School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 2 2017

Project 1: Lexical Normalisation of Twitter Data

Due:	1pm (13h00 UTC+10), Wed 06 Sep 2017
Submission materials:	Source code, `README`; PDF Report (Submission mechanisms described on the LMS)
Assessment criteria:	Method, Critical Analysis, Report Quality
Marks:	The Project will contribute 20% of your overall mark for the subject.

Overview

The goal of this Project is to assess the performance of some spelling correction methods on the problem of tweet normalisation, and to express the knowledge that you have gained in a technical 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 must:
   - 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 for each token
   - Evaluate the matches, with respect to the truly intended 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 1000–1600 words, which must:
   - Give a short description of the problem and data set
   - **Briefly** summarise some relevant literature
   - Briefly explain the approximate matching technique(s), and how it is (they are) used
   - Present the results, in terms of the evaluation metric(s) and illustrative examples
   - Contextualise the system's behaviour, based on the (admittedly incomplete) understanding from the subject materials
   - **Clearly** demonstrate some knowledge about the problem

Terms of Use

By using this data, you are becoming part of the research community — consequently, as part of your commitment to Academic Honesty, you **must** cite the curators of the dataset in your report, as the following publication:

> Bo Han and Timothy Baldwin (2011) Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA. pp. 368–378.

Reports that do not cite this work constitute plagiarism, and will be correspondingly assigned a mark of 0.

Please note that the dataset is a sub-sample of actual data posted to Twitter, with almost no filtering whatsoever. Unfortunately, the Internet is a place where freedom of speech is both empowering and harmful: consequently, 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. (For example, it is generally not necessary to actually read the tweets themselves.)

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

Method: (20% of the marks available)
You will attempt a representative sample of approximate matching techniques, which is adequate for deriving some knowledge about the problem of tweet normalisation. You will evaluate your method(s) formally.

Critical Analysis: (50% of the marks available)
You will explain the practical behaviour of your systems, referring to the theoretical behaviour where appropriate. You will support your observations with evidence, in terms of illustrative examples and evaluation metrics. You will derive some knowledge about the problem of tweet normalisation.

Report Quality: (30% of the marks available)
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 (1000-1600 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.

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.