School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 1 2019

Project 1: *very tweetz. such non-canonickal. amayze*

Release: 6.15pm, Mon 01 Apr 2019
Due: Report: 9am, Mon 29 Apr 2019
Reviews/Reflection: 5pm, Fri 03 May 2019

Marks: The Project will contribute 20% of your overall mark for the subject; you will be assigned a mark out of 20, according to the criteria below.

Overview

In this Project, you will be tasked with leveraging spelling correction methods for the problem of **lexical normalisation**: finding a canonical form for each token with a document. In this case, the documents will correspond to short messages ("tweets") from the social media platform Twitter[1], which pose some unique issues. You will be tasked with gaining some knowledge about this challenging problem, and then expressing that in a written technical report.

This Project aims to reinforce concepts in approximate matching and evaluation, and to strengthen your skills in data analysis and problem solving.

Deliverables

1. One or more programs, implemented in one or more programming languages, which will:
   - Predict the best match(es) for each of the input tokens (`misspell`), with respect to a reference collection (`dict`)
   - Evaluate the predictions, with respect to the corresponding canonical form (`correct`), based on one or more evaluation metrics
2. A `README` that **briefly** details how your program(s) work(s). You may use any external resources for your program(s) that you wish: you must indicate these, and where you obtained them, in your `README`. The program(s) and `README` are required submission elements, but will not typically be directly assessed.
3. A technical report, of 1100–1350 words ($\pm 10\%$), comprising 16 of the 20 marks;
4. Reviews of two papers written by your peers, each of 250-350 words ($\pm 10\%$), comprising 3 of the 20 marks;
5. A reflection about your own paper, of 250-350 words ($\pm 10\%$), comprising 1 of the 20 marks.

[1] `http://twitter.com`

Terms of Use

Although we have manipulated the data into a suitable format, we do not own these datasets. As part of your commitment to Academic Honesty, in your report, you **must** cite its curators:

> Baldwin, Timothy, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu (2015) Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, Beijing, China, pp. 126–135.

Reports that do not contain the corresponding citation will have commensurate deductions applied to them.

Please note that the main dataset is a sub-sample of actual data posted to Twitter, with almost no filtering whatsoever. Despite the nature of the data collection (most notably as isolated terms, stripped of context), it is likely that some of the information within is distasteful or offensive. Using this data in a teaching capacity does not constitute endorsement of the corresponding viewpoints by the University of Melbourne or any of its employees.

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

Report

You will write an **anonymous** report (i.e. no name or student ID, in the header, text or file name) in **PDF format**. This report will detail your analysis, and it should focus mostly on the application of approximate matching methodologies to this problem. This should be a structured technical report, roughly in the style of the sample papers; a sample structure might be as follows:

1. A short description of the problem, data set, and what you aim to discover;
2. A **brief** summary of some published literature related to spelling correction and/or lexical normalisation;
3. An overview of your approximate matching method(s) — you can assume that the reader is familiar with the methods discussed in this subject, and instead focus on how they relate to this task and your hypothesis(es);
4. The results, in terms of the evaluation metric(s) and illustrative examples;
5. A discussion of how the results provide evidence for the normalisation of certain tokens;
6. Some conclusions about the problem of using approximate matching methods to approach lexical normalisation of short messages from social media.

You should include a bibliography and citations to relevant research papers. A good place to begin is with the citation from the Terms of Use (which in turn references many related works), but also the bibliography from the Approximate Matching lecture slides, for example:

> Zobel, Justin and Philip Dart. (1996). Phonetic String Matching: Lessons from Information Retrieval. In *Proceedings of the Eighteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*. Zürich, Switzerland. pp. 166–173.

Assessment Criteria

Report (16 marks out of 20)

Method: (30% of the report mark)
You will make one or more suitable hypotheses regarding the sorts of lexical normalisation that could feasibly be performed using approximate string matching, and design experiments using one or more spelling correction methods which could plausibly test your hypothesis(es). You will use the data to evaluate the method(s) logically and formally. You will describe your implementation in a manner that would make your work reproducible.

Critical Analysis: (40% of the report mark)
You will analyse the effectiveness of your system(s), referring to the underlying theoretical behaviour of the method(s) where appropriate. You will attempt to confirm or reject your hypotheses, using supporting evidence in terms of illustrative examples and evaluation metrics. You will derive some knowledge about the problem of lexical normalisation on short messages from social media.

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 will express your ideas clearly and concisely, and remain within the word limits. 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 out of 20)

For each paper you review, you will have 250–350 words to respond to three "questions":

- 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

Your review should focus on the content of the report, and not so much the style or structure.

1 mark will be assigned to each completed review, and 1 mark will be assigned for overall effort. Completing the reviews is expected to take about 2–3 hours in total.

Reflection (1 mark out of 20)

You will be required to "review" your own paper, according to the same three questions above. Completing the self-review is worth 1 mark, and is expected to take about 1 hour.

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 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.
- Late submissions of the reviews/reflection cannot be processed through the submission system, and therefore create a major hassle. Any late submission of these components will have a flat deduction of 50% applied.