

Natural Language Processing

Lab 1.2

January 18, 2024

Dr Phil Smith

This lab sheet is to practice the concepts taught so far with a focus on byte pair encoding and minimum edit distance.

1. What is tokenization in the context of natural language processing?
2. Explain Byte Pair Encoding with an example.
3. Create a basic Byte Pair Encoding algorithm in Python that will work with a test corpus.
4. What is the minimum edit distance and its significance in NLP?
5. Compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 1) of *leda* to *deal*. Show your work (using the edit distance grid).
6. Figure out whether *drive* is closer to *brief* or to *divers* and what the edit distance is to each. Use Levehnstein distance.
7. Now implement a minimum edit distance algorithm and use your hand-computed results to check your code.

a	4	3	3	2	3
d	3	2	2	2	3
e	2	2	1	2	3
l	1	1	2	3	3
#	0	1	2	3	4
	#	d	e	a	/

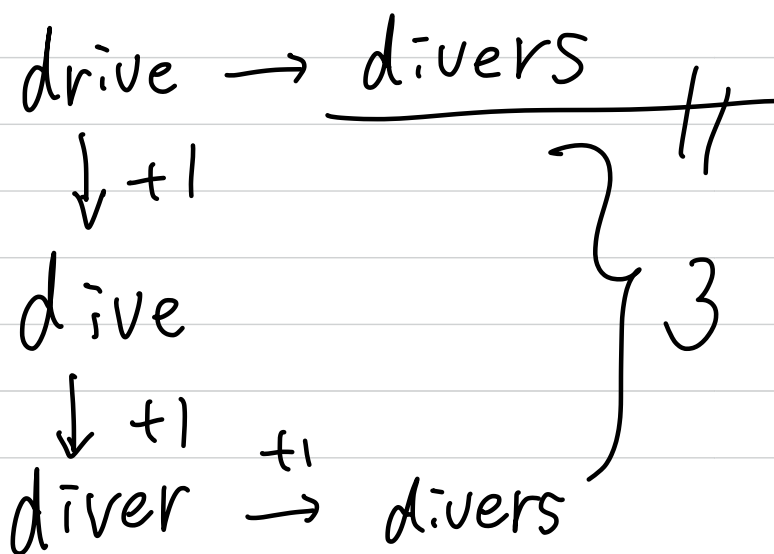
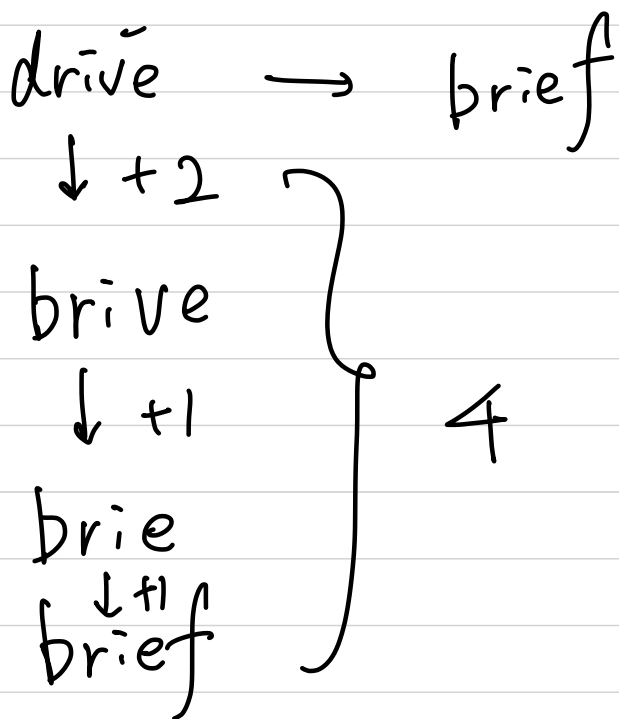
$l \rightarrow de$

$le \rightarrow d$

$leda \rightarrow dea$

$\begin{matrix} l & e & d \\ \downarrow & & \downarrow \\ d & e & a \end{matrix}$

$\begin{matrix} & & d & e & d & a \\ & & \downarrow & & & \\ & & d & e & a \end{matrix}$



(5)

1	4					
a	3					
e	2					
d	1	2	3			
#	0	1	2	3	4	
	#	1	e	d	a	

