Lab 1: Introduction into Natural Language Processing (5 points total)

In this exercise you will have to implement text preprocessing, the Bag of Words model and calculate cosine similarity between two large texts.

Text 1: “Distributed Representations of Words and Phrases and their Compositionality” [link].

Text 2: “Attention is All You Need” [link].

Step 1: Implement word-based or subwordy tokenization using only regular expressions and pure Python. It’s expected from you:

- to remove special characters and tags from the documents,
- to identify and justify the list of stop-words and remove them;

Step 2: Implement the Bag of Words model using pure Python and NumPy only. Shrink the Bag of Words model and justify your choice (clue: think about how the number of word appearances is distributed in your model).

Step 3: Generate vectors for both documents and calculate cosine similarity between them. Use pure Python and Numpy only.

Step 4: Try to interpret measured cosine similarity.

Requirements:

- To use only: pure Python + Python Standard Libraries + NumPy + Re + one library of your choice to read the .pdf file,
- Explain and justify each step of your solution.

Submission:

- Deadline: Friday (June 7), 23:59
- Submit exactly one .ipynb file