Corporate Research Project

Master in Data Sciences and Business Analytics (DSBA)
ESSEC Business School – CentraleSupélec
Academic year: 2022/2023

Project Description

Contact information:
Fragkiskos Malliaros, Assistant Professor, CentraleSupélec
Email: fragkiskos.malliaros@centralesupelec.fr

1. Company information and contact person(s) in charge of the project and the data (including name, email and phone number)

Société Générale
17 Cours Valmy, Nanterre 92000
Supervisors: Lamiae EL FILALI, Grigory SHARKOV
Email addresses: lamiae.el-filali@socgen.com & grigory.sharkov@socgen.com

2. Title of the project

Automatic Entity Matching

3. Short description of the project, including motivation, challenges and expected results (in case needed, an extensive description can be added in the appendix)

Context

The activities of the Société Générale Group are regulated by numerous laws, both national and international. This multiple, dispersed and heterogeneous information leads us to pose the question: given the names, dates of birth/creation, country of residence/registration of two entities (natural or legal persons), how can we confirm that they correspond to the same entity? Addressing this question is one of the keys to the success of the LAB-FT project (Lutte Anti-Blanchiment, Financement du Terrorisme (English: Anti-Money Laundering, Financing of Terrorism)).

The comparison of names by classical methods, the *Levenshtein distance* for example, is not very efficient. The use of acronyms, reductions, different languages or synonyms confuses the task. Thus, "Bill GATES" and "GATES, William" can be the names of one and the same person.

We seek to test different approaches to develop a model that will be able to respond to this problem. If the progress of the work allows, the students could also work on a model for predicting the type of person (natural / legal) from the name of it.

Expected results

The students will work with a list of names of natural persons, aiming to develop a machine or deep learning model that will take two names as input and return a measure of similarity (0 – completely different names, 1 – identical names). The selected model must be selected based on a set of performance criteria. Different approaches can be tested by the students.

The choice of model will be described in a research report presented at the end of the project. The evaluation criteria will be the following:

- Performance of the best model on the test dataset (ROC, AUC, Precision, Accuracy, ..., criteria freely chosen by the students).
- The variety of approaches tested during the project.
- Quality of presentation of the results.

4. Overview of the dataset (list of variables, volume, etc.)

The available dataset for this project is divided into two parts:
1. Initial dataset: unlabeled list of names selected for this project (list of names of public persons)
2. Evaluation data: labeled list of names.

5. Which data science techniques and software tools (e.g., Python, R, SQL) are expected to be used?

The students will test different approaches to measure the similarity between two nouns:

- Text distance (alternatives to the Levenshtein distance which can be considered as a benchmark are welcome).
- Different techniques for creating representations of names, such as *WordToVec* or *CharToVec*.
- Supervised learning techniques.
- Deep Learning techniques.

6. Other useful information (e.g., how frequent can you meet the group of students, can the group work on site, etc.)

The main stages of the project will be:

1. Construction of training databases: construction of the dataset, by using external databases is expected.
2. Model(s) learning.
3. Evaluation of model(s).
4. Preparing the final report.

Appendix

(any additional information you would like to include)

1. Chen Zhao and Yeye He. 2019. Auto-EM: End-to-end Fuzzy Entity-Matching using Pre-trained Deep Models and Transfer Learning. In Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10. 1145/3308558.3313578
2. P. Christen. Development and user experiences of an open-source data cleaning, deduplication and record linkage system. ACM SIGKDD Explorations Newsletter, 11(1):39–48, 2009.
3. M. Stonebraker and I. F. Ilyas. Data integration: The current status and the way forward. IEEE Data Eng. Bull., 41(2):3–9, 2018.
4. A. Doan and A. Y. Halevy. Semantic integration research in the database community: A brief survey. AI magazine, 26(1):83, 2005.