

NLP Homework 2, Part 1

Vahid Kharazi

November 27, 2015

1 MINIMUM EDIT DISTANCE

According to wikipedia and course slides, we have to implement minimum edit distance algorithm:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

To set a custom cost to this algorithm, we should change initialize matrix with insertion and deletion cost and change +1 with *cost*:

Initialization:

$$D(0,0) = 0$$

$$D(i,0) = D(i-1,0) + \text{del}[x(i)]; \quad 1 < i \leq N$$

$$D(0,j) = D(0,j-1) + \text{ins}[y(j)]; \quad 1 < j \leq M$$

Recurrence Relation:

$$D(i,j) = \min \begin{cases} D(i-1,j) + \text{del}[x(i)] \\ D(i,j-1) + \text{ins}[y(j)] \\ D(i-1,j-1) + \text{sub}[x(i),y(j)] \end{cases}$$

Test: There is an example in course slides to compute edit distance between intention and execution. here is my result of this example:

```

(venv)vahid@kharazi:~/dev/nlp/two$ python one.py
0  1  2  3  4  5  6  7  8  9
1  2  3  4  3  4  5  6  7  8
2  3  4  5  4  5  6  7  8  9
3  4  5  6  5  6  7  8  9  10
4  5  6  7  6  7  8  9  10  11
5  6  7  8  7  8  9  10  11  12
6  7  8  7  8  9  8  9  10  11
7  6  7  8  9  10  9  8  9  10
8  7  8  9  10  11  10  9  8  9
9  8  7  8  9  10  11  10  9  8

-----

o  o  o  o  o  o  o  o  o  o
o  TLD TLD TLD D  L  L  L  L  L
o  TLD TLD TLD T  TLD TLD TLD TLD TLD
o  TLD TLD TLD TD TLD TLD TLD TLD TLD
o  TLD TLD TLD T  TLD TLD TLD TLD TLD
o  TLD TLD TLD T  TLD TLD TLD TLD TLD
o  TLD TLD D  TL TLD D  L  L  L
o  D  L  TL TLD TLD T  D  L  L
o  T  TLD TLD TLD TLD T  T  D  L
o  T  D  L  L  LD TL  T  T  D

-----

o  o  o  o  o  o  o  o  o  o
o  TLD ,  TLD D  L  L  L  L  L
o  TLD TLD ,  T  TLD TLD TLD TLD TLD
o  TLD TLD TLD ,  TLD TLD TLD TLD TLD
o  TLD TLD TLD ,  TLD TLD TLD TLD TLD
o  TLD TLD TLD T  ,  TLD TLD TLD TLD
o  TLD TLD D  TL TLD ,  L  L  L
o  D  L  TL TLD TLD T  ,  L  L
o  T  TLD TLD TLD TLD T  T  ,  L
o  T  D  L  L  LD TL  T  T  ,

-----

['i', 'n', 't', 'e', 'n', 't', 'i', 'o', 'n']
['s', 's', '-', 'i', 's', '-', '-', '-', '-']
['e', 'x', 'e', 'c', 'u', 't', 'i', 'o', 'n']
(8, ['s', 's', '-', 'i', 's', '-', '-', '-', '-'])
(venv)vahid@kharazi:~/dev/nlp/two$

```

The first matrix is D, Second matrix is back-track(T = Top, L = Left, D = Diag).
Finally you can find the path of minimum edit distance: