МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №5
по курсу «Автоматическая обработка текстов»

«Лемматизация»

Выполнил:
студент группы ИУ9-11М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2018

1 Цель работы

Обработать текст и построить частотный список получившихся лемм.

2 Текст программы

```python
import re
import pymorphy2
from collections import Counter

morph = pymorphy2.MorphAnalyzer()

file_name = "a_dance_with_dragons.txt"
N_WORDS_TO_PRINT = 50

def read_words(filename: str) -> list:
    contents = open(filename, 'r').read().casefold()
    clean = re.sub(r'[^a-z0-9]', ' ', contents)      # add cyrillic
    splitted = re.compile(r'\s+').split(clean)
    return splitted[:-1] # remove last empty word

def main():
    word_list = read_words(file_name)

    normalized = list(map(lambda word: morph.parse(word)[0].normal_form,
        word_list))
    sorted_by_rank = Counter(normalized).most_common()

    with open('out.txt', 'w+') as out:
        for word, count in sorted_by_rank[:N_WORDS_TO_PRINT]:
            print(f'{word}\t\t{count}', file=out)

if __name__ == '__main__':
    main()
```

Листинг 1: Исходный код программы

3 Результаты тестирования

В качестве художественного произведения был взят роман «Танец с драконами» из цикла «Песнь Льда и Пламени» Джорджа Р. Р. Мартина. В ходе работы текст был обработан и лемматизирован с помощью библиотеки `pymorph2`. Результаты представлены в таблице 1.

Из документации к Pymorphy2: «*Pymorphy2 использует алгоритм нахождения нормальной формы, который работает наиболее быстро (берется первая форма в лексеме) - поэтому все причастия нормализуются в инфинитивы.*»

4 Выводы

В ходе работы была произведена обработка и лемматизация литературного произведения. Общее количество уникальных слов сократилось, а частота встречи лемм гораздо выше частоты встречи аналогичных слов.

Таблица 1: Наиболее часто встречаемые леммы

лемма	частота
и	10198
он	7911
в	6640
не	5337
на	4774
быть	3990
с	3654
что	3420
они	3158
это	2546
но	2418
она	2358
как	2094
из	2002
а	1777
весь	1772
за	1645
ты	1618
свой	1577
мы	1537
вы	1357
её	1356
тот	1352
когда	1298
мочь	1291
то	1269
от	1246
один	1169
бы	1111
мой	1110
чтобы	1066
лорд	1049
если	1031
человек	1021
так	973
сказать	956
этот	928
джон	871
рука	835
ещё	830
тириона	696