Unsupervised segmentation for Email Classification

SmartClassifier is a product developed by CIB Analytics Consulting, which aims at automatically classifying incoming emails. Depending on the difficulty of the task, the models used for text classification are based on usual machine learning methods or deep learning.

Generic client service mailboxes receive service requests from clients on various topics. SmartClassifier is a valuable tool to help the desk better manage and address the requests they receive by adding the automatic classification and routing layer. However, SmartClassifier current models are built on a purely supervised approach, which assumes that targeted mailboxes already have an established classification with an history depth of classified emails long enough to perform a training on this basis.

However, a significant portion of servicing desks did not organize their mailboxes according to this criteria. In order to enlarge the scope of candidate mailboxes for SmartClassifier, we propose to develop models based on an **unsupervised** approach. In this approach, for each mailbox studied, a dedicated classification is built upon which client service requests are categorized. The aim of the internship is to **design, implement and validate** such methods. Studies will be performed on large corpus of 100 different mailboxes, each one having between 10k and 100k e-mails.

The challenges of the internship are the following:

- Being able to **identify topics of classification** that seem relevant to the desk team, so that they feel comfortable using them.
- Being able to find automatically an optimal number of classes in order to find the **best segmentation** without adding too much complexity (not too many classes).
- Being able to accommodate constraints from the users, such as imposing specific topics based on keywords.

This is a **research-oriented** internship and the intern will have laterality to explore his own ideas. However, during the internship, the intern will also have the opportunity to work in a production environment, directly contributing to the improvement of the SmartClassifier product. Therefore interns willing to familiarize themselves with best practices of **software** and **ML models deployments** in production will have the opportunity to balance the internship with tasks more closely related to the typical work of a data scientist, including direct interactions with users of our services.

References:

[1] Hinton, Geoffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press. ISBN 978-0262581684.

[2] Hinton, G (2010-08-02). "A Practical Guide to Training Restricted Boltzmann Machines".

[3] Hinton, Geoffrey (September 2009). "Deep Belief Nets" (video).

[4] Buhmann, J.; Kuhnel, H. (1992). "Unsupervised and supervised data clustering with competitive neural networks". [Proceedings 1992] IJCNN International Joint Conference on Neural Networks. Vol. 4. IEEE. pp. 796–801.

Practical Information:

Supervisor: WOUTERS Denis, Data Scientist, denis.wouters@bnpparibas.com
Location: Grands Moulins de Pantin, 9 Rue du Débarcadère, 93500 Pantin (Possible partial work from home)
Compensation: Competitive salary
How to apply: Please submit your application at paris.cib.analytics.consulting.careers@bnpparibas.com