Project assignment SECU0057

Applied Data Science

Applied Data Science Project

- Weight for final grade: 70%
- Learning outcomes tested: (1) demonstrating knowledge of a broader range of analytical techniques used in the field of Security and Crime Science, (2) performing data science analyses on crime and/or-security related issues, (3) applying the data science pipeline on crime and/or-security related issues, (4) interpreting and effectively reporting the results of said techniques
- Deadline: 16 April 2020, 4pm.
- Page limit: See below.

> *This assessment is the capstone project of the module. It requires you to address a research problem in the full data science workflow (e.g., obtaining the data, processing the data, modelling the data, building predictive models, reporting on the findings, interpreting the outcomes). You will write a brief report on your project (a template will be provided) and you have to submit the R code needed to reproduce your findings. After passing this assessment, you will have the demonstrated the skills to solve a problem using data science techniques.*

Assessment topic

This year's topic is authorship attribution. Recently, an area of digital forensics closely connected to NLP that seeks to identify who wrote a piece of text/novel/song/etc. has received considerable attention (e.g., Why Molière most likely did write his plays, Researcher uses AI to unravel the mystery of Shakespeare's co-author, and Plecháč, 2019).

With more techniques at the researchers' disposal, this area is likely to impact on how we do digital forensic investigations (e.g. when trying to find who wrote a post). The project requires you to conduct your own authorship attribution project including (web) data collection, text processing and text mining, and approaches to identify authorship (e.g. through machine learning). *The exact nature of the project is up to you.* **The only restrictions on the topic and scope are: you must not use novels or Twitter as the source of data, and you must have at least 500 data points (e.g. texts/paragraphs/etc.) for each of at least 10 authors.**.

Further details:

- the project should be on authorship attribution (not author profiling)
- all steps in this project must be reproducible with your code supplement
- the project should have a large-scale focus (i.e. not a case study with one document)
- all three core areas of this module should be used: web data collection, text mining, machine learning

Feedback

A full project is a major step in your data science skills career. To help you in the process, you will receive feedback from us on a concept draft of your project. The deadline for the concept draft is **9 March 2020 (4pm)** via Turn-it-in on moodle. The requirements for the feedback submission are available on moodle. *The submission of the concept draft accounts for 10% of the Applied Data Science Project.*

The code supplement

Submit your R code in the form of a commented R notebook. To ensure that no code is lost and that we can review all code equally, submit your code as an anonymised view-only version on the Open Science Framework. Create a private repository, upload your code as an R notebook and create an anonymised, view-only link that you include in your report (for a guide on creating that link, see here). For details on reporting your code as an R notebook, you can consult these guides: guide 1, guide 2.

The report

For this assignment, you are asked to report your findings in the form of a short paper. Specifically, you should use the template of the ACL conference proceedings (these can be downloaded here for Latex and Word or can be imported into Overleaf here).

Additional requirements for the report:

- use the ACL style guidelines (easiest through the templates)
- the **page limit is 8 content pages** + unlimited pages for the reference list
- use the ACL referencing style (this is available in reference managers like Zotero or handled directly in Overleaf) - i.e. adhere to their font type, font size and heading guidelines.
- use the anonymous submission version which contains line numbers
- the paper must contain only your examination number in the author line
- include a footnote with an anonymised view-only link to your code on the OSF (see above).
- submit the report via moodle as a pdf file using the following file name: *SECU0057_12345.pdf* (replace 12345 with your examination number)

Deliverables

- Concept draft (9 Mar 2020)
- Project report (16 April 2020)
- Code supplement (16 April 2020)

Grading criteria

Criterion	Meaning	Weight
Originality of project	The degree to which the student demonstrates insight and is able to use an innovative approach to address the question of authorship attribution.	15%
Quality of data science techniques	The degree to which the techniques used in this project are appropriate to answer the research question and are utilised and interpreted properly.	15%
Quality of the R code	The degree to which the R code is well-documented, reproducible (with provided data if needed), and correct.	20%
Report "Introduction" section	The quality of the review of related works and the logical flow of the argument in the introduction section.	10%
Report "Method" section	The clarity of the method section that details the steps of data acquisition, preprocessing and analysis.	10%
Report "Results" section	The suitability and clarity of the results section.	10%
Report "Discussion" section	The quality of the overall interpretation as well as limitations and suggestions for future work.	15%
Concept draft submission	Whether or not the concept draft was submitted in the required format.	5%