Department of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 2 2019

Project 1: Word Blending in Twitter

Released:	Friday 16 Aug
Due:	**Research Paper:** Friday 13 Sep – 5PM **Reviews:** Wednesday 18 Sep – 5PM
Marks:	The project will be marked out of 20 (according to the given criteria), and will contribute 20% of your total mark.

Overview

The goal of this project is to develop and critically assess methods for detecting word blends among frequent terms in Twitter data, and to express the knowledge that you have gained about this task in a short research paper. Twitter users use language innovatively, and coining new terms by blending two existing words is a common phenomenon, known as lexical blending. Consider the following examples:

Component 1	Component 2	Blend word
Britain	exit	Brexit
spoon	fork	spork
breakfast	lunch	brunch

You will detect occurrences of blend words among a pre-processed list of tokens from a Twitter data set, using a reference set of English words from a dictionary, and using methods for approximate string matching as encountered in the lectures. We will also provide you with a set of tweets the token list was extracted from, which you may (but are not expected to) use. You will evaluate the output of your algorithm(s) against a list of true word blends. The project aims to reinforce concepts in approximate matching and evaluation, and to strengthen your skills in data analysis and problem solving.

The goal of this assignment is **not** to develop a system which achieves near-perfect precision (in fact, this is impossible – we are developing knowledge technologies after all!).

Deliverables

1. One or more programs, implemented in the programming language(s) of your choice, which must:

    - Process the data input file(s), to identify word blend candidates
    - Identify word blend candidates among a set of tokens, with the help of a reference

collection of tokens (dictionary)

- Evaluate the matches, with respect to the list of true word blends, 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. An **anonymous** short research paper of 1100–1350 words (±10%), as a single file in PDF format, which should include:

   - A short description of the problem and data set
   - A brief summary of some relevant literature
   - A brief explanation of the approximate matching techniques used
   - Presentation of your results in terms of the evaluation metrics discussed and illustrative examples
   - A discussion on the knowledge you have gained about the problem at hand, and about the (un)suitability of the approaches you have adopted

4. Reviews of two research papers written by your peers, each of 250-350 words (±10%), comprising 4 out of the 20 marks and a critical self-reflection on your own work.

Terms of Use

As part of the terms of use of Twitter, in using the data you agree to the following:

- The Twitter dataset is based on the data set presented in

  > Jacob Eisenstein, Brendan O'Connor, Noah A. Smith,and Eric P. Xing. 2010. A latent variable model forgeographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1277–1287

  You need to cite this paper in your research paper.

- The list of blend words was compiled using resources presented in the following publications

  > Deri, A. and Knight, K. (2015) How to Make a Frenemy: Multitape FSTs for Portmanteau Generation. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 206–210
  >
  > Das, K. and Ghosh, S. (2017) Neuramanteau: A Neural Network Ensemble Model for Lexical Blends. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 576–583
  >
  > Cook, P. and Stevenson, S. (2010) Automatically Identifying the Source Words of Lexical Blends in English. In *Computational Linguistics, Volume 36(1)*

  You need to cite these papers in your research paper.

- You are strictly forbidden from reproducing documents in the document collection in any publication, other than in the form of isolated examples.

Additionally note that the document collection is a sub-sample of actual data posted to Twitter, without any filtering whatsoever. As such, the opinions expressed within the documents in no way express the official views of The University of Melbourne or any of its employees, and my using them does not constitute endorsement of the views expressed within. We recognize that some of you may find certain of the documents in bad taste and possibly insulting, but please look beyond this to the task at hand. The University of Melbourne accepts no responsibility for offence caused any content contained in the documents.

Assessment Criteria

(1) Short research paper: (15 marks out of 20)

- **Method: (30% of the paper mark)**

  You will make one or more suitable hypotheses regarding the coinage of blend words, and design experiments using one or more approximate matching 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 paper mark)**

  You will analyze the effectiveness of your system(s), referring to the underlying theoretical behavior 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 paper mark)**

  You will produce a 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.

NOTE: A marking rubric is available on LMS to indicate what we will be looking for in each of these categories when marking.

(2) Reviews and self-reflection (5 marks out of 20)

You will have 250–350 words to respond to three “questions” for two research papers of your peers (2 marks each) and for your own paper (1 mark):

- Briefly summarize 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

Completing the reviews is expected to take about 3–4 hours in total.

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.