COMP90051 Statistical Machine Learning
Project 2 Specification

Due date: 5pm Tuesday 18^{th} October 2022 (competition closes 12pm noon) Melbourne timezone **Weight:** 25%

Competition link: `https://www.kaggle.com/t/90c01d1a83664fd3b57e1889c2b2ad44`

1 Overview

Authorship attribution is the task of identifying the author of a given document. It has been studied extensively in machine learning, natural language processing, linguistics, and privacy research and is widely applied in real-world settings. Establishing the provenance of historical texts is helpful in placing important documents in their historical contexts. A whistleblower may assume that their identity is private when posting online anonymously or pseudonymously without attaching their real name. Where authorship attribution can be used to reveal their identity, as a kind of privacy attack, it is important to study the possibility of such attacks so that more robust defences can be designed. Finally plagiarism detection makes use of authorship attribution with applications from educational assessment to intellectual property law in business.

Your task:

Your task is to come up with test predictions for an authorship attribution problem given a training set and test inputs. You will participate as part of a group of students in a Kaggle competition, where you upload your test predictions. Your mark (detailed below) will be based on your test prediction performance and a short report documenting your solution.

The training data is a set of academic papers published in a time period (spanning 19 years), the given paper information includes the year it was published, the words in the title and abstract, the venue it was published in, and its authors. All the information in the discrete data has been given randomly assigned IDs, except year of publication.

The test data is a list of 800 papers, all published in the year after the training period. Your task is to predict for each of the test papers, which of a set of *100 prolific authors* were involved in writing the paper. The correct answer may be zero, one or many of these authors.

2 Dataset

train.json – contains 25.8k papers. This file is in JSON format as a list of papers, where each paper is a dictionary with keys:

- **authors:** a set of the IDs of the authors of the paper, with values in $\{0,\ldots,21245\}$;
- **year:** the year the paper was published, measured in years from the start of the training period;
- **venue:** the publication venue (name of journal/conference series), mapped to a unique integer value $\{0,\ldots,464\}$ or "" if there is no specified venue;
- **title**: the sequence of words in paper title, after light preprocessing, where each word has been mapped to an index in $\{1,\ldots,4999\}$; and
- **abstract**: the sequence of words in paper abstract, proceessed as above, using the same word-integer mapping.

Authors with IDs < 100 are the *prolific authors*, the target of this classification task. Many of the papers in **train.json** don't include any prolific authors; you will have to decide whether (and how) to use these instances in training. Note that we include some papers in the test set (described below) which have no prolific authors (no more than 25% of papers), so you will need to be able to handle this situation.

test.json – contains 800 papers, stored in JSON format with the fields **year**, **venue**, **title** and **abstract** as described above, along with one additional item:

- **identifier:** The unique identifier of the paper, used to ensure your predictions are aligned correctly in Kaggle;
- **coauthors:** The IDs of the co-authors of the paper, with values in $\{100,\ldots,21245\}$ (profilic authors with IDs < 100 are excluded). This field may be empty if there are no co-authors.

2.1 Kaggle Submission Format

You will need to submit your predictions on the 800 test papers to Kaggle at least once during the project (but ideally several times). To accomplish this, you will place your 800 predictions in a file of a certain format (described next) and upload this to Kaggle.

If your predictions are $\{1\}$ for first test paper, $\{2,3\}$ for the second test paper, and $\{\}$ for the third test paper, then your output file should be as follows in CSV format:

```
Id,Predicted
0,1
1,2 3
2,-1
```

Note that the special `-1` label used for the empty set prediction, and that the `Id` field is the *identifier* value of the corresponding paper.

The test set will be used to generate a F1-score for your performance[1] you may submit test predictions multiple times per day (if you wish). Section 6 describes rules for who may submit—in short you may only submit to Kaggle as a team not individually. During the competition the F1 on a 50% subset of the test set will be used to rank you in the **public leaderboard**. We will use the other 50% of the test set to determine your **final F1 and ranking**. The split of test set during/after the competition is used to discourage you from constructing algorithms that overfit on the leaderboard. The training data "train.json", the test set "test.json", and a sample submission file "sample.csv" will be available within the Kaggle competition website. In addition to using the competition test data, so as to prevent overfitting, we encourage you to generate your own test validation data from the training set, and test your algorithms with that validation data also.

3 Report

A report describing your approach should be submitted through the Canvas LMS **by 5pm Tuesday 18th October 2022** . It should include the following content:

1. A brief description of the problem and introduction of any notation that you adopt in the report;
2. Description of your final approach(s) to authorship attribution, the motivation and reasoning behind it, and why you think it performed well/not well in the competition; and
3. Any other alternatives you considered and why you chose your final approach over these (this may be in the form of empirical evaluation, but it must be to support your reasoning—examples like "method A, got F1 0.6 and method B, got F1 0.7, hence I use method B", with no further explanation, will be marked down).

Your description of the algorithm should be clear and concise. You should write it at a level that a postgraduate student can read and understand without difficulty. If you use any existing algorithms, *please do not rewrite the complete description, but provide a summary* that shows your understanding and references to the relevant literature. In the

[1] F1 is a balance between precision and recall, in this case computed for each author label, and then averaged (this corresponds to the `average='samples'` option in `sklearn.metrics.f1_score`). Note that the special value of `-1` is treated as an author label, and thus forms part of the computed metric.

report, we will be interested in seeing evidence of your thought processes and reasoning for choosing one algorithm over another.

Dedicate space to describing the features you used and tried, any interesting details about software setup or your experimental pipeline, and any problems you encountered and what you learned. In many cases these issues are at least as important as the learning algorithm, if not more important.

Report format rules. The report should be submitted as a PDF, and be no more than three pages, single column. The font size should be 11pt or above and margins should be at least 1.5cm on all sides, i.e., like this document. If a report is longer than three pages in length, we will only read and assess the report up to page 3 and ignore further pages. (Don't waste space on cover pages. References and appendices are included in the page limit—you don't get extra pages for these. Double-sided pages don't give you extra pages—we mean equivalent of three single-sided. *Three pages means three pages total.* Learning how to concisely communicate ideas in short reports is an incredibly important communication skill for industry, government, academia alike.)

4 Submission

The final submission will consist of three parts:

- **By 12pm noon Tuesday 18th October 2022** , submitted to the Kaggle competition website: A valid submission to the Kaggle competition. This submission must be of the expected format as described above, and produce a place somewhere on the leaderboard. Invalid submissions do not attract marks for the competition portion of grading (see Section 5).
- **By 5pm Tuesday 18th October 2022** , submitted to the Canvas LMS:
  - A written research report in PDF format (see Section 3).
  - A zip archive[2] of your source code[3] of your authorship attribution algorithm including any scripts for automation, and a README.txt describing in just a few lines what files are for. Again, do not submit data. You may include Slack/Github logs. (We are unlikely to run your code, but we may in order to verify the work is your own, or to settle rare group disputes.)
  - Your Kaggle team name—without your exact Kaggle team name, we may not be able to credit your Kaggle submission marks which account for almost half of project assessment.

The submission link will be visible in the Canvas LMS prior to deadline.

Note that after about a week into Project 2 you will need to also submit a **Group Agreement**. While not a formal legal contract, completing the agreement together is a helpful way to open up communication within your team, and align each others' expectations.

5 Assessment

The project will be marked out of 25. Note that there is a hurdle requirement on your combined continuous assessment mark for the subject, of 25/50, of which Project 2 will contribute 25 marks. **Late report submissions will incur a deduction of 2 marks per day—it is not possible to mark late competition entries.**

The assessment in this project will be broken down into two components. The following criteria will be considered when allocating marks.

Based on our experimentation with the project task, we expect that all reasonable efforts at the project will achieve a passing grade or higher.

[2] Not rar, not 7z, not lzh, etc. Zip! Substantial penalties will be applied if you don't follow this simple instruction.

[3] We would encourage you to use Python, but we will also accept submissions in Matlab, R, or otherwise. You are welcome to use standard machine learning libraries, such as sklearn, pytorch, etc, but the code submitted should be your own. Do not submit data as Canvas has limited storage space.

Kaggle Competition (**12/25**):

Your final mark for the Kaggle competition is based on your rank in that competition. Assuming N teams of enrolled students compete, there are no ties and your team comes in at R place (e.g. first place is 1, last is N) with an F1 of $F \in [0,1]$ then your mark is calculated as[4]

$$9 \times \frac{\max\{\min\{F, 0.65\} - 0.25, 0\}}{0.4} + 3 \times \frac{N-R}{N-1} \ .$$

Ties are handled so that you are not penalised by the tie: tied teams receive the rank of the highest team in the tie (as if no entries were tied). This expression can result in marks from 0 to 12. For example, if teams A, B, C, D, E came 1st, 4th, 2nd, 2nd, 5th, then the rank-based mark terms (out of 3) for the five teams would be 3, 0.75, 2.25, 2.25, 0.

This complicated-looking expression can result in marks from 0 all the way to 12. We are weighing more towards your absolute F1 than your ranking. The component out of 9 for F1 gives a score of 0/9 for F1 of 0.2 or lower; 9/9 for F1 of 0.65 or higher; and linearly scales over the interval of F1 [0.25, 0.65]. We believe that a mid-way F1 is achievable with minimal work, while results about 0.65 are good, but will require more effort. *For example, an F1 of 0.65 for a student coming last would yield 9/12; or 10.5/12 if coming mid-way in the class.*

External unregistered students may participate, but their entries will be removed before computing the final rankings and the above expression, and will not affect registered students' grades. We do not however actively invite such participation.

The rank-based term encourages healthy competition and discourages collusion. The other F1-based term rewards students who don't place in the top but none-the-less achieve good absolute results. Therefore you can achieve a high H1 grade overall irrespective of your placing in the ranking.

Note that invalid submissions will come last *and* will attract a mark of 0 for this part, so please ensure your output conforms to the specified requirements.

Report (**13/25**):

The marking rubric in Appendix A outlines the criteria that will be used to mark your report.

6 Additional Competition Rules

Teams: You are required to form a team of three students including yourself. Piazza proj2_groups folder is available to help with this—e.g. you can post if you're looking there. It isn't required that teams sit in one workshop.

Group account submissions to Kaggle: Only submissions from your group account are permitted on Kaggle. You should not attempt to submit from your individual account or create additional accounts to circumvent the daily submission limit. We have set the submission limit high so no advantage is gained from circumvention. Moreover submitting too many times is likely to risk overfitting to the leaderboard portion of the test data.

Auxiliary data prohibited: The use of any additional datasets to help with your solution is prohibited. You should only use the data that we have supplied for you. You should not search for original data sources, and we have deployed a number of obfuscation measures to prevent this.

Plagiarism policy: You are reminded that all submitted project work in this subject is to be your own individual work. Automated similarity checking software will be used to compare submissions. It is University policy that academic integrity be enforced. For more details, please see the policy at `http://academichonesty.unimelb.edu.au/policy.html`.

[4] Note that Kaggle is set to take two "scored private submissions" per team. These means that by default, your top two submissions based on public leaderboard score are chosen, then after competition close these two submissions will be scored on the entire test set, and the best total test scoring submission will make up your F1 F. If you prefer non-default submissions for your two "scored private submissions" Kaggle permits you to select others.

A Marking scheme for the Report

Critical Analysis (Maximum = 8 marks)	**Report Clarity and Structure** (Maximum = 5 marks)
8 marks Final approach is well motivated and its advantages/disadvantages clearly discussed; thorough and insightful analysis of why the final approach works/not work for provided training data; insightful discussion and analysis of other approaches and why they were not used	5 marks Very clear and accessible description of all that has been done, a postgraduate student can pick up the report and read with no difficulty.
6.4 marks Final approach is reasonably motivated and its advantages/disadvantages somewhat discussed; good analysis of why the final approach works/not work for provided training data; some discussion and analysis of other approaches and why they were not used	4 marks Clear description for the most part, with some minor deficiencies/loose ends.
4.8 marks Final approach is somewhat motivated and its advantages/disadvantages are discussed; limited analysis of why the final approach works/not work for provided training data; limited discussion and analysis of other approaches and why they were not used	3 marks Generally clear description, but there are notable gaps and/or unclear sections.
3.2 marks Final approach is marginally motivated and its advantages/disadvantages are discussed; little analysis of why the final approach works/not work for provided training data; little or no discussion and analysis of other approaches and why they were not used	2 mark The report is unclear on the whole and the reader has to work hard to discern what has been done.
1.6 mark Final approach is barely or not motivated and its advantages/disadvantages are not discussed; no analysis of why the final approach works/not work for provided training data; little or no discussion and analysis of other approaches and why they were not used	1 mark The report completely lacks structure, omits all key references and is barely understandable.