University of Windsor
Faculty of Science
School of Computer Science

COMP-8730
Natural Language Processing and Understanding
Winter 2024

#	Title	Due Date	Grade Release Date
Assign. 1	Spell Correction Using MED	Jan. 28, AoE	Feb. 05

The objectives of the assignments are to practice on topics covered in the lectures as well as improve the student's critical thinking and problem-solving skills in ad hoc topics that are closely related but not covered in the lectures. Lecture assignments also help students with research skills, including the ability to access, retrieve, and evaluate information (information literacy).

Lecture Assignment

Given a dictionary D and a spelling error corpus C for the English language, calculate the average success at k (s@k) for the minimum edit distance (MED) algorithm for all misspelled tokens in C.

a) Use WordNet[1] as the dictionary *D*. A python interface to WordNet is PyDictionary[2] which is available opensource[3].
b) Use Birkbeck[4] spelling error corpus. This corpus includes the most common misspelled tokens and the correct spell in pairs. For instance ('*desing*', '*design'*).
c) Success at *k* (s@*k*) measures whether the correct spell of the token happens to be in the top-*k* (most similar, least distant) list of tokens that are retrieved by the MED from the dictionary *D*. For instance, given '*desing'* from Birkbeck corpus, the top-5 most similar (least distant) tokens to '*desing'* based on MED from WordNet are ['*desi', 'design', 'designer', 'designate', 'despair'*]. Then, s@1 is 0 since the correct spell from Birkbeck is '*design'* which is not happening at the first item. However, s@*k* for $k \geq 2$ is 1. Report the average s@*k* for $k=\{1, 5, 10\}$ using PyTrec_Eval_Terrier[5].
d) *Comparing each misspelled word with all words in a dictionary takes time. You have to come up with workarounds such as parallel runs.*

Submission Guidelines

- Submission must be written as a report in English, in the current ACM two-column conference format in LaTeX. Overleaf templates[6] are available from the ACM Website[7] (use the *sigconf* proceedings template).
- The report must be 1 page in length, no more no less, including figures, tables, references, and *authored by the students from same team of research project*.
- The code should be available in a *public* online repo (preferably Github) and the link should be mentioned as a footnote to the report's title. See the example below. The results reported in the report must be reproducible (multiple runs with the same setting should result in the same results.)
- *I dare you, I double dare you![8] on plagiarism, code copy, or any issues with academic integrity.*
- Submission must be in one single zip file named Assign01_firstname1_{firstname2}.zip, including:
  1. LaTeX files
  2. the pdf file

A sample submission has been attached to this manual in Brightspace, also available online[9].

[1] https://en.wikipedia.org/wiki/WordNet
[2] https://pypi.org/project/PyDictionary/
[3] https://github.com/geekpradd/PyDictionary/tree/master/PyDictionary
[4] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643
[5] https://pypi.org/project/pytrec-eval-terrier/
[6] https://www.overleaf.com/gallery/tagged/acm-official
[7] https://www.acm.org/publications/proceedings-template
[8] https://www.youtube.com/watch?v=Qc8-G7gbkms
[9] https://www.overleaf.com/read/xdmhfgmfjfwk