110-2
Natural Language Processing

HW - 1

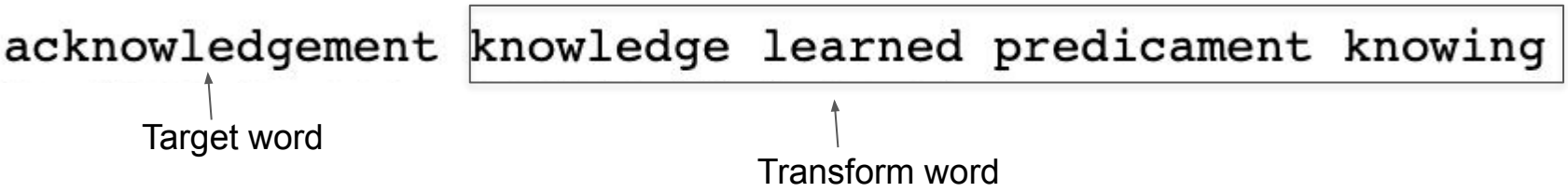
TA: Kuei-Chun Kao

Minimum Edit Distance

- The minimum edit distance between two strings
- Is the minimum number of editing operations
  - Insertion
  - Delete
  - Substitution
- Needed to transform one into the other

Inputs

- The input consists of sets of words (one set per line in lowercase). The first word in each line is the target word. All other words in the line must be transformed to the target word.
- Test input file will be provided. (input.txt)
- Two additional files are provided:
  - costs1.csv - Levenshtein substitution costs for lowercase alphabet
  - costs2.csv - weighted substitution costs for lowercase alphabet

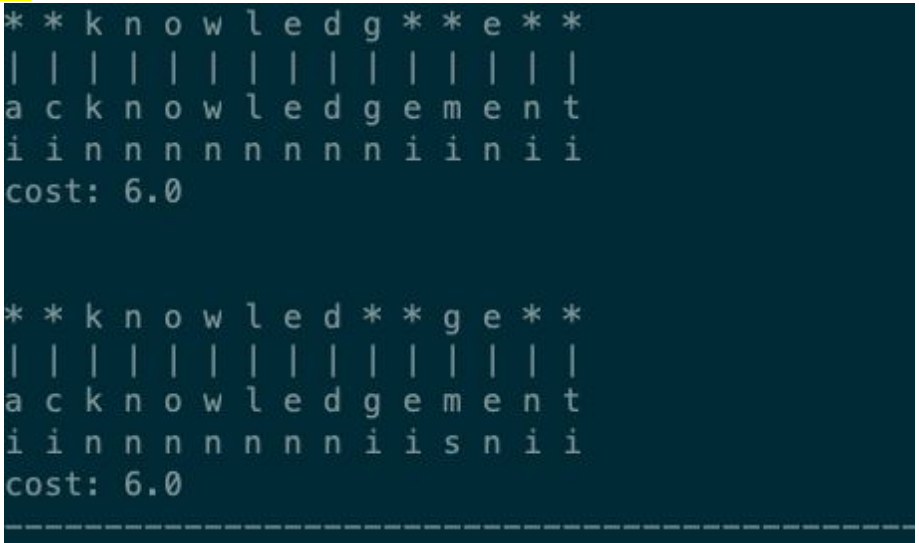


Workflows

- The cost of insertions and deletions is 1 in all cases. Substitution costs will be read from input files. (cost1.scv & cost2.csv)
- For each pair of source and target words, the minimum edit distance is calculated **(using both Levenshtein and weighted matrix costs)**, and the cost and backtrace of operations are in the output.
- Use dynamic programming and backtracking.
- When constructing the backtrace, any one of the possible cells that provide the minimum cost to the cell being processed is randomly selected.

Outputs

- 4 output lines for each of the method
  - Line 1 shows the source word
  - Line 2 contains a vertical bar ("|") for each operation (one per character)
  - Line 3 shows the target word
  - Line 4 show the operations for each character. Letter ‘n’ indicates a null operation (rather than a space). Letter ‘i’ means insert. Letter ‘s’ means substitution. Letter ‘d’ means delete.



Requirements

- Python only
- No plagiarism!
- At the top of your Source code

```
#Author: Kuei-Chun Kao

#Student ID: 1234567

#HW ID: hw1

#Due Date: 01/30/2020
```

Submission

- Deadline
  - Submit Zip to E3 before **3/18 11:59 PM**
  - No Late Submission, thanks!
- Format
  - Source code: Hw1_<StudentID>.py (py only)
  - Report file: Hw1_<StudentID>.pdf (pdf only)
  - Make sure the .py file contains the **correct execution results and formats**.
  - If can’t compile correctly, no score for you
  - Also zip input.txt & cost1.csv & cost2.csv
  - Zip file: Hw1_<StudentID>.zip (zip only)
- Any question can ask me on E3, answer your question ASAP

Grading policy

- Above requirements (60%)
- Test your program to ensure different backtracking results (10%)
- Code comments and formats (10%)
- Report (20%)