




# **Влияние лемматизации на извлечение палиндромов из текста на примере цикла Клайва. С. Льюиса "Хроники Нарнии"**

Подготовила: Денисенко С.



# Задачи

1. Предобработка текста
2. Частотный анализ:
  - Сделать частотный словарь и график распределения топ-20 слов
  - Посчитать количество слов (русская версия и оригинал)
  - Посчитать TTR текста (русская версия и оригинал)
  - Сравнить результаты
3. Лемматизация текста (русская версия и оригинал)
4. Поиск палиндромов в исходном тексте (русская версия и оригинал)
5. Поиск палиндромов в лемматизированном тексте (русская версия и оригинал)
6. Сравнение результатов
7. Частеречный анализ словаря палиндромов, полученных из лемматизированной версии текста (русская версия)



Предполагается, что лемматизация текста способствует извлечению палиндромов из текста, благодаря приведению слова к его изначальному виду.

Ожидаемый результат: в лемматизированном тексте палиндромов должно быть больше, чем в исходном.



## Основные библиотеки

- NLTK
- pymorphy3
- matplotlib
- pandas
- wordcloud



## Предобработка текста

```
def preprocessing(text):  
    lower_text = text.lower()  
    without_punkt = re.sub(r'^\w\s]', '', lower_text)  
    clean_text = re.sub(r'\n+', ' ', without_punkt)  
    return clean_text
```



# Частотный анализ

## 1.1. Подсчет общего количества слов в тексте + ttr

```
def ttr(text):  
    ###посчитаем количество слов  
    word_count = len(text.split()) #  
    unique_word_count = len(set(text.split()))  
    ###посчитаем TTR  
    ttr = unique_word_count/word_count  
    return ' - общее количество слов в тексте ' + str(word_count) + '.' +  
    '\n - уникальных слова в тексте ' + str(unique_word_count) + '.' +  
    '\n - type token ratio составляет ' + str(ttr)
```



```
print('В переведенной на русский язык версии текста: \n' + ttr(clean_ru_text))
```

В переведенной на русский язык версии текста:

- общее количество слов в тексте 215218.
- уникальных слова в тексте 29205.
- type token ratio составляет 0.1356996162031057

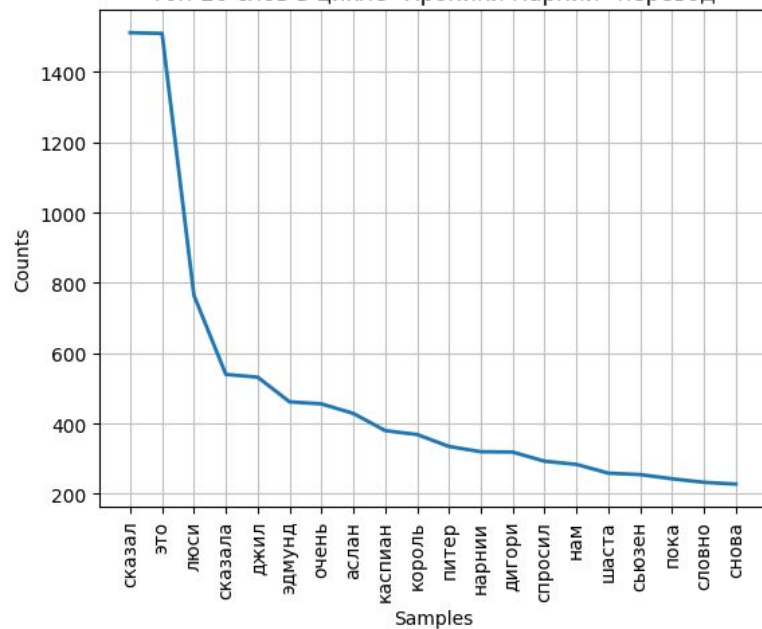
```
print('В оригинальной версии текста: \n' + ttr(clean_eng_text))
```

В оригинальной версии текста:

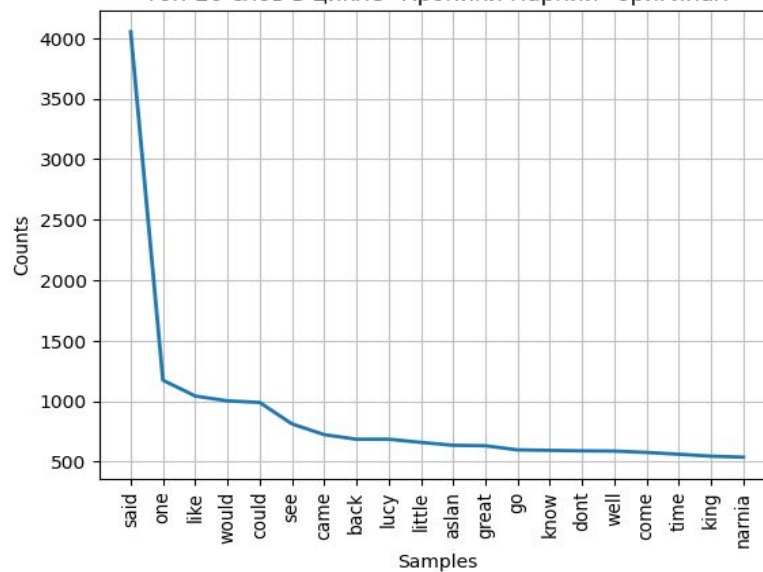
- общее количество слов в тексте 319507.
- уникальных слова в тексте 12602.
- type token ratio составляет 0.03944201535490615

## 1.2 Графики распределения самых частотных слов

Топ-20 слов в цикле "Хроники Нарнии" перевод



Топ-20 слов в цикле "Хроники Нарнии" оригинал







# Лемматизация

## 1.1. Для русского языка

```
def lemmas_ru(text):  
    lemmas = []  
    for token in word_tokenize(text):  
        token_lemma = morph.parse(token)[0].normal_form  
        lemmas.append(token_lemma)  
    lemmatized_text = ' '.join(lemmas)  
    return lemmatized_text
```



## 1.2. Для английского языка

```
def lemmas_eng(text):  
    lemmatizer = WordNetLemmatizer()  
    tokens = word_tokenize(text)  
    lemmatized_words = [lemmatizer.lemmatize(word) for word in tokens]  
    lemmatized_text = ' '.join(lemmatized_words)  
    return lemmatized_text
```



## Поиск палиндромов

```
def find_palindrome(text):  
    list_of_palindrome = []  
    for i in text.split():  
        if len(i) > 2 and i == i[::-1]:  
            list_of_palindrome.append(i)  
    return list_of_palindrome
```







	token	lemma	pos_tag
0	aaa	aaa	INTJ
1	ooo	ooo	NOUN
2	хоххоххох	хоххоххох	NOUN
3	угу	угу	INTJ
4	топот	топот	NOUN
5	тот	тот	ADJF
6	летел	лететь	VERB
7	ими	они	NPRO
8	корок	корка	NOUN
9	еще	ещё	ADVB
10	кок	кок	NOUN
11	эээ	эээ	INTJ
12	охохо	охохо	ADVB
13	тащат	тащить	VERB
14	еде	еда	NOUN
15	рррррр	рррррра	NOUN
16	иди	идти	VERB
17	шалаш	шалаш	NOUN
18	как	как	CONJ
19	рррр	рррра	NOUN
20	течет	течь	VERB
21	шшш	шшш	None

22	обо	о	PREP
23	мадам	мадам	NOUN
24	око	око	NOUN
25	ого	ого	INTJ
26	ага	ага	INTJ
27	еле	еле	ADVB
28	или	или	CONJ
29	уху	уха	NOUN
30	mmm	mmm	NOUN
31	комок	комок	NOUN
32	воров	вор	NOUN
33	дед	дед	NOUN
34	oooo	oooo	NOUN
35	огого	огий	ADJF
36	лил	лить	VERB
37	елееле	елееле	ADVB
38	дмд	дмд	None
39	ухху	ухха	NOUN
40	мэм	мэм	None
41	мэээм	мэээма	NOUN
42	оно	оно	NPRO
43	тут	тут	ADVB
44	зараз	зараз	ADVB



**Спасибо за внимание!**