|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NLP 2025 Lab 1 Report: Tokenization** | | | | |
|  | | | | |
|  | **First Author** | **Second Author** | **Third Author** |  |
|  | Affiliation / Address line 1 | Affiliation / Address line 1 | Affiliation / Address line 1 |  |
|  | email@domain | email@domain | email@domain |  |
|  | | | | |
|  | | | | |

Abstract

|  |  |  |
| --- | --- | --- |
| **Type of Text** | **Font Size** | **Style** |
| paper title | 15 pt | **bold** |
| author names | 12 pt | **bold** |
| author affiliation | 12 pt |  |
| the word “Abstract” | 12 pt | **bold** |
| section titles | 12 pt | **bold** |
| document text | 11 pt |  |
| captions | 10 pt |  |
| abstract text | 10 pt |  |
| bibliography | 10 pt |  |
| footnotes | 9 pt |  |

Table : Font guide.

DELETE (OR REPLACE) AFTER YOU READ. This is a template for the report of the NLP 2025 labs. It is adapted from the Association for Computational Linguistics (ACL) conference[[1]](#footnote-1) word template[[2]](#footnote-2). Fill in your names in the authors list and provide your answers to the exercises listed in the accompanying notebook. The answer to each exercise should be in a separate section. For the coding exercises, provide code snipped and a short explanation. For the open questions, you can add figures, plots, tables, etc. Make sure to include a short discussion.

You need to submit a PDF of your report (use the templates provided in latex (preferred) or word and a copy of your notebook with the code. Make sure that all figures/results/plots are shown in your (execucted) notebook. While you may talk with others about this lab, we ask that you write your solutions with your group only. If you do discuss specific tasks with others please include their names in the appendix. Honor code applies to this lab. For more check Syllabus §7.2. It is mandatory to list and disclose any website (or other resource) you used (e.g. stackoverflow) as well as any genAI tools (e.g. chatGPT) used. Failure to disclose these is a violation of academic integrity. For more check Syllabus §7.3. THIS TEXT WILL BE SELF-DESTROYED AFTER YOU READ IT.

Formatting instructions

In this short section we will show you several examples of elements that you can adapt when writing your answers. Feel free to remove this section before submitting your report.

To write a code snippet you use the Courier New font. For example:

def foo():

print('Hello World')

Here is the example of a table. Each table and figure should have a reference from text. For example, Table 1 and Figure 1 can be referenced using References/Cross-reference option in Word.

|  |
| --- |
|  |
| Figure 1: A figure with a caption that runs for more than one line**.** |

Exercise 1: Questions about the datasets

Your answers here.

Exercise 2: Write the text cleaning function

Exercise 3: Build the vocabulary

Exercise 4: Frequency of pairs of words

Exercise 5: Tokenize the dataset

Exercise 6: Questions about the tokenization

Exercise 7: Counting the characters

Exercise 8: Calculate the frequency statistics of adjacent symbol pairs

Exercise 9: BPE algorithm

Exercise 10: Comparing tokenizers

1. Collaborators outside our group

Our group talked with Elon Musk on how to build the BPE tokenizer for Grok...

1. Use of genAI

Use of genAI tools (e.g. chatGPT), websites (e.g. stackoverflow): list websites where you found code (or other info) as well as include information on how you used genAI tools (e.g. prompts). Failure to disclose use of genAI tools is violation of adademic integrity.

1. <https://github.com/acl-org/acl-style-files> [↑](#footnote-ref-1)
2. <https://github.com/acl-org/acl-style-files> [↑](#footnote-ref-2)