Lab Class NLP:IV

By June 23, 2025, solutions for the following exercises have to be submitted: 1, 2, 3, and 4.

Exercise 1 : Byte-Pair Encoding

Byte-pair encoding (BPE) is a common tokenization technique used in NLP to segment words into subword units. It is based on the idea of merging the pairs of characters or subword units.

(a) BPE Rule Finding: Assume that you have already preprocessed some toy dataset and split the text into strings. The following table shows the frequency of each string in your preprocessed text data:

string	freq
the	10
cat	8
sat	3
on	6
mat	5

Note: Do not include whitespace, start-of-word, and end-of-word tokens in your solution.

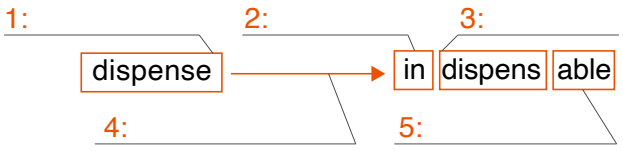
Train the BPE algorithm on this toy dataset with the number of merge operations R set to 5. Create an initial index I_0 and vocabulary V_0 based on the dataset. Then, manually compute the 5 BPE merge operations. At each step j, write the new merge rule R_j, vocabulary V_j, and index I_j.

(b) Tokenization: Apply the BPE tokenizer you trained in the task (a) to the following strings: "`month`", "`other`", "`thecat`", "`cats`". Write down each step of the tokenization process.

(c) What are the advantages of using BPE for tokenization? Write down at least two advantages. Use your results from (b) to support your answer.

Exercise 2 : Morphology

(a) Fill in the blanks in the following illustration with the correct morphological terms from the given set.



`derivation` `prefix` `suffix` `inflection` `stem` `infix` `morpheme` `lexeme` `root` `lemma` `affix`

(b) What affixes are in the word "reactors"? Identify whether each affix is *derivational* or *inflectional*.

(c) Using the word "reactors" explain the difference between *root* and *stem*.

(d) You are given the following excerpt of rules from the Porter stemmer.

Index	Ruleset	Premise	Suffix	Replacement
(I)	1a	`null`	SSES	SS
(II)	1b	(*v*)	ING	`null`
(III)	1b	(*v*)	IZ	IZE
(IV)	1c	(*v*)	Y	I
(V)	2	(m>0)	BILITI	BLE
(VI)	2	(m>0)	IVENESS	IVE
(VII)	2	(m>0)	IZATION	IZE
(VIII)	3	(m>0)	NESS	`null`
(IX)	4	(m>1)	AL	`null`
(X)	4	(m>1)	IVE	`null`
(XI)	4	(m>1)	ABLE	`null`
(XII)	4	(m>1)	ITI	`null`
(XIII)	4	(m>1)	IZE	`null`
(XIV)	5	(m>1)	E	`null`

Stem the following words using the Porter stemmer with the rules given above. Note down the index of the rules you apply in order.

1. recognizability
2. recognization
3. recognizing
4. universalness
5. universe
6. university

(e) Are any of the words in the previous exercise under- or over-stemmed? Which problems (if any) can arise if the words are under- or over-stemmed?

Exercise 3

Fill in the following cloze. Use the terms from the box below. Terms may be used multiple times and not all terms must be used.

lemma • stem • Krovetz Stemmer • morphology • text preprocessing • stemming • effectiveness • canonical form • morpheme • root • Porter Stemmer • efficiency • lemmatization • maximal unit • morphological analysis • minimal unit

_____ aims to convert text into a _____ to improve information retrieval _____. _____ is the study of the structure and formation of words, where a _____ is the _____ of meaning. The word's _____ is the derivational base of a word, while a _____ is its inflectional base. _____ is the identification of a word's morphemes and their role.

_____ maps a word token to its word _____ by removing inflection, while _____ maps it to its dictionary form, or _____. The _____ applies nine sets of rules, each containing between 1 and 20 rules. In contrast, the _____ combines a dictionary-based approach with rules.

Exercise 4 : Tokens and Types

What (word-)tokens and which types appear in the following text?

```
The bear could not bear the cold, so it went back into its cave.
```