School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 1 2018

Project 2: Which emoji is missing?

Due:	Stage I: 3pm (15h00 UTC+10), Tue 22 May 2018 Stage II: 5pm (17h00 UTC+10), Fri 25 May 2018
Submission materials:	Stage I: Predictions; PDF Report Stage II: Reviews (via Turnitin PeerMark)
Assessment criteria:	Output; Critical Analysis, Report Quality; Reviews
Marks:	The Project will contribute 20% of your overall mark for the subject.

Overview

The goal of this Project is to critically analyse the effectiveness of some (supervised) Machine Learning methods on the problem of determining which emoji was used in a tweet, and to express the knowledge that you have gained in a technical report. This aims to reinforce concepts in data mining and evaluation, and to strengthen your skills in data analysis and problem solving.

Deliverables

1. The predicted labels of the test tweets.
2. An **anonymous** technical report, of 1000–1500 words, which must:
   - Give a short description of the problem and data set
   - **Briefly** summarise some relevant literature
   - Indicate which ML methods and feature representation(s) you are employing
   - Present the results, in terms of evaluation metric(s) and, ideally, illustrative examples
   - Contextualise the system's behaviour, based on the (admittedly incomplete) understanding from the subject materials
   - **Clearly** demonstrate some knowledge about the problem
3. In Stage II, reviews of two reports written by other students, 200–400 words each, which:
   - Briefly summarise what the author has done
   - Indicate what you think the author has done well, and why
   - Indicate what you think could have been improved, and why

Terms of Use

As part of the Terms of Use of Twitter, in using the data, you must agree to the following:

- You are strictly forbidden from re-distributing (sharing) the dataset with others, or re-using it for any purpose other than this project.
- You are strictly forbidden from re-producing messages from the collection in any publication, other than in the form of isolated examples.

Please note that the dataset is a sub-sample of actual data posted to Twitter, with almost no filtering whatsoever. Unfortunately, some of the information expressed in the tweets is undoubtedly in poor taste. We would ask you to please look beyond this to the task at hand, as much as possible.

The opinions expressed within the tweets in no way express the official views of the University of Melbourne or any of its employees; using the data in a teaching capacity does not constitute endorsement of the views expressed within. The University accepts no responsibility for offence caused by any content contained within this data.

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

Assessment Criteria

Output: (1 mark)
You will generate and submit the predicted labels of the test tweets.

Report: (16 marks), of which:
Critical Analysis: (70% of the report mark)
You will explain the practical behaviour of your system(s), referring to the theoretical behaviour of the Machine Learning methods where appropriate. You will support your observations with evidence, in terms of evaluation metrics, and, ideally, illustrative examples. You will derive some knowledge about the problem of identifying which emoji was used in a tweet.

Report Quality: (30% of the report mark)
You will produce a formal report, which is commensurate in style and structure with a (short) research paper. You must express your ideas clearly and concisely, and remain within the word limit (1000-1500 words). You will include a short summary of related research.

We will post a marking rubric to indicate what we will be looking for in each of these categories when marking.

Reviews: (3 marks)
You will write a review for each of two reports written by other students; you will follow the guidelines stated above.

Why emojis?

It's unlikely that this specific task — identifying which emoji has been removed from a tweet — is all that interesting, at a casual glance. Consequently, you could simply regard this as an odd dataset, and proceed with Machine Learning methods. However, we are asking you to find some knowledge, which means identifying what the problem truly is.

Automatically identifying a missing emoji *is* related to various real-world problems, for example:

- **text prediction**[1]: if we knew which emoji was most likely to be used, then we could suggest it to the user
- **sentiment analysis**: use of various emojis appears to be highly correlated with, or good predictors of the emotional feelings of the writer, and *vice versa*

If you are desperately looking for published articles for your literature summary, then searching Google Scholar (`scholar.google.com`) will turn up plenty related to sentiment analysis on Twitter.

Data

The data files are described in some detail in the README. In addition to this, here is a short description of how the data was collected:

- We sent rate-limited queries through the Twitter API[2] over the course of several days in April 2018
- These queries consisted of one of the 10 emojis[3], and results were capped at 10000 tweets per emoji; some other emojis like U+1F485 and U+1F494 were rejected, for being too rare
- The returned results were filtered, to remove tweets containing two or more of the 10 queried emojis, and tweets not containing the emoji in the visible text
- The remaining tweets were stripped of emojis (and some other special characters), shuffled, and randomly assigned to training, development and test sets

Machine Learning

If you have never done any Machine Learning before, we would strongly recommend Weka (`https://www.cs.waikato.ac.nz/ml/weka/`): it has an easy-to-use Graphical User Interface, it is coded in Java (so it runs on any platform), and it is adequate for a small dataset like this one. There is an LMS Discussion Forum post, which walks through typical Weka GUI usage.

Of course, you may use other sophisticated ML packages — most notably, `scikit-learn` is quite popular, if you have reasonable facility with Python.

You may use any of the supplied raw text, CSV, or ARFF file formats — if you wish to generate your own features from the raw text, you may do so (i.e. there is no requirement to use the given `top10` or `most100` representations).

[1] Modern smartphones actually do this for emojis!

[2] `https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets`

[3] Strangely, when querying, the emojis need to be encoded in UTF-8, but the JSON results are encoded in UTF-16.

Changes/Updates to the Project Specifications

If we require any (hopefully small-scale) changes or clarifications to the project specifications, they will be posted on the LMS. Any addendums will supersede information included in this document.

Academic Misconduct

For most people, collaboration will form a natural part of the undertaking of this project. However, it is still an individual task, and so reuse of ideas or excessive influence in algorithm choice and development will be considered cheating. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy (`http://academichonesty.unimelb.edu.au/policy.html`) where inappropriate levels of collusion or plagiarism are deemed to have taken place.

Late Submission Policy

You are strongly encouraged to submit by the time and date specified above, however, if circumstances do not permit this, then the marks will be adjusted as follows:

- Each business day (or part thereof) that this project 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.

Note that submitting the report late will mean that you will probably lose the opportunity for your report to participate in the reviewing process, which means that you will receive less feedback.