ALTEGRAD 2019-20 Data Challenge

DaSciM team, École Polytechnique

1 Introduction

The goal of this challenge is to solve a web domain classification problem. You are given a subgraph of the French web graph where nodes correspond to domains. A directed edge between two nodes indicates that there is a hyperlink from at least one page of the source domain to at least one page of the target domain. Furthermore, your are provided with the text extracted from all the pages of each domain. A subset of these domains were manually classified into 8 categories and split between a training set and a test set. Your task is to predict the categories to which the domains of the test set belong.
The competition is hosted here on the Kaggle in-class platform. **All general questions must be asked on the Kaggle forum of the competition**.

2 Data description

2.1 Dataset

The dataset of French domains was generated from a large crawl of the French web that was performed by the DaSciM team. You are given the following files:

1. **edgelist.txt**: a subgraph of the French web graph. It has $28,002$ vertices and $319,498$ weighted, directed edges. Nodes correspond to domain ids and there is an edge between two nodes if there is a hyperlink from at least one page of the source domain to at least one page of the target domain.
2. **text directory**: for each domain, a .txt file containing the text of all the pages of the domain. The text was extracted from the HTML source code of the pages.
3. **train.csv**: $2,125$ labeled domain ids. One domain id and category per row. The list of categories is shown in Table 1.
4. **test.csv**: 560 domain ids the category of which is to be predicted. One domain id per row.
5. **graph_baseline.csv**: output of the provided graph baseline. Submissions have to follow this **exact format**.

2.2 Code

To get you started, you are provided with two simple baselines:

1. **graph_baseline.py** - based on simple graph features and the logistic regression classifier. Achieves a log loss score of 1.75 on the public leaderboard.

2. **text_baseline.py** - based on simple text features and the logistic regression classifier. Achieves a log loss score of 1.25 on the public leaderboard.

Category	# of train domains
business/finance	626
entertainment	579
tech/science	290
education/research	209
politics/government/law	200
health/medical	92
news/press	83
sports	46

Table 1: Labels of the 8 categories.

3 Evaluation

The metric for this challenge is the multi-class logarithmic loss. It is defined as the negative log-likelihood of the true class labels given a probabilistic classifier's predictions:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij})$$

where N is the number of samples (i.e., web hosts), C is the number of classes (i.e., the 8 different topics), y_{ij} is 1 if sample i belongs to class j and 0 otherwise, and p_{ij} is the predicted probability that sample i belongs to class j.

4 Rules

Read carefully and make sure you follow the rules below. Teams breaking the rules will get penalized.

- **Team size** limit is 3 students. This limit is strict, no exceptions will be made.
- **Deadline**: all submissions must be uploaded by **Sunday, March 8, at midnight Paris time**. *No extension will be granted.*
- **Submissions** must be uploaded here: `https://forms.gle/WMv4pW6YmvsbMC3z8`. Make sure that:
  - you upload **one submission per team**, as as a single archive **named after the team**
  - the archive contains **only**: (1) the team's code stored in a `/code/` subdirectory and (2) the team's report **as a .pdf file**
  - team name matches that on the Kaggle leaderboard
  - the **team name** and the **names of all team members** appear on the first page of the report
  - your submission does not contain the original data or the description of the competition
- **Reproducibility**: your code must allow the jury to reproduce your Kaggle submission.

- **Report format**: PDF reports **must follow the IJCAI format**: `https://www.ijcai.org/authors_kit`. In particular, they must have a maximum of six pages, plus at most one for references.
- **Questions**: all general questions **must be asked on the Kaggle forum of the competition**.
- **Reverse engineering**: your predictions must be generated only from the data provided. Finding the category of a domain using external sources or tools is strictly forbidden and will be considered cheating.

5 Grading scheme

Grading will be on 100 points total:

- **30 points** will be allocated based on the raw Kaggle performance, provided that the submission is reproducible. That is, using only your code and the data provided on the competition page, the jury should be able to train your final model and use it to generate the predictions you submitted for scoring.
- The content of the report will be worth **60 points**. Regardless of the performance achieved, the jury will reward here the research efforts, creativity and experimental rigor put into your work.
- Finally, **10 points** will be allocated to the organization and readability of the code and report. Best submissions will (1) clearly deliver the solution, providing detailed explanations of each step, (2) provide clear, well organized and commented code, (3) refer to relevant research papers.