# Natural Language Processing (NLP)

# IN-COURSE ASSESSMENTS

# Overview

The string distance is a metric of how alike two strings are to each other. The Bayesian method can be viewed as a way of applying such an algorithm to the spelling error correction problem; the candidate word which is ‘closest’ to the error in the sense of having the highest probability given the error, is chosen.

Your project is to build a probabilistic model for detecting of spelling errors. It should correct spelling errors which result in **non-words** and **real words**, for example correcting *graffe* to *giraffe*. Search and find a corpus in any sciences field, such as medical science, business, etc., consisting of at least 100,000 words, and build the system based on this corpus.

# Academic Integrity

Copying or paraphrasing someone's work (code included) or permitting your own work to be copied or paraphrased, even if only in part, is not allowed, and will result in disciplinary action. Your grade should reflect your own work. Basically, 'plagiarism' means representing someone else's work as if it is your own. This is a very serious academic offence for all students within the University regulations and is particularly reprehensible for a researcher. Please do not even consider it. Remember that accidental plagiarism (or the appearance of it) may be avoided by referencing your work properly. This gains you credit, not loses it! The simple rule is that you must not represent the ideas of other people (whether they are published works or the work of other students) as your own.

The golden rule on plagiarism is **DO NOT DO IT!**

# GROUP ASSIGNMENT - 60%

1. **Introduction of Candidate Techniques [10 Marks]**

Conduct appropriate review of existing systems and techniques to understand the problem domain and identify suitable candidate techniques. Most candidate techniques would have been coded and explored, with good techniques incorporated into packages and libraries for generic use. Identify available libraries / packages in Python (and other relevant sources) and coding techniques that would aid in the implementation of the required system. Ensure that all sources used are cited correctly.

1. **Edit Distance [10 Marks]**

The most common method used in determining “similar” words in a spellcheck system is **Edit Distance**. There are variations of how this can be implemented in order to improve accuracy. Describe the variations of Edit Distance that may be useful candidate techniques for this project.

# Formulation / Design [15 Marks]

You will need to design a system to implement advanced NLP techniques to solve the given problem. Design a system to meet the following requirements:

1. Your system must have an appropriate graphical user interface (GUI). There must be small editor, with a size of 500 words in the GUI to allow user to write a short piece of text. The Figure below depicts a sample of GUI needed for your application: You may come up with a more suitable GUI.

Check Spelling

Text Editor (500 characters)

and

base

board

boss

cast

board

Dictionary:

Search:

1. Your application must be able to find the spelling errors and suggest a few words to the user to modify the text.
2. The spelling errors that need to be addressed by your system are:
3. **Non-words** (wrong spelling, where the word does not exist)
4. **Real-words** (wrong spelling due to wrong context, but the misspelt word does exist)
5. The techniques used for the detection of the spelling errors must include *bigram*, *minimum edit distance*, and any other suitable popular techniques used in NLP.
6. Provide the following functionality in your application:

* Ability to show a sorted list of all words in the corpus with the facility of exploring the list and search for a specific word
* Ability to highlight the misspelled words, and click on the word to suggest the correct words *(with their minimum edit distance from the wrong word)*

# Implementation [15 Marks]

Detail the implementation of your system by showing all interesting implementation efforts and optimisations undertaken. Provide a discussion of the findings.

# Results [10 Marks]

Provide explanation supported by screenshots to show how the system was able to perform the requirements stated above. Show proper testing, describing the strengths and any limitations.

**Submission Criteria**

1. This assignment is to be completed in groups. Details of the groups to be provided during class.
2. Deliverable: A write-up of approximately **3000** words, split into the sections mentioned above.
3. Submission deadline: **[date]**

**DEMONSTRATION (INDIVIDUAL COMPONENT) – 40%**

1. **Demo [20 Marks]**

You will need to demonstrate the developed system, discussing the strengths and limitations.

1. **Advance Concepts [20 Marks]**

Using your knowledge with regards to **Parts of Speech (POS), Information Retrieval (IR)** and **Semantics**,providea clear understanding of how the system could be improved further in order to obtain better accuracy.

**Submission Criteria**

1. This is an **individual component** of the group work and to be completed individually. The assessment would be in terms of the sections above.
2. Deliverable: **2-5 minutes** screen cam recording containing the demonstration of the system and brief talking points of item 2 above. For the latter, displaying the talking points on the screen with voiceover explanation will suffice.
3. Submission deadline: **[date]**