Exam assignment

Language Processing 2, semester F19, teachers Jürgen Wedekind and Manex Agirrezabal

Write a paper on **Authorship attribution**

Your paper must include a practical investigation of the 2018 dataset on cross-domain authorship attribution, available on Absalon (watch out, the year is different). In your investigation you must utilize at least three machine learning-based algorithms, using the provided dataset for training and testing. You should also try different feature configurations, such as, just unigrams, unigrams and bigrams, word embeddings, and so on. You can use both K-Fold validation or a train/test split. You may use any other external package as features.

NB! Please meet *all* requirements below. **Any deviance may affect your grade.**

Formal requirements

- Follow the formal exam requirements stated in the curriculum (see link below)
- *https://hum.ku.dk/uddannelser/aktuelle_studieordninger/it_cognition/it_cognition_msc_2015*

Content requirements

- Your paper must be in English (check carefully for spelling errors and broken syntax).
- The paper should have (*i*) a descriptive title and (*ii*) a by-line with your name(s) and student ID(s).
- The paper must include the following main sections:
- **Introduction / Background** [explaining the purposes and methods of Authorship Attribution in broader terms. Check relevant literature for this.[1]]
- **Main part: Choosing ML classifiers and features** [this section must report on actual test-runs utilizing the provided dataset and applying machine learning algorithms.]
- **Quantitative evaluation** [showing how results can be evaluated automatically. Write about the validation method, K-Fold CV (stratified or not), train/test split, and so on.]
- **Qualitative evaluation** [evaluating and discussing the quantitative results wrt. validity, reliability and/or relevance in a "real-world" perspective.]
- **Conclusion** [presenting in a condensed form your results and observations.[2]]
- References

[1] Check related work section in https://pan.webis.de/clef19/pan19-web/author-identification.html

[2] You may, but need not, include your observations, e.g. strengths and weaknesses, and/or future directions into a separate section.