School of Computing and Information Systems
The University of Melbourne
COMP30027, Machine Learning, 2018

Project 2: this blog reads like it was written by a child!

Task:	Build a blog-authorship classifier
Due:	Stage I: Wednesday 9 May, 11am UTC+10
	Stage II: Wednesday 16 May, 11am UTC+10
Submission:	Stage I: Report (PDF) to Turnitin; test output(s) and code to LMS
	Stage II: Reviews (via Turnitin PeerMark)
Marks:	The Project will be marked out of 20, and will contribute 20% of your total mark.
Groups:	Groups of 1 or 2, with commensurate expectations for each (see below).

1 Overview

The goal of this Project is to build and critically analyse some supervised Machine Learning methods, to automatically identify the age of the author. This is a sub-problem of the larger (and more difficult) text authorship problem, which has been well-studied, but a solution remains elusive.

This aims to reinforce the largely theoretical lecture concepts surrounding learners, data representation, and evaluation, by applying them to a sophisticated problem. You will also have an opportunity to practice your general problem-solving skills, written communication skills, and creativity.

2 Deliverables

1. Stage I: the output(s) of your classifier(s), comprising predictions of labels for the test instances
2. Stage I: one or more programs, written in Python[1], which implement machine learning methods that build the model, make predictions, and evaluate where necessary
3. Stage I: an anonymous written report, of 1000-1500 words (for a group of one person) or 2000-2500 words (for a group of two people)
4. Stage II: reviews of two reports written by other students, of 200-400 words each
5. Stage II: a written reflection piece

3 Terms of Use

The data has kindly been provided to us, under the provision that any resulting publication should cite the curators of the dataset:

> Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James Pennebaker (2006) Effects of Age and Gender on Blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. Stanford, USA.

We will flatly **refuse to mark submissions** that do not cite this paper, as it is in breach of the Terms of Use.

[1] We will waive the Python requirement under certain circumstances.

Please note that the dataset is a sample of actual data posted to the World Wide Web. As such, it may contain information that is in poor taste, or that could be construed as offensive. We would ask you, as much as possible, to look beyond this to the task at hand. (For example, it is generally not necessary to read individual blog posts.)

The opinions expressed within the data are those of the (anonymised) authors, and in no way express the official views of the University of Melbourne or any of its employees; using the data in an educative capacity does not consitute endorsement of the content contained therein.

If you object to these terms, please contact us (`nj@unimelb.edu.au`) as soon as possible.

4 Data

The data files are available via the LMS, and are described in a corresponding README.

Briefly, you will be provided with a set of training documents, and a set of development documents. These have been labelled with a "class" according to the age of the author, in one of four (inclusive) ranges: 14-16, 24-26, 34-36, 44-46; the development (and test) documents also contain some documents by authors not in these ranges. Near the end of the Project period, you will be provided with a set of test documents, which will not be labelled.

Your job is to come up with one or more implemented Machine Learning system(s), which are trained using the training dataset, and evaluated using the development dataset. You will then run the trained classifier over the test dataset, and submit the corresponding predicted labels. One possible data representation has been provided, which you may use or ignore according to your needs.

5 Assessment

The Project will be marked out of 20, and is worth 20% of your overall mark for the subject. The mark breakdown will be:

Ranking of your best-performing classifier	3 marks
Report	12 marks
Reviews	3 marks
Reflection	2 marks
TOTAL	20 marks

The report will be marked according to the accompanying rubric; the details of the Stage II assessment will be announced via the LMS.

The mark for the system ranking will be calculated by equal-frequency binning the systems in the final system ranking, and assigning a score to each group based on the output which occurs in the highest-ranking bin. This procedure will be applied separately for groups of 1 member, and groups of 2 members. We may assign a bonus mark to remarkable submissions.

Since all of the documents exist on the World Wide Web, it is inconvenient but possible to "cheat" and identify some of the author ages from the test documents using non-Machine Learning methods. If there is any evidence of this, the system ranking will be ignored, and you will instead receive a mark of 0 for this component. The code will not be directly assessed, but if you do not submit it, it will be assumed that you are attempting to circumvent the Machine Learning requirement, and you will receive a mark of 0 for the system ranking component.

6 Submission

All submission will be via the LMS. Stage I submissions will open one week before the due date. Stage II submissions will be open as soon as the reports are available, immediately following the Stage I submission deadline.

7 Individual vs. Two-Person Participation

You have the option of participating as a group of one member, or in a group of two. In the case that you opt to participate individually, you will be required to enter at least 1 and up to 4 distinct systems, while groups of two will be required to enter **at least** 3 and up to 4 distinct systems, of which one is to be an ensemble system (stacking) based on the other systems. The report length requirement also differs, as detailed below:

Group size	Distinct system submissions required	Report length
1	1–4	1,000–1,500 words
2	3–4	2,000–2,500 words

If you wish to form a two-person group, **both** members need to send email to Jeremy (`nj@unimelb.edu.au`) by Friday 20 April, indicating this. Note that once you have signed up for a given group, you will not be allowed to change groups. If you do not contact Jeremy, we will assume that you will be participating as an individual, even if you were in a two-person group for Project 1.

8 Report

The report should be 1,000-1,500 words (groups of one person) or 2,000-2,500 words (groups of two people) in length and provide a basic description of:

1. the task, and a short summary of some related literature
2. what you have done, including any learners that you have used, or features that you have engineered[2]
3. evaluation of your classifier(s) over the development documents

You should also aim to have a more detailed discussion, which:

4. Contextualises the behaviour of the method(s), in terms of the theoretical properties we have identified in the lectures
5. Attempts some error analysis of the method(s)

And don't forget:

6. A bibliography, which includes Schler et al. (2006), and other related work

Note that we are more interested in seeing evidence of you having thought about the task and determined reasons for the relative performance of different methods, rather than the raw scores of the different methods you select. This is not to say that you should ignore the relative performance of different runs over the data, but rather that you should think beyond simple numbers to the reasons that underlie them.

We will provide LaTeXand RTF style files that we would prefer that you use in writing the report. Reports are to be submitted in the form of a single PDF file. If a report is submitted in any format other than PDF, we reserve the right to return the report with a mark of 0.

Your name and student ID should not appear anywhere in the report, including the metadata (filename, etc.).

[2]This should be at a conceptual level; a detailed description of the code is not appropriate for the report.

9 Reviews

During the reviewing process, you will read two submissions by other students. This is to help you contemplate some other ways of approaching the Project, and to ensure that students get some extra feedback. For each paper, you should aim to write 200-400 words total, responding to three "questions":

- Briefly summarise what the author has done
- Indicate what you think that the author has done well, and why
- Indicate what you think could have been improved, and why

10 Changes/Updates to the Project Specifications

We will use the LMS to advertise any (hopefully small-scale) changes or clarifications in the Project specifications. Any addendums made to the Project specifications via the LMS will supersede information contained in this version of the specifications.

11 Late Submissions

Late submissions will seriously create havoc with the reviewing process. You are strongly encouraged to submit by the date and time specified above. If circumstances do not permit this, then the marks will be adjusted as follows:

- Each business day (or part thereof) that the report is submitted after the due date (and time) specified above, 10% will be deducted from the marks available, up until 5 business days (1 week) has passed, after which regular submissions will no longer be accepted. A late report submission will mean that your report might not participate in the reviewing process, and so you will probably receive less feedback.
- There is no mechanism by which the reviews may be uploaded to the system after the deadline, consequently, it is a major hassle to accept late submissions. Any late submission of the reviews will incur a 50% penalty (i.e. 1.5 of the 3 marks), and will not be accepted more than a week after the reviewing deadline.
- The reflective task will largely be non-sensical to attempt after the deadline. We will reluctantly accept late submissions at a 50% penalty (1 of the 2 marks) up until a week after the task deadline.

12 Academic Honesty

While it is acceptable to discuss the Project with in general terms, excessive collaboration with students outside of your group is considered cheating. We will be vetting system submissions for originality and will invoke the University's Academic Misconduct policy (`http://academichonesty.unimelb.edu.au/policy.html`) where either inappropriate levels of collaboration or plagiarism are deemed to have taken place.

13 Important Dates

Release of training and development data	11 April 2018
Deadline for group registration	20 April 2018
Release of test data	2 May 2017
Deadline for submission of results over test data	9 May 2018 (11:00am)
Deadline for submission of written report	9 May 2018 (11:00am)
Deadline for submission of reviews	16 May 2018 (11:00am)