will be used to test your fake news detector [1].

THIRD PART Split the downloaded dataset into two subsets, *e.g.* 70% of the entries will be used to train the model and the rest to test the model's predictive power. A *PassiveAggressive Classifier* may be used to fit the model.

FOURTH PART Try to calculate the model's accuracy and present a *confusion matrix* to reckon the model's predictive power.

FIFTH PART Try the model with other news, not in the original Kaggle dataset. Depending on the state of the previous items, the implementation of a GUI could be considered.

REFERENCES

[1] Kaggle. Fake News: build a system to identify unreliable news articles. https://www.kaggle.com/c/fake-news/data. Accessed: 2020-05-14.

[2] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[3] Karen Hao. What is machine learning? in mit technology review. https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/, 2018.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[6] Brenda P. Winnewisser, Manfred Winnewisser, Ivan R. Medvedev, Markus Behnke, Frank C. De Lucia, Stephen C. Ross, and Jacek Koput. Experimental confirmation of quantum monodromy: The millimeter wave spectrum of cyanogen isothiocyanate ncncs. *Phys. Rev. Lett.*, 95:243002, Dec 2005.