University of Windsor
Faculty of Science
School of Computer Science

COMP-8730
Natural Language Processing and Understanding
Winter 2022

#	Title	Due Date	Grade Release Date
Assignment 02	Spell Correction using LM	March 07, 2022, AoE	March 14, 2022

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).

Assignment

Given a set of language models L that are trained on a corpus C and a spelling error corpus E for the English language, calculate the average success at k (s@k) using each language model $l \in L$ for E.

a) Train L={n-Gram language models} on C={news genre of Brown's corpus *or any reasonable corpus*} for n={1,2,3,5,10}[1].
b) Use `APPLING1DAT.643` in Birkbec[2] spelling error corpus that includes the most common misspelled tokens, the correct spells, and the sentences that the misspelled token happened, in triples. For instance (`steped stepped when I first *`). *You can use any other well-know corpus*.
c) Success at k ($s@k$) measures whether the correct spell of the token happens to be in the top-k (most probable) list of tokens that are retrieved by a language model. For instance, given '`when I first`', the top-5 most probable tokens based on unigram language model would be ['`went`', '`saw`', '`started`', '`stepped`', '`looked`']. Then, $s@1$ is 0 since the correct spell from Birkbeck is '`stepped`' which is not happening at the first item. However, $s@k$ for $k \geq 4$ is 1. Report the average $s@k$ for k={1, 5, 10} using PyTrec_Eval[3] for each $l \in L$.
d) Hint: unplugging the MED (Assign 1) and plug the trained LM. There should be no change in evaluation.

Submission Guidelines

- Submission must be written as a report in English, in the current ACM two-column conference format in LaTeX. Overleaf templates[4] are available from the ACM Website[5] (use the sigconf proceedings template).
- The report must be 1 page in length, no more no less, including figures, tables, references, and single-authored by the student.
- The code should be available in an 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.)
- Submission must be in one single zip file named `COMP8730_Assign02_UWindId.zip`, including:
  1. the LaTeX files
  2. the pdf file

A sample submission has been attached to this manual in Blackboard, also available online[6].

[1] https://www.kaggle.com/alvations/n-gram-language-model-with-nltk
[2] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643
[3] https://github.com/cvangysel/pytrec_eval
[4] https://www.overleaf.com/gallery/tagged/acm-official
[5] https://www.acm.org/publications/proceedings-template
[6] https://www.overleaf.com/read/dbrhbpqghfjc