Supervised and Unsupervised learning approaches for Authorship Identification- Supplementary Material

Tom Vaingart and Ayala Shaubi-Mann

1 Code repository

All code relate to this project written in Python, version 3.6.
Repository used for version control (github): https://github.com/tomvin6/author-identification

2 Set up project

- Make sure you have Python 3.6 installed and set up as interceptor.
- Clone github repository to your local environment using command "git clone https://github.com/tomvin6/author-identification.git".
- You might need to install python packages used in this project (as SKlearn, NLTK, etc.) in case you do not have them installed in your Python environment.
- As supervised output models are very big files, exceeded github allowed space in an unpaid account, we stored them in a ZIP file, on google drive.
  Download directory from this link:
  https://drive.google.com/file/d/1h33VI7Xyoh1XuiKXPq7agtZU609osFvg/view?usp=sharing
  Unzip directory to src/baseline_classifiers/ xgboost_stacked_sub_mod_dumps.

3 Running experiments

To execute experiments in supervised and unsupervised models, see our README files containing documentation on input params and used command lines for each experiment:

- Data analysis:
  https://github.com/tomvin6/author-identification/blob/master/src/data_analysis/README.md
- For supervised models: https://github.com/tomvin6/author-identification/blob/master/src/baseline_classifiers/README.md
- For unsupervised models:
  https://github.com/tomvin6/author-identification/blob/master/src/unsupervised/README.md

4 Experiments outputs

- Data analysis and general output charts are in exps/output_charts_data_analysis. Including instances distribution to author, top used words per author, top used entities per author, sentence length boxplot, word count boxplot, noun use boxplot, punctuation use boxplot.
- Supervised models files: stored in models.zip file.
  Models can be loaded and used in different python projects, or executed using one of documented command lines (as specified in README file).
- Supervised models outputs are under exps/supervised. Including confusion matrix, classified rows, learning curves and features importance.
- Unsupervised model outputs are under exps/unsupervised, including cluster labels, 2D visualization and dandogram.