School of Computing and Information Systems
The University of Melbourne
COMP30027, Machine Learning, 2017

Project 2: Language Identification

Version 1.0

1 TL;DR

Task:	Build a language identification classifier
Due:	Friday 19 May 5pm
Submission:	Source code (Python) and report (PDF) to Turnitin; test output(s) to Kaggle
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).

2 Basic Task Description

The basic task is to build a language identification system focusing on Twitter posts, to gain hands-on familiarity with the various machine learning methods we have been discussing (and continue to discuss) over the course of the subject. Language identification is the task of predicting the (predominant) language a given document is written in. Complications in the language identification task as defined in this project are: (a) the development and test datasets contain languages which are not attested in the training data, which must be explicitly identified as being in an `UNKNOWN` language; and (b) while the provided training data contains Twitter data in part, it also includes three other domains, with documents of very different nature to Twitter.

You will be provided with: (1) a set of training documents; (2) a set of development documents; and (3) closer to the end of the project (see below), a set of test documents. For each training and development document, you will additionally have access to the language(s) of that document (out of a total of 20 languages, plus `UNKNOWN` for the development documents); for the test documents, this will not be provided.

Your job is to come up with implemented language identification system(s) based on the training and development datasets, to then run over the test documents to predict the language of each. Your project submission will take the form of a file containing the predicted language for each of the test documents, and the code for the systems you submit outputs for. You will also be required to write up your methodology, results and findings in the form of a written research report, to be submitted at the same time as the results of the different runs.

All documents we will look at have been manually annotated for language content, and are guaranteed to be encoded in `utf-8`. See Table 1 for an exhaustive listing of the languages in the combined document collection.

You will develop your language identification system(s) over the training and development documents, and run up to 4 different systems (see below for details) over the test documents. You may assume that the basic distribution of languages across the three document sets is comparable (for the Twitter component of the training data, at least), but expect there to be (potentially different) languages that are present in the development and test data, but not attested in the training data.

Note that we have made no attempt to sanitise the text in any way, and we accept no responsibility for the content of any of the data (text or links).

3 Assessment

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

ID	Language	ID	Language	ID	Language
ar	Arabic	he	Hebrew	nl	Dutch
bg	Bulgarian	hi	Hindi	ru	Russian
de	German	it	Italian	th	Thai
en	English	ja	Japanese	uk	Ukrainian
es	Spanish	ko	Korean	ur	Urdu
fa	Persian	mr	Marathi	zh	Chinese
fr	French	ne	Nepali	unk	"Unknown"

Table 1: The set of languages contained in the training, development and test document collections

The mark breakdown for the project will be:

Ranking of your best-performing classifier	5 marks
Creativity	5 marks
Technical soundness	5 marks
Report clarity	2.5 marks
Report structure	2.5 marks
TOTAL	20 marks

For details, see the Project 2 marking sheet.

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 individual based on the output which occurs in the highest-ranking bin.

4 Individual vs. Team Participation

You have the option of participating as an individual or in a team 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 teams of two will be required to enter **at least** 3 and up to 4 distinct systems, on of which is to be an ensemble system based on the other systems. The report length requirement also differs, as detailed below:

Team 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 team, each member needs to send email to Jeremy (`nj@unimelb.edu.au`) by 5:00pm 28 April, 2017 stating the name and unimelb username of your partner. You will then be assigned a team name for submission to Kaggle. 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, and allocate you an individual team name, which we will email to your `unimelb` email account. It is your responsibility to track down these emails and notify us if you do not receive them.

5 Data Files

All necessary data files to carry out the project are contained in the following tarball, accessible from the student machines:

```
http://people.eng.unimelb.edu.au/tbaldwin/subjects/comp30027-2017s1/projects/
project2.tgz
```

This contains the following files:

- `dev.json`: a file containing the development documents (as individual JSON records, one per line)
- `train.json`: a file containing the training documents (as individual JSON records, one per line)

6 Submission

The final submission will consist of three basic parts:

1. the output of your different classifiers (at least 1, and up to 4) over: (a) the development documents, and (b) the test documents. Ensure that your output files are formatted correctly (see below) by running the `langid-evaluate.prl` script over them
2. the Python code for your classifiers, in a single file
3. a written research report in PDF format

7 Report

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

1. the task
2. the different aspects of the task you have focused on
3. the technical details of all you have implemented (*at a conceptual rather than code level!*)
4. evaluation of your classifier(s) over the development documents, and any error evaluation

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, 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 (and, indeed, a bonus mark will be awarded for the best-performing system over the test document set), but rather that you should think beyond simple numbers to the reasons that underlie them.

We will provide LaTeXstyle files to use in writing the report. We will also publish some links to relevant papers on language identification that you may use as sources of ideas in your submission, but be sure to correctly attribute any papers through the use of references.

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 team name and member names should be clearly indicated in the report header.

8 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 the hard-copy version of the project.

9 Late Submissions

We will not accept late submissions, even on documented medical or personal grounds. If there are documentable medical or personal circumstances which have taken time away from your project work, you should contact Tim via email at the earliest possible opportunity (this generally means well before the deadline). We will assess whether special consideration is warranted, and if granted, will scale back the expectations proportionately. No requests for special consideration will be accepted after the submission deadline.

10 Academic Honesty

While it is acceptable to discuss the project with other teams in general terms, excessive collaboration 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.

11 Important Dates

Release of training and development data	24 April, 2017
Deadline for team registration	28 April, 2017
Release of test data (without annotations)	1 May, 2017
Deadline for submission of results over test data	19 May, 2017 (5:00pm)
Deadline for submission of written report	19 May, 2017 (5:00pm)