School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 2 2018

Project 1: Waht kinda typoz do poeple mak?

Release: 7.15pm (19h15 UTC+10), Tue 07 Aug 2018
Due: Report: 1pm (13h00 UTC+10), Tue 04 Sep 2018
Reviews/Reflection: 1pm (13h00 UTC+10), Tue 11 Sep 2018

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 over some given data, to identify the causes of typographical errors made by typical authors, and to express the ensuing knowledge in a short technical/research report. This 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:
   - Determine the best match(es) for a token, with respect to a reference collection (dictionary)
   - Process the data input file(s), to determine the best match(es) for each (misspelled) token
   - Evaluate the matches, with respect to the truly intended (correct) words, using 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 (±10%), comprising 15 of the 20 marks;
4. Reviews of three papers written by your peers, each of 250-350 words (±10%), comprising 4 of the 20 marks;
5. A reflection about your own paper, of 250-350 words (±10%), comprising 1 of the 20 marks.

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, you **must** cite its curators; in the case of Wikipedia:

> Wikipedia contributors (n.d.) Wikipedia:Lists of common misspellings. In *Wikipedia: The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Wikipedia:Lists_of_common_misspellings&oldid=813410985

In the case of Birkbeck[1], if you use it:

> Mitton, Roger (1980) Birkbeck spelling error corpus. In *University of Oxford Text Archive*, http://ota.ox.ac.uk/headers/0643.xml

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

Please note that the main dataset is a sub-sample of actual data posted to Wikipedia, with almost no filtering whatsoever. Due to the nature of the data collection (most notably as isolated terms, stripped of context), it is unlikely — but not impossible — that 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 and data set;
2. A **brief** summary of some published literature related to spelling correction and/or typographical errors;
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 are applied to this task;
4. The results, in terms of the evaluation metric(s) and illustrative examples;
5. A discussion of how the results provide evidence supporting the presence/absence of theoretical types of typographical errors;
6. Some conclusions about the problem of using approximate matching methods to identify typographical errors.

You should include a bibliography and citations to relevant research papers. A good place to begin is probably with the bibliography from the Approximate Matching lecture slides, in particular:

> 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.

[1] Distributed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License: https://creativecommons.org/licenses/by-nc-sa/3.0/

Assessment Criteria

Report (15 marks out of 20)

Method: (30% of the report mark)
You will make one or more suitable hypotheses regarding the cause(s) of typographical errors, and design experiments using one or more spelling correction methods which could plausibly test your hypotheses. 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 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 identifying the causes of typographical errors.

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 (4 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

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 3–4 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 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.