Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчет

по заданию №1
«Сбор собственного корпуса текстовых документов»

по дисциплине «Компьютерная лингвистика»

Автор: Лакиза Александр Николаевич

Факультет: ИКТ

Группа: К3242

Преподаватель: Чернышева Анастасия Владимировна

Санкт-Петербург 2021

Цель работы: Собрать собственный корпус текстовых документов при помощи инструментов программирования для использования его в будущем в качестве материала для изучения дисциплины «Компьютерная лингвистика»

Требования к корпусу: Корпус должен быть собран из документов на русском языке, в корпус должно входить не менее 200 документов и не менее 10000 слов

Ход работы: Я решил поработать с Википедией и её api-шкой для языка Python, так как никогда до этого при сборе информации с ней не работал.

```
1. wiki = wikiapi.Wikipedia("ru")
2. page = wiki.page("Титаник")
```

Указываем, на каком языке хотим получать данные и какую страницу будем первой парсить. Я решил, что в этом задании я буду собирать короткие «саммарис» каждой страницы, на которую есть гиперссылка в статье о Титанике.

```
1. links = page.links
2. ru_links = [i for i in sorted(list(links.keys())) if ord(i[0]) >= 1040
3.             and ord(i[-1]) >= 1040]
4. ru_links = [i for i in ru_links if ':' not in i]
5. ru_links = [i for i in ru_links if ',' not in i]
```

Я положил названия всех ссылок в массив links, а затем решил удалить все ссылки, которые содержать английские буквы на первом и последнем месте (этого достаточно, чтобы оставить лишь ссылки на русском), которые содержат знаки двоеточие и запятую, так мы избавимся от лишних ссылок таких, как «Шаблон: …» и ссылок на статьи о людях. (Примечательно, что некоторые статьи о людях содержат запятую между именем и фамилией, а некоторые – нет). Заодно мы сократили количество ссылок с >500 до чуть больше 350

```
1. data = {}
2.
3. for i in ru_links:
4.     page = wiki.page(i)
5.     if page.exists():
6.         data[i] = page.summary
7.     time.sleep(1)
8.     print(i, 'is done')
```

Инициализируем словарь для дальнейшего хранения всех «саммарис» и пишем цикл для того, чтобы спарсить «саммари» с каждой ссылки. В течение работы над сбором данных не раз столкнулся с проблемой большого количества запросов на сайт, что в последствие выдавало ошибку. Справиться с проблемой я решил в лоб и просто добавил `time.sleep(1)`.

`data['corpus_length'] = len(data)` Последним элементом в словарь добавил информацию о количестве спарсенных «саммарис». У меня получилось их 363 штуки.

```
1. with open("titanic_data.json", "w") as f:
```

```
    # Записываем наш корпус в формате .json
    json.dump(data, f, ensure_ascii=False)
```

И записываем наш корпус в формате json.

Говоря, о параметрах корпуса, уже было сказано, что у меня получилось 363 источника, тематика – Титаник, корабельная тема, немного географии, среднее количество слов могу указать лишь навскидку, так как «саммарис» содержать много разных примечаний (например, перевод с греческого), которые надо редактировать, чтобы посчитать количество слов. По моим грубым подсчётам в корпусе около 20 тысяч слов.

Ссылка на исходный код и json файл с корпусом: https://github.com/alexanderlakiza/cs224