## Speech and Image Processing - Lab 2

Date: 16 September 2019

**Task description:**

**Part (a)** (8 Marks)

Write a python program that takes two single-line lowercase English files reference.txt and hypothesis.txt, and outputs the file result.txt containing Levenshtein distance of these two files as below. The distance should be word level and not character level.

\*\*\*\*\*\*\*\*\*\*reference.txt\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

this is some text and we would like to see if it has been identified correctly by speech recognition system

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*hypothesis.txt\*\*\*\*\*\*\*\*\*\*\*\*\*

this is a text and we would like to check what has been identified by the speech recognition

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*result.txt\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Levenshtein distance is 7

Insertions 1

Deletions 3

Substitutions 3

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Part (b)** (2 marks)

Modify the above program so that it ignores 10 common words such that

* Insertions and deletions involving these common words are ignored
* Substitutions are ignored when both initial and final word are one of 10 common words

List of 10 common words:

the, of, and, a, be, this, there, an, been, some

Now the result2.txt for task in part (a) becomes:

\*\*\*\*\*\*\*\*\*result2.txt\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Levenshtein distance is 5

Insertions 0

Deletions 3

Substitutions 2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Deliverables: Word file with code and screenshots of output.**