Big Data Applications

‘Spam-detection Machine Learning Classifier with Pyspark ’

(Coursework Resit)

Simone Zanetti

MSc Data Science

Introduction

Spam or Junk email is a term that can be attributed to those electronic unrequested messages sent in bulk to a list of individuals. In particular, these type of electronic communications can be usually sent for a series of different reasons, and between the main ones there are commercial purposes (i.e. a company is sending commercial advertisements to communicate promotions, and so on) or phishing purposes. Specifically, the latter reason represents a technique that tries to captivate the attention of a user to direct them to write personal information at a false and fraudulent website which resembles the appearance and quality of the legitimate site.

In this context, while techniques related to phishing email are improving with the passing of the time and the evolution of technological opportunities - with the risk of more people being the victim of this illicit activity - techniques that can be able to detect a spam email and redirect it to another folder are simultaneously improving.

With all the due respect to their senders, junk emails can represent a true burden for myself and most of the other people. It is not rare to receive over 30 or 40 junk emails per day, with the consequence that these messages may create extreme disorder in your email repository, and you may run the risk to not easily identify the truly important communications. Moreover, in a world full of inputs and information, where everyone is fighting so hard to get your attention, it is fundamental to be able to decide the type and the size of information we want to be exposed to, and exclude the others. At this regards, the machine learning systems of detection of Spam email play a fundamental role to facilitate people's life and to try to help us/replace us in doing this task, by analysing electronic messages as soon as we receive them and label them as Junk or Ham[1] emails, directing them in the proper folder of our email account address.

Despite the detection of junk emails can represent probably one of the simplest Machine Learning application to explain to a non-expert, yet this task can cover a very complete set of skills and be very useful to improve the abilities of a Data Scientist. The reason for this lies in the fact that this application aims to attempt to obtain a Machine Learning algorithm out of a text file. Specifically, if we consider Big Data in terms of the 3V's[2]identified by Laney (Laney, 2001), which are Volume, Velocity and Variety, we can conclude that Spam detection is working with a type of Big data in terms of Variety. In this context, it is unavoidable to be able to work with string types and text files, and being able to find the proper way to input them into a Machine Learning model that can find hard to ingest something different from numbers. Moreover, text files may be also big in terms of Volume and require techniques of analysis based on distributed data processing systems, that can be able to overcome the limits of working with a single machine like our personal computer. Finally, this task may allow reaching a great value of accuracy and experiment with different Machine Learning methods, as well as finding several examples and cases online that can be absolutely helpful to improve the set of skills that a Data Scientist must have.

All the reasons above-mentioned are at the basis of my choice of working with this type of task in occasion of the resit coursework of the Big Data Analysis Module. The

[1] term identifying the opposite of spam emails, so in this case they are legitimate messages.

[2] Nowadays, some people talk about 4V's, 5V's or even 6V's in some cases (Patgiri and Ahmed, 2016)

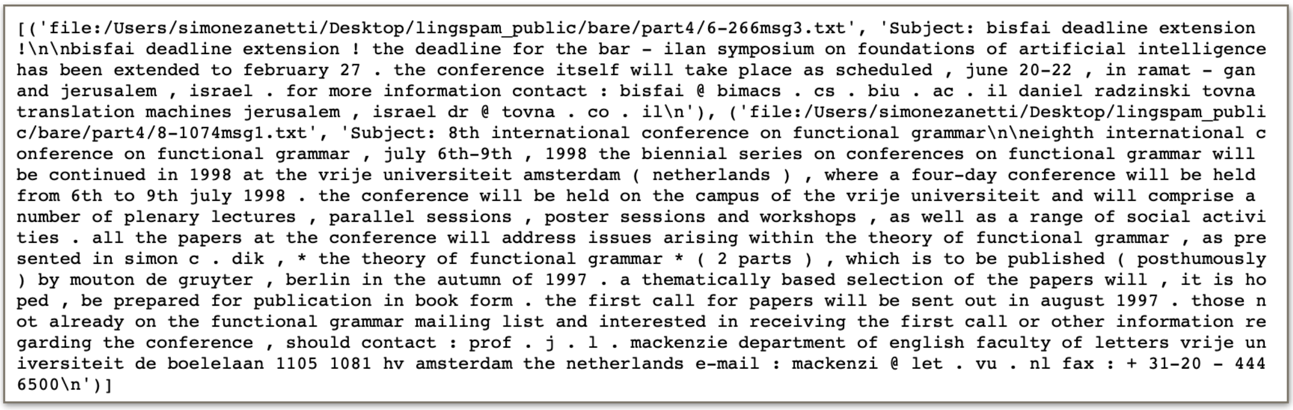
goal is the one of being able to work with a corpus of emails, that are organised in a dataset I could find and access online, and being able to import them and handle them, preparing them for the Machine Learning models I will then apply.

Dataset

The dataset used for this task is the Ling-Spam corpus, which is available online[3] and is defined in the paper Androutsopoulos et al. 2000. The dataset consists in four versions, that correspond of four subdirectories which are summarised as follows:

- **bare**: Lemmatiser disabled, stop-list disabled.
- **lemm**: Lemmatiser enabled, stop-list disabled.
- **lemm_stop**: Lemmatiser enabled, stop-list enabled.
- **stop**: Lemmatiser disabled, stop-list enabled

Each one of these 4 directories includes 10 subdirectories (part1, part2, ...) that correspond to the 10 separations of the corpus that were utilised in the 10-fold experiments. In each repetition, one part was reserved for testing and the other 9 were used for training[4].
For the scope of this analysis, the subdirectory bare will be taken into consideration. Each file/message contains the message itself and the title/name of the file. In this context, those messages whose title is spmsg* are those identified as spam. As a consequence, the other messages are those who are legitimate. This aspect is extremely important as it will allow me to have a target variable to identify on my data and train my model to predict. Each message included in the dataset when will be imported into the frame will have the following starting form (that will necessarily be modified to allow the Machine Learning algorithms to obtain a clean input):



```
[('file:/Users/simonezanetti/Desktop/lingspam_public/bare/part4/6-266msg3.txt', 'Subject: bisfai deadline extension
!\n\nbisfai deadline extension ! the deadline for the bar - ilan symposium on foundations of artificial intelligence
has been extended to february 27 . the conference itself will take place as scheduled , june 20-22 , in ramat - gan
and jerusalem , israel . for more information contact : bisfai @ bimacs . cs . biu . ac . il daniel radzinski tovna
translation machines jerusalem , israel dr @ tovna . co . il\n'), ('file:/Users/simonezanetti/Desktop/lingspam_publi
c/bare/part4/8-1074msg1.txt', 'Subject: 8th international conference on functional grammar\n\neighth international c
onference on functional grammar , july 6th-9th , 1998 the biennial series on conferences on functional grammar will
be continued in 1998 at the vrije universiteit amsterdam ( netherlands ) , where a four-day conference will be held
from 6th to 9th july 1998 . the conference will be held on the campus of the vrije universiteit and will comprise a
number of plenary lectures , parallel sessions , poster sessions and workshops , as well as a range of social activi
ties . all the papers at the conference will address issues arising within the theory of functional grammar , as pre
sented in simon c . dik , * the theory of functional grammar * ( 2 parts ) , which is to be published ( posthumously
) by mouton de gruyter , berlin in the autumn of 1997 . a thematically based selection of the papers will , it is ho
ped , be prepared for publication in book form . the first call for papers will be sent out in august 1997 . those n
ot already on the functional grammar mailing list and interested in receiving the first call or other information re
garding the conference , should contact : prof . j . l . mackenzie department of english faculty of letters vrije un
iversiteit de boelelaan 1105 1081 hv amsterdam the netherlands e-mail : mackenzi @ let . vu . nl fax : + 31-20 - 444
6500\n')]
```

[3] By obtaining a copy of this corpus I agree to acknowledge the use and origin of the corpus in any published work of yours that makes use of the corpus, and to notify the person below about this work.

[4] From the README.txt of this dataset.

Hypothesis

The goal of the analysis is the one of being able to determine the best Machine Learning algorithm to detect whether a message needs to be identified as Spam or Legitimate mail. This type of machine learning scope undoubtedly represents a Supervised Learning task. In particular, the goal will be the one of setting the target variable as either 0 (ham mail) or 1 (spam mail), creating the conditions to perform a Binary Classification task. To do so, it will be necessary at first to be able to apply a series of techniques of Natural Language Processing that will support me to clean my text data and make them feasible for coding them into a Machine Learning algorithm that can be able to understand the inputs. Moreover, the necessity of dealing with text data that can potentially have a high volume and require a set of intense operations suggests the possibility of working with the Spark framework, that will allow me to speed up the analysis, both in the phase of importation of data that in the one of production of Machine Learning algorithms. Consequently, classifiers will be tested in a distributed way using Spark. To conclude, the possibility to exploit the potentiality of Spark within the Python environment leads to the consequence that I will use Pyspark, which represents the Spark Python API (PySpark) able to expose the Spark programming model to Python. In this context, the hope is the one of obtaining good accuracy by testing different machine learning algorithms.

Planned Analysis

The planned analysis can be summarised in the following bullet points:

- **Access and Import the File:** the dataset which is organised in different folders needs to be accessed to be imported. In this phase, it is necessary to define the format in which I want the data to be stored. Rdd, Resilient Distributed Dataset, seems a good choice for storing text files which are not very big in dimension. Rdd is the primary data abstraction in Apache Spark and the core of Spark.

- **Clean the Text files**: In this phase, it is necessary to apply Natural Language Techniques to perform some operation of cleaning into the text files imported. These can be described by the necessity of operating Tokenisation, which is a technique demanded when dealing with text files organised in a string. Tokenisation is the process of separating your text into minimal meaningful units. It is a necessary step before any kind of processing. The essential tokenizer (like in NLTK) will split your text into sentences and your sentences into typographic tokens. That means isolating punctuation. Moreover, the new tokenised file will be removed of the punctuation. This will help to remove 'noise' to the future input of the machine learning models. However, it is important to note that some junk emails have a characteristic of having a lot of punctuation (ex. Great news for you !!!!!! You have won 50k $!!!!). In this context, it may be more useful to keep some types of punctuation.

- **Convert Text files into Vectors of fixed dimensions and Encode the Target into a Binary variable**: a text file in itself may not be very useful for a machine learning algorithm, which may have troubles in interpreting the strings. As a consequence, it is important to find a way for text to be converted in a format easier to understand for the algorithm. Dense vectors will be created out of each message, and from the title identifying whether a message is spam or ham I will identify a category defined by two factors (1 for spam, 0 if ham).

- **Normalise the Values and Turn the tuple Label/Dense Vector into LabeledPoints local vector**: This stage represents a part of the preprocessing phase, in which the values are normalised to be combined on a common scale. This is very important for example in the context of some models which rely on distance as a method of analysis, such as Support Vector Machines or K-Nearest Neighbours. In the end, Labeled-points local vectors will be created out of the label 0 or 1 and Dense vector. This will help in the next phase, in which models will be employed to obtain the best classification algorithm.

- **Classification models**: This phase is the proper point in which the models are applied and the accuracy will be tested on both the training data and test data. For this occasion, as well as almost everything that is done in this analysis, Pyspark will be used to set the different algorithms.

To conclude, it is important to consider that the analysis will be held in two directions:

1. From one point of view, I will attempt to develop an analysis in a highly sequential way, which means that each function and syntax will be developed and presented step by step to expose the process I made to grasp and optimise this task, same when the process may seem redundant and too verbose.

2. From the opposite point of view, I will try to conclude each phase by defining a series of functions that can allow the task to be reproducible and scalable in an efficient way on other data. This will help me to eventually modify some parameters in an easy way and perform all the analysis in few clicks, as soon as some modifications I may need to do later. This will also make everything easily readable and cleaner.

Literature

Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering". In Potamias, G., Moustakis, V. and van Someren, M. (Eds.), Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain, pp. 9-17, 2000.

Laney, D. (2001). [online] Blogs.gartner.com. Available at: hbps://blogs.gartner.com/doug- laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and- Variety.pdf [Accessed 09 August 2019].

Patgiri, R. and Ahmed, A. (2016). Big Data: The V's of the Game Changer Paradigm (PDF Download ... - MAFIADOC.COM. [online] Available at: hbps://mafiadoc.com/big-data-the- vs-of-the-game-changer-paradigm-pdf-download-_5979d23d1723dd93e84daeae.html [Accessed 09 August 2019].