+FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

**Report**

**on the practical task No. 2**

**“Natural language processing”**

Performed by

Chernobrovkin T.V. (412642)

Academic group J4133c

Accepted by

Gladilin P.E.

St. Petersburg

2023

# Goal

Learn and put into practice the basic principles of NLP.

# Problem

1. Download Alice in Wonderland by Lewis Carroll from Project Gutenberg's website http://www.gutenberg.org/files/11/11-0.txt

2. Perform any necessary preprocessing on the text, including converting to lower case, removing stop words, numbers / non-alphabetic characters, lemmatization.

3. Find Top 10 most important (for example, in terms of TF-IDF metric) words from each chapter in the text (not "Alice"); how would you name each chapter according to the identified tokens?

4. Find the Top 10 most used verbs in sentences with Alice. What does Alice do most often?

# Theory

Natural Language Processing (NLP) is a machine learning technology that gives computers the ability to interpret, manipulate and understand human language. Organizations today have large amounts of voice and text data from various communication channels such as emails, text messages, social media news feeds, video, audio, and more. They use NLP software to automatically process this data, analyze the intent or sentiment in the message, and respond to human communication in real time.

# Materials and methods

In this task, all calculations were performed on the student's personal laptop in Google Colab. The work was performed in the Python programming language.

# Results

1. The text of Lewis Carroll's "Alice in Wonderland" was downloaded from the website Project Gutenberg <http://www.gutenberg.org/files/11/11-0.txt>.
2. Performed text preprocessing including conversion to lower case, removal of stop words, digits/nonalphabetic characters, and lemmatization.
3. The Top 10 most important words from each chapter of the text were found (Figure 1). We also came up with new chapter titles according to the resulting word list (Figure 2)..

Изображение выглядит как текст, Шрифт, снимок экрана, документ

Автоматически созданное описание

Figure 1 - Top 10 most important words

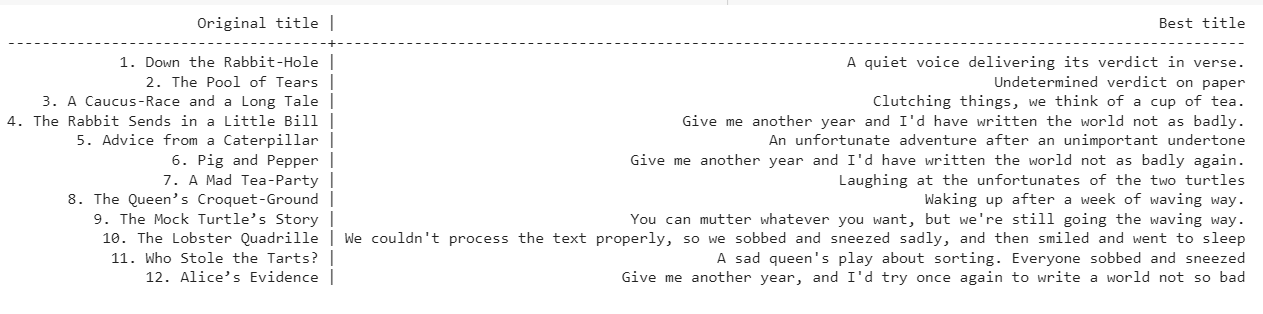


Figure 2 - New names

1. The top 10 most frequently used verbs with Alice in sentences were found (Figure 3). It turned out that among the top 10 actions, Alice spoke most often and thought least often.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Figure 3 - Alice's most frequent actions

# Conclusion

NLP algorithms were studied and practiced in this assignment. The book "Alice in Wonderland" was processed. The Top 10 most important words from each chapter of the text and the 10 most frequently used verbs with Alice were found.

# Appendix

GitHub link:

<https://github.com/LesostepnoyGnom/homework_ML/blob/main/task2/Task_2_Chernobrovkin_J4133c.ipynb>