+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

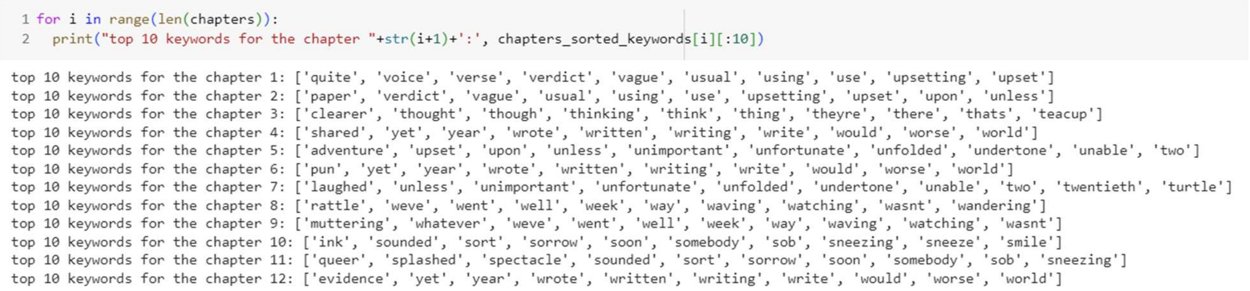
Natural Language Processing (NLP) is a machine learning technology that gives computers the ability to interpret, manipulate and understand human language. Organizations today have large amounts of voice and text data from various communication channels such as emails, text messages, social media news 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

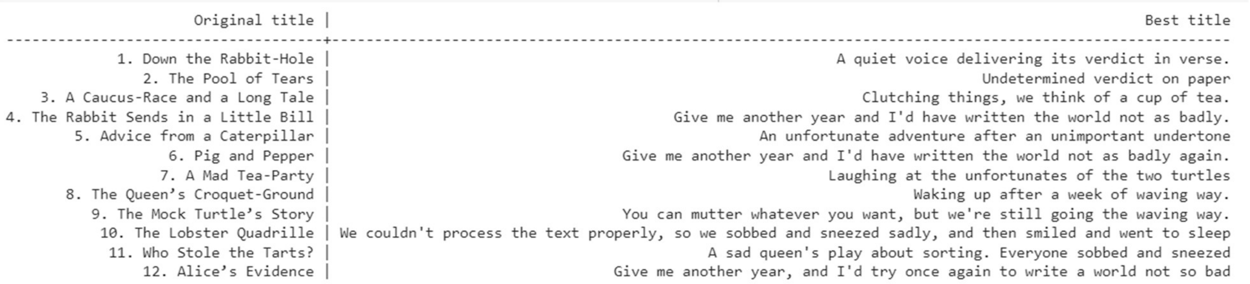
1. The text of Lewis Carroll's "Alice in Wonderland" was downloaded from the website Project Gutenberg http://www.gutenberg.org/files/11/11-0.txt.
2. Performed text preprocessing including conversion to lower case, removal of stop words, digits/nonalphabetic characters, and lemmatization.
3. The Top 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)..



```python
for i in range(len(chapters)):
  print("top 10 keywords for the chapter "+str(i+1)+':', chapters_sorted_keywords[i][:10])
```

```
top 10 keywords for the chapter 1: ['quite', 'voice', 'verse', 'verdict', 'vague', 'usual', 'using', 'use', 'upsetting', 'upset']
top 10 keywords for the chapter 2: ['paper', 'verdict', 'vague', 'usual', 'using', 'use', 'upsetting', 'upset', 'upon', 'unless']
top 10 keywords for the chapter 3: ['clearer', 'thought', 'though', 'thinking', 'think', 'thing', 'theyre', 'there', 'thats', 'teacup']
top 10 keywords for the chapter 4: ['shared', 'yet', 'year', 'wrote', 'written', 'writing', 'write', 'would', 'worse', 'world']
top 10 keywords for the chapter 5: ['adventure', 'upset', 'upon', 'unless', 'unimportant', 'unfortunate', 'unfolded', 'undertone', 'unable', 'two']
top 10 keywords for the chapter 6: ['pun', 'yet', 'year', 'wrote', 'written', 'writing', 'write', 'would', 'worse', 'world']
top 10 keywords for the chapter 7: ['laughed', 'unless', 'unimportant', 'unfortunate', 'unfolded', 'undertone', 'unable', 'two', 'twentieth', 'turtle']
top 10 keywords for the chapter 8: ['rattle', 'weve', 'went', 'well', 'week', 'way', 'waving', 'watching', 'wasnt', 'wandering']
top 10 keywords for the chapter 9: ['muttering', 'whatever', 'weve', 'went', 'well', 'week', 'way', 'waving', 'watching', 'wasnt']
top 10 keywords for the chapter 10: ['ink', 'sounded', 'sort', 'sorrow', 'soon', 'somebody', 'sob', 'sneezing', 'sneeze', 'smile']
top 10 keywords for the chapter 11: ['queer', 'splashed', 'spectacle', 'sounded', 'sort', 'sorrow', 'soon', 'somebody', 'sob', 'sneezing']
top 10 keywords for the chapter 12: ['evidence', 'yet', 'year', 'wrote', 'written', 'writing', 'write', 'would', 'worse', 'world']
```

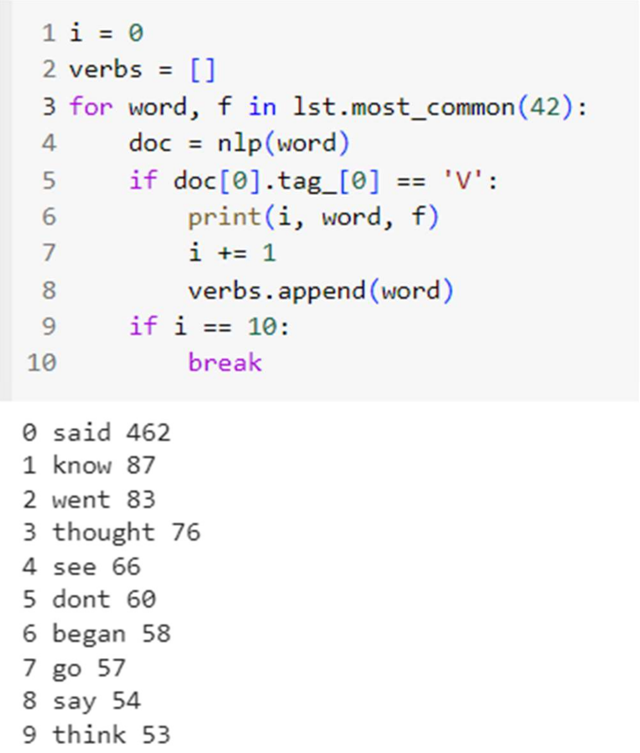
Figure 1 - Top 10 most important words



```
                      Original title |                                                                                              Best title
-------------------------------------+---------------------------------------------------------------------------------------------------------
             1. Down the Rabbit-Hole |                                                          A quiet voice delivering its verdict in verse.
                2. The Pool of Tears |                                                                           Undetermined verdict on paper
    3. A Caucus-Race and a Long Tale |                                                             Clutching things, we think of a cup of tea.
4. The Rabbit Sends in a Little Bill |                                       Give me another year and I'd have written the world not as badly.
        5. Advice from a Caterpillar |                                                 An unfortunate adventure after an unimportant undertone
                   6. Pig and Pepper |                                 Give me another year and I'd have written the world not as badly again.
                  7. A Mad Tea-Party |                                                         Laughing at the unfortunates of the two turtles
       8. The Queen’s Croquet-Ground |                                                                   Waking up after a week of waving way.
          9. The Mock Turtle’s Story |                                 You can mutter whatever you want, but we're still going the waving way.
           10. The Lobster Quadrille | We couldn't process the text properly, so we sobbed and sneezed sadly, and then smiled and went to sleep
            11. Who Stole the Tarts? |                                           A sad queen's play about sorting. Everyone sobbed and sneezed
                12. Alice’s Evidence |                                Give me another year, and I'd try once again to write a world not so bad
```

Figure 2 - New names

4. 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.



```python
i = 0
verbs = []
for word, f in lst.most_common(42):
    doc = nlp(word)
    if doc[0].tag_[0] == 'V':
        print(i, word, f)
        i += 1
        verbs.append(word)
    if i == 10:
        break
```

```
0 said 462
1 know 87
2 went 83
3 thought 76
4 see 66
5 dont 60
6 began 58
7 go 57
8 say 54
9 think 53
```

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