

Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
**"Национальный исследовательский университет
Высшая школа экономики"**
Департамент прикладной математики, бакалавр

РАБОТА НА ТЕМУ:

Корпусные исследования академических текстов.

Выполнил:

Колодин Матвей Алексеевич

Руководитель группы:

Власова Екатерина Александровна

Москва, 2022

Содержание

1	Введение	3
2	Алгоритм решения задачи	3
2.1	Сбор коллекции	3
2.2	Создание текстовой коллекции	5
2.2.1	Очистка pdf-файла и сохранение результата в формате txt	5
2.2.2	Предобработка текстовых данных	12
2.3	Преобразование текстовых коллекций в лингвистический набор данных.	17
2.4	Примечание	24

1 Введение

Прежде всего, необходимо четко сформулировать условие задачи, решение которой будет представлено ниже. Имеется некоторый набор источников, в которых содержится научная литература (в формате pdf). Данные необходимо представить в текстовом формате и обработать.

Для решения задачи будет необходимо пройти 3 этапа - сбор данных, создание коллекции в формате txt и ее предобработка, преобразование текстовых коллекций в лингвистический набор данных.

Инструкция, в целом, имеет рекомендательный характер - иметь другой подход не воспрещается, но тот путь, который будет описан далее - наиболее оптимальный из всего, что я смог найти (но это не отменяет того, что вы можете найти путь лучше/удобнее/быстрее).

После небольшого предисловия перейдем к решению задачи.

2 Алгоритм решения задачи

2.1 Сбор коллекции

Пусть есть некоторый источник, например: сетевой научный журнал Вестник. Как скачать файл, конечно же, зависит от сайта - но в целом алгоритм более чем схожий:

1. Находим необходимую статью на сайте:

Гражданская активность в России: институты, мотивации, восприятие

Козырева П. М., Смирнов А. И.
Динамика субъективного властного статуса. С. 13-30
DOI: 10.19181/vis.2022.13.2.787

 [Текст статьи](#)

Рис. 1: Пример статьи, которую необходимо скачать.

2. Нажимаем на ссылку со статьей, в результате чего открывается новая страница с pdf-файлом. На данной странице скачиваем файл. Таким образом скачиваем все необходимые статьи, после чего переходим к 3 пункту.

Гражданская активность в России: институты, мотивации, восприятие

Козырева П. М., Смирнов А. И.
Динамика субъективного властного статуса. С. 13-30
DOI: 10.19181/vis.2022.13.2.787


 [Текст статьи](#)

Рис. 2: Открываем ссылку на статью.

3. Создаем папку, в которой будут храниться файлы и переносим их туда.
4. После чего переименовываем все файлы по следующей схеме:

«Название журнала» + «Год выпуска» + «Номер файла» + «Название»
(обычно название состоит из имен авторов и наименования их работы).
Таким образом, имена файлов должны выглядеть примерно следующим образом:
"Вестник института социологии.2022.№2 Козырева П. М., Смирнов А. И. Динамика субъективного властного статуса"

Важно соблюдать единообразие в именах файлов, стоящих перед частью «Название» (в данном примере - это "Вестник института социологии.2022.№2 ")

В результате чего получается коллекция, которая выглядит примерно следующим образом:

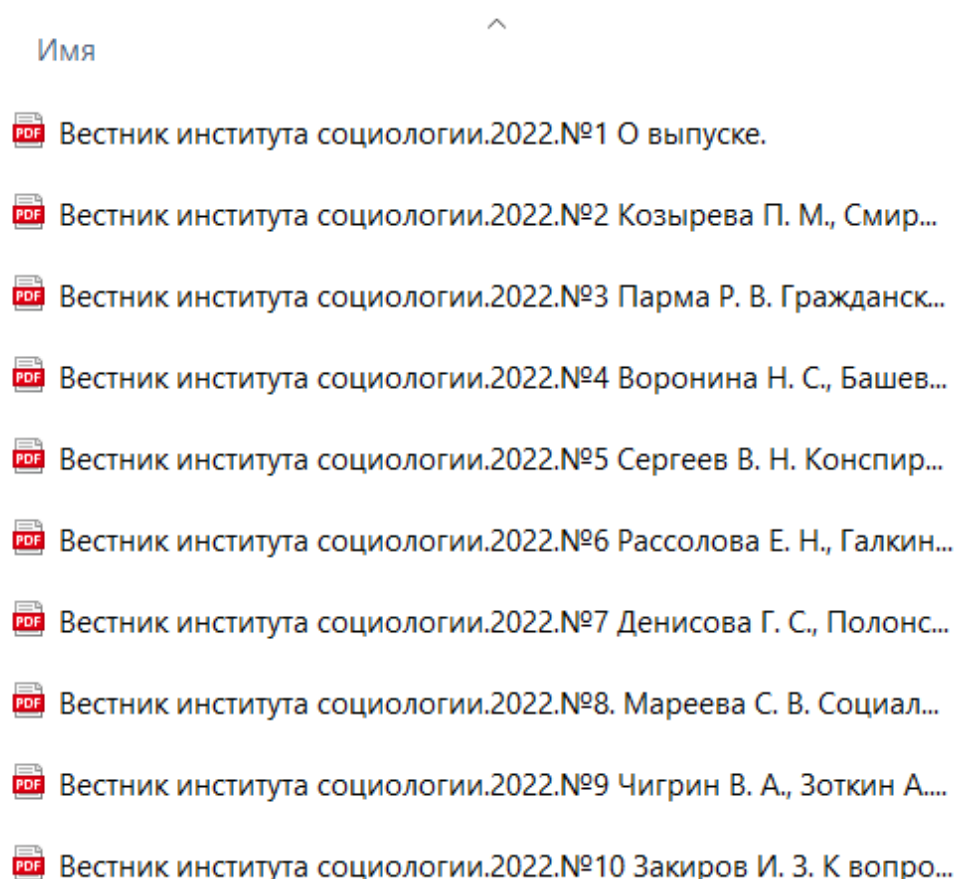


Рис. 3: Пример коллекции pdf-файлов.

После создания коллекции можно переходить к этапу ее обработки.

2.2 Создание текстовой коллекции

Как было сказано выше, данный этап будет делиться на две части - очистку pdf-файлов от ненужных элементов (графиков, картинок и т.д) и создание текстовых файлов.

2.2.1 Очистка pdf-файла и сохранение результата в формате txt

Для выполнения данного этапа будет необходимо иметь на компьютере программу FineReader. Поэтому перед тем как приступить к выполнению, скачайте программу с официального сайта или другого удобного вам источника. После установки запускаем программу, после чего открываем pdf-файл:

Просмотр и редактирование PDF-документов

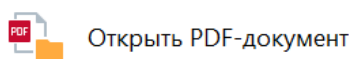


Рис. 4: Начало работы с pdf-файлом в FineReader.

Далее выполняем обработку поэтапно:

1. Номера страниц, колонтитулы и прочие объекты, которые находятся на краях, проще всего обрезать. Для этого на верхней панели выбираем инструмент «Обрезать»:

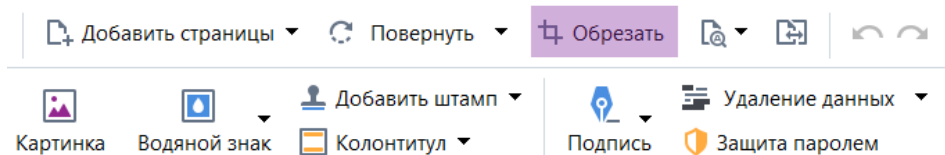


Рис. 5: Инструмент «Обрезать» на верхней панели.

После чего выбрать режим применения ко всем страницам:

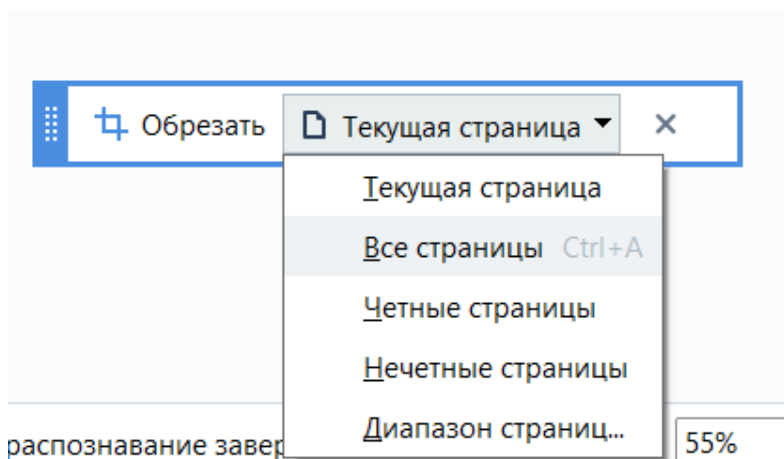


Рис. 6: Выбор режима инструмента «Обрезать».

Далее необходимо выбрать правильные размеры области - нужно не потерять полезные данные и отсеить как можно больше мусора.

Я рекомендую ориентироваться на первые три страницы (можно и больше, в зависимости от статьи, но обычно этого хватает) - по ним будут понятны границы текста на странице. На первой странице я стараюсь максимально близко подогнать границы области к картинкам, колонтитулам, чтобы была наибольшая область с полезной нам информацией.

По следующим двум страницам стоит корректировать область. К сожалению, стиль страниц в журналах хоть и очень схож, но бывает так, что область текста на странице смещается. По этой причине достаточно выбрать такую область, чтобы для первых трех страниц она была корректна, в большинстве случаев, такое выделение подходит и к оставшимся страницам.

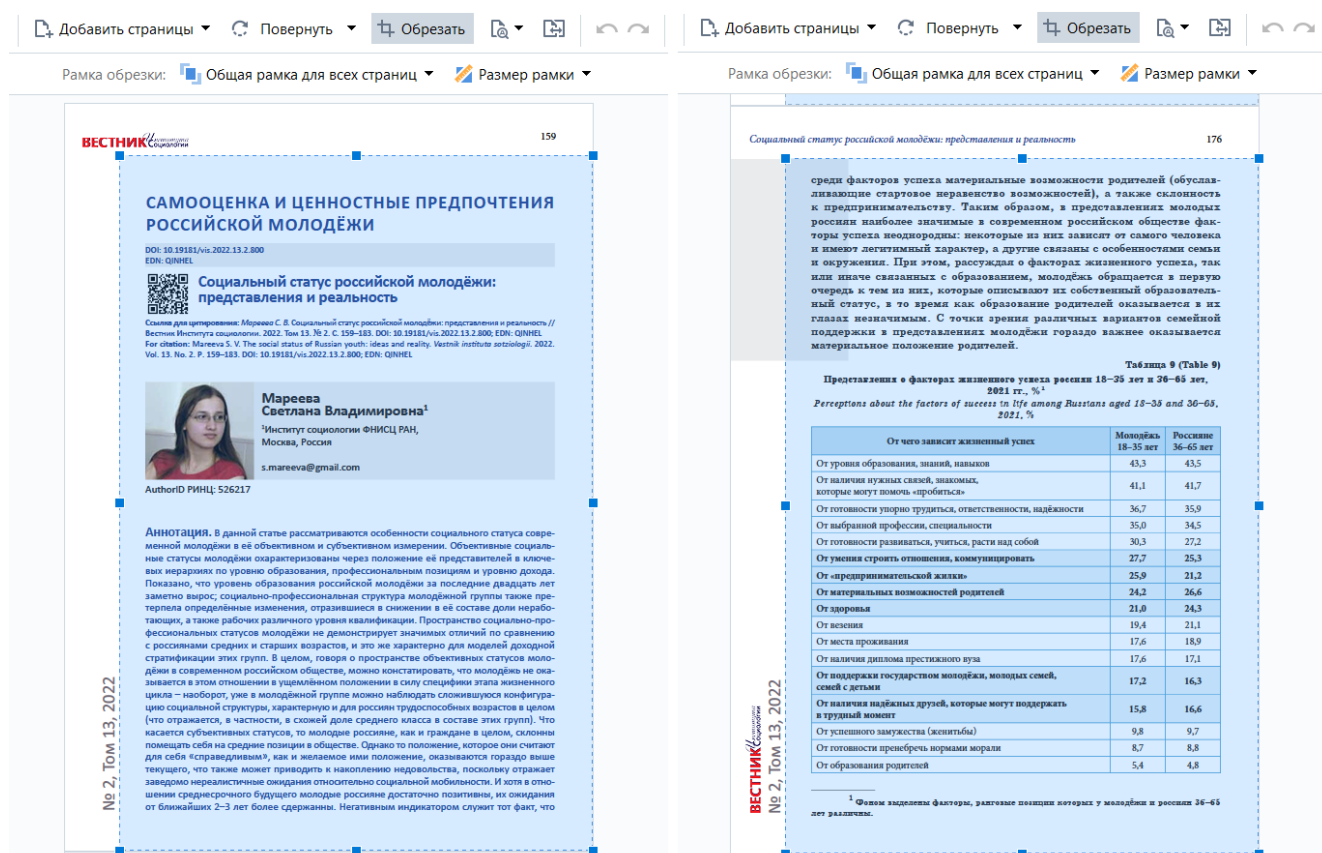


Рис. 7: Пример хорошо выбранной области и ее корректности для остальных страниц файла.

После чего необходимо нажать кнопку обрезать рядом с тем окошком, где был выбор для каких страниц применять обрезку. И нажимаем на крестик в этом же окошке, чтобы выйти из режима обрезки:

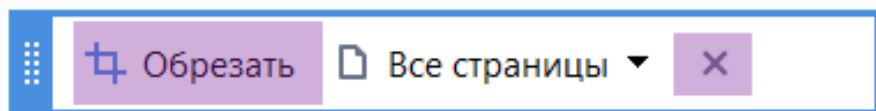


Рис. 8: Кнопка обрезать и закрыть инструмент.

2. На следующем этапе будет происходить удаление ненужных элементов при помощи инструмента «Стереть».

P.S. Может случаться так, что отображается не полный набор инструментов, для этого достаточно нажать на кнопку «Инструменты» в верхнем правом углу:



Рис. 9: Панель инструментов.

Ей необходимо выделять то, что нужно удалить: **графики и таблицы** (в том числе пояснения к ним), **ссылки и примечания внизу страниц** (поскольку при обрезке страниц, чтобы не потерять полезную информацию, где-то останется и то, что необходимо удалить), **фотографии и прочие элементы**, которые не касаются текстового содержания.

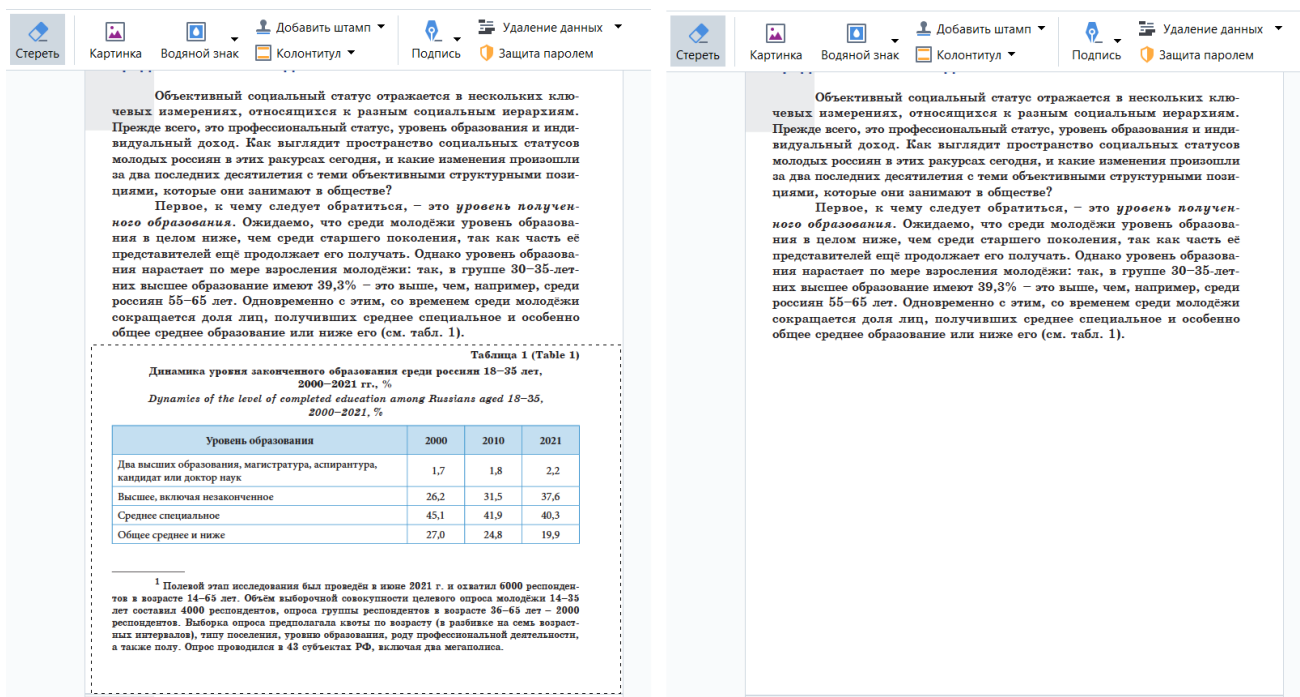


Рис. 10: Пример использования инструмента «Стереть».

Удалять необходимо вплоть до страницы с библиографическим списком, там все сделать можно чуть проще - просто удалив страницы. Объясню более подробно: когда мы дошли до страницы со списком литературы, мы должны удалить все, что находится ниже него **на данной** странице при помощи стирания:

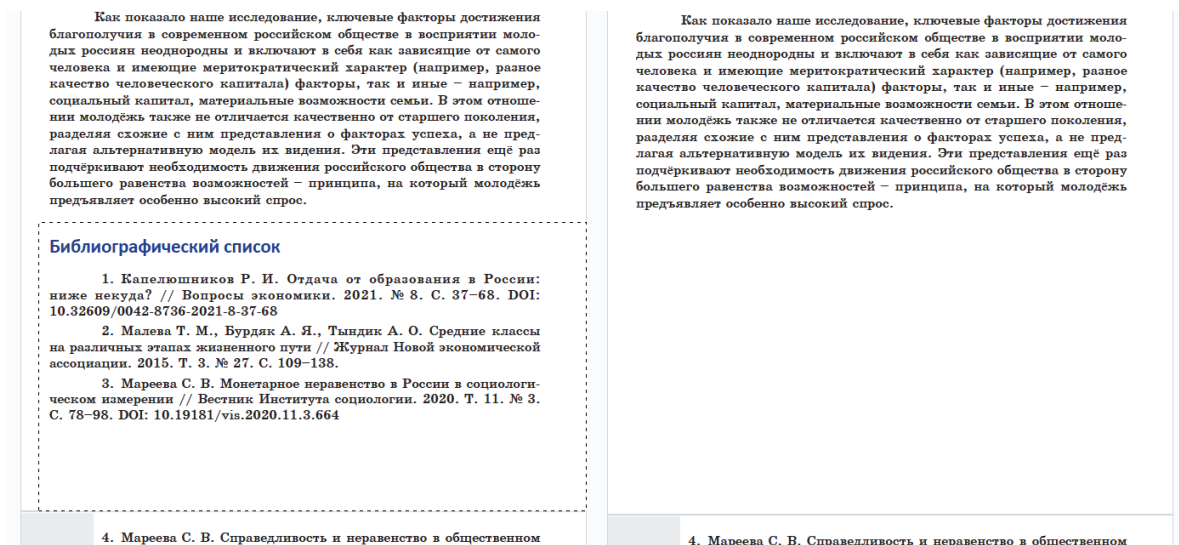


Рис. 11: Использование инструмента «Стереть» на странице со списком.

После чего нажимаем сверху слева на кнопку «Страницы»:

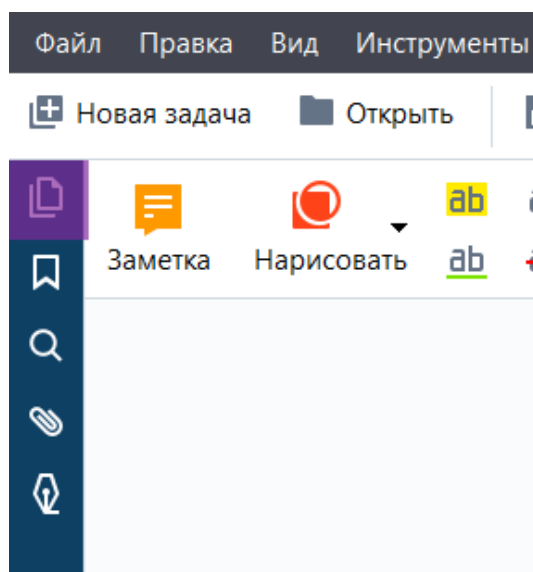


Рис. 12: Раздел «Страницы».

Скорее всего, будут выделены все страницы. Необходимо нажать на любую страницу (например, самую нижнюю), чтобы снять выделение:

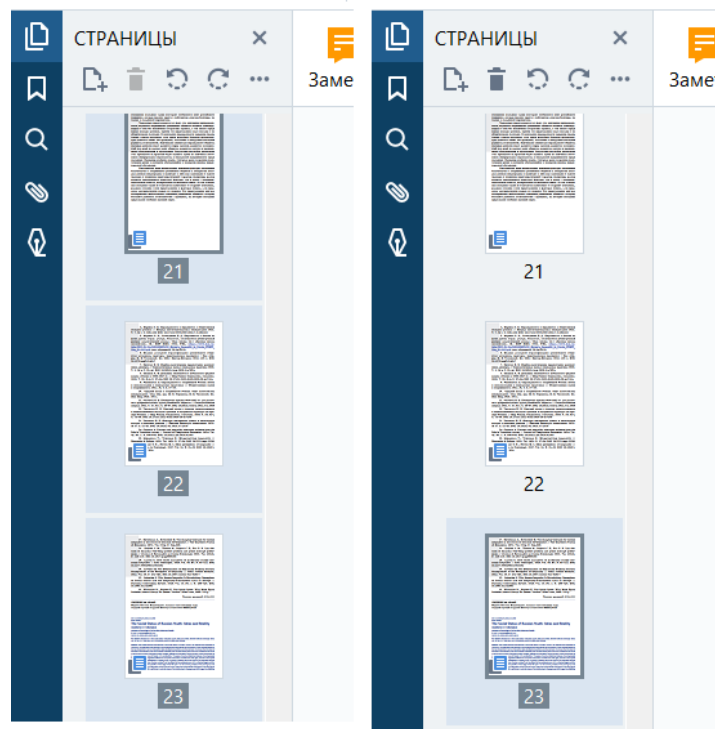


Рис. 13: Снятие выделения со всех страниц.

Нажимая «Ctrl»+«ЛКМ», выделяем те страницы, которые находятся ниже библиографического списка. После чего нажимаем правой кнопкой мыши по любой из выделенных страниц и нажимаем удалить:

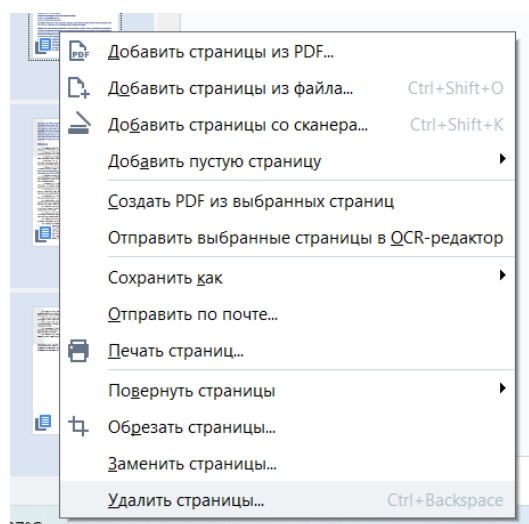


Рис. 14: Удаление ненужных страниц.

В результате чего мы получим почти (не удалены ссылки внутри текста в квадратных скобках) чистый pdf-файл. Теперь перейдем к следующему этапу

3. Теперь же начинается наиболее важный этап - перевод pdf-файла в txt-Файл.

Необходимо максимально внимательно выполнять данный этап, чтобы был получен наилучший результат за короткое время.

Для начала необходимо проверить, сходятся ли настройки. Нажимаем на верхней панели «Инструменты», а уже там вкладку «Настройки»:

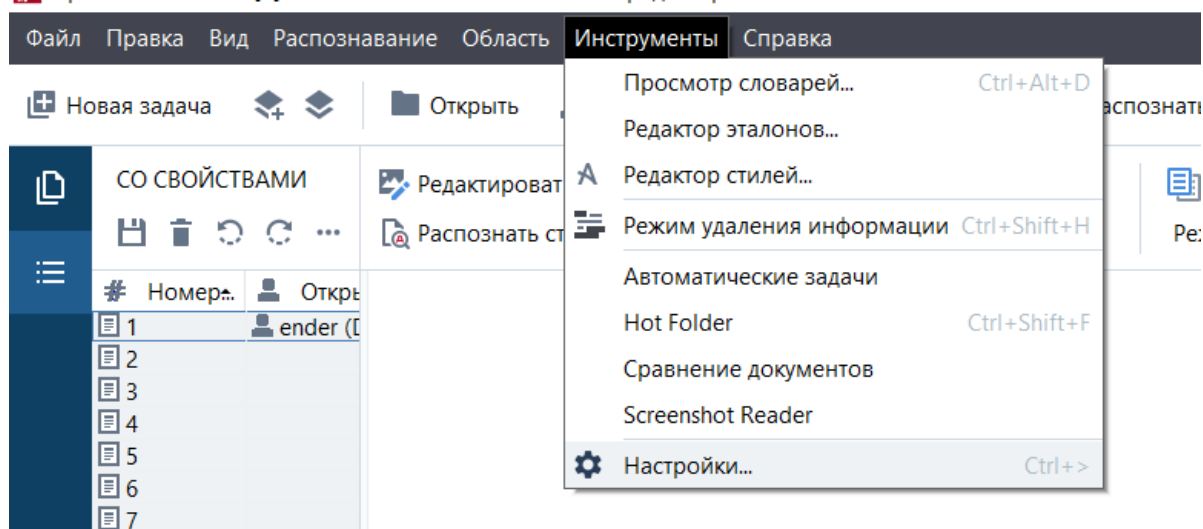


Рис. 15: Вкладка настроек.

После чего необходимо проверить, сходятся ли следующие настройки:

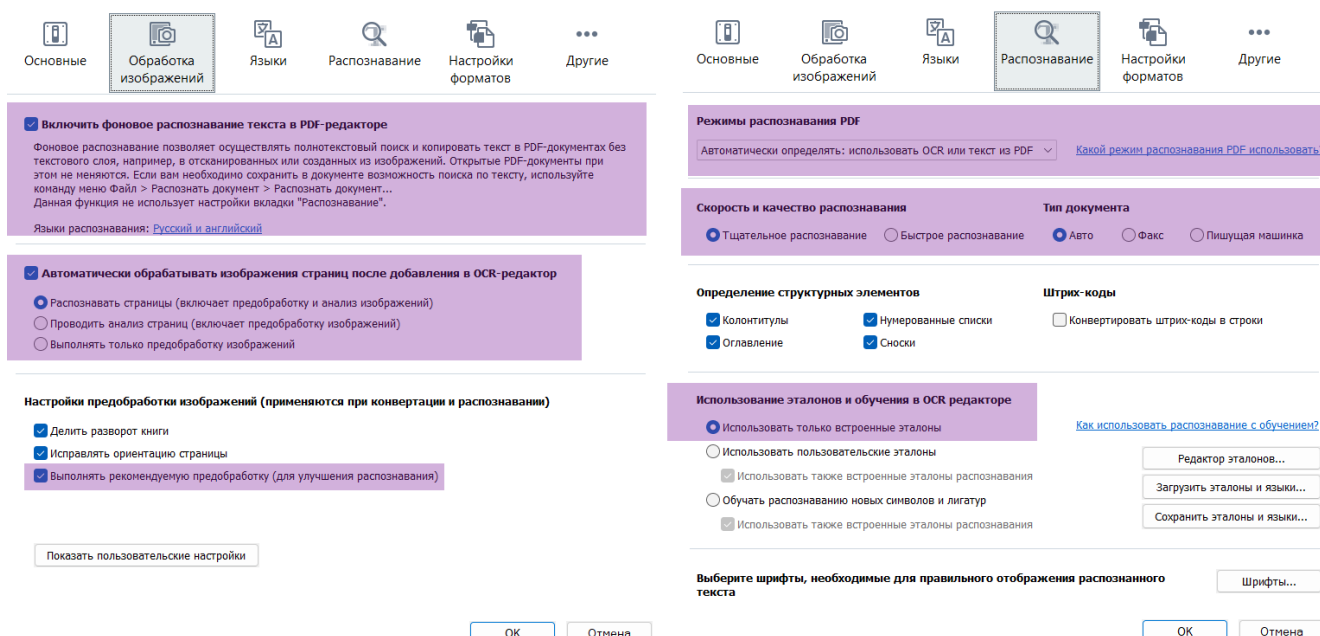


Рис. 16: Наиболее оптимальные настройки FineReader.

Далее закрываем вкладку настроек и нажимаем на кнопку «Распознать и проверить в OCR-редакторе»:

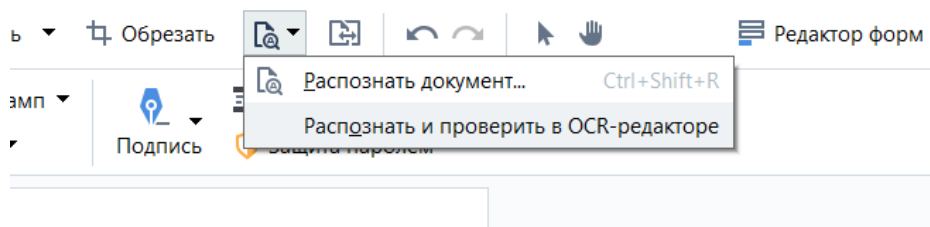


Рис. 17: Распознавание текста в OCR-редакторе.

Нужно набраться терпения, поскольку распознавание текста займет некоторое время. В результате будет запущен OCR-редактор, благодаря которому и будет извлечен текст.

В самом редакторе можно проверить правильность распознавания текста, но, все же, я рекомендую это делать в Word, поскольку более удобно визуально воспринимать, и большая часть букв, в распознавании которых сомневается OCR-редактор, оказываются верными. К тому же, там удобно исправлять орфографические ошибки (там есть словарь, хоть и не Ожигова, но есть) и Word нужно будет использовать для удаления ссылок в квадратных скобках.

Чтобы текст получился в наиболее удачном виде (без переносов слов, лишних символов и т.д.), в поле выбора режима обработки выбираем «Простой текст»:

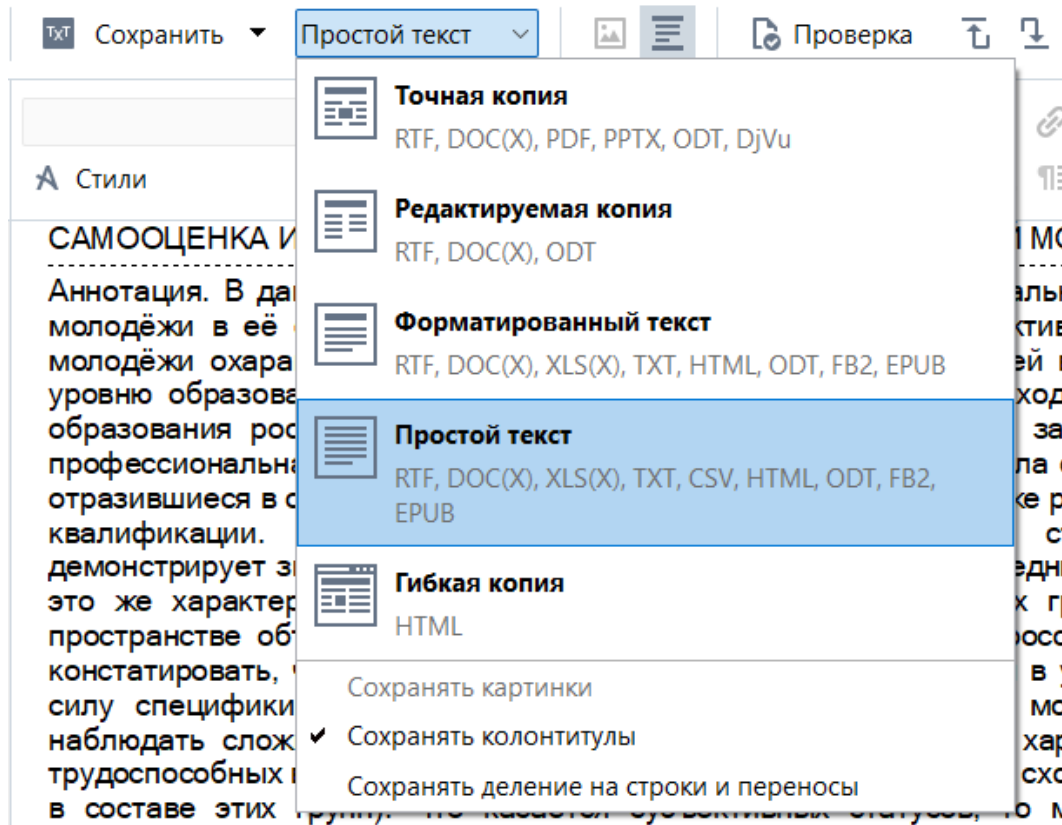


Рис. 18: Выбор режима распознавания.

Теперь же мы можем выгрузить наш проект в формате txt:

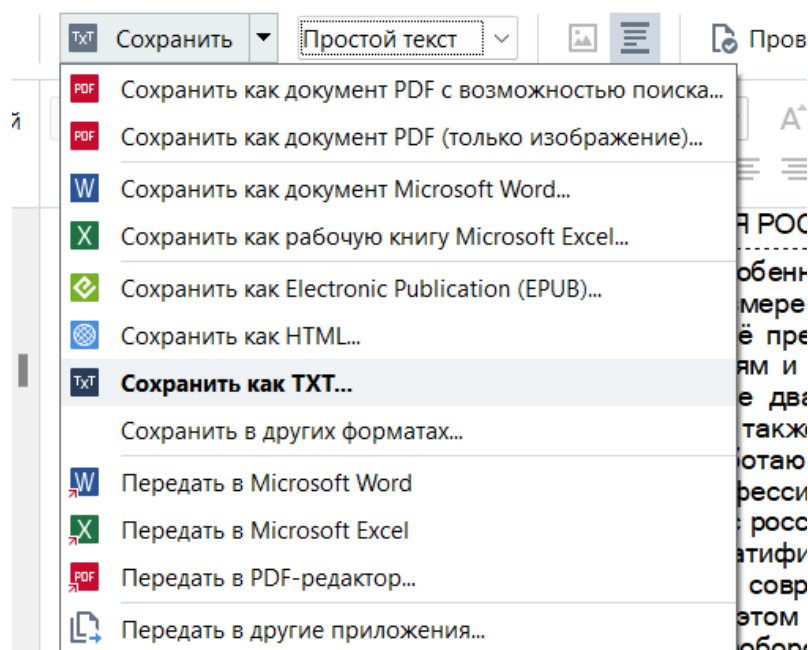


Рис. 19: Сохранение файла в формате txt.

После чего txt файл нужно будет назвать (пока имя неважно) и он откроется автоматически.

2.2.2 Предобработка текстовых данных

Теперь подготовим среду Word'a для работы - нужно отключить расставление переносов. Для начала создадим новый документ, после чего отключение переносов (в версии 2019 года) можно сделать двумя способами:

- Перейти в раздел «Макет», нажать на «Расстановку переносов» и выбрать «Нет»:

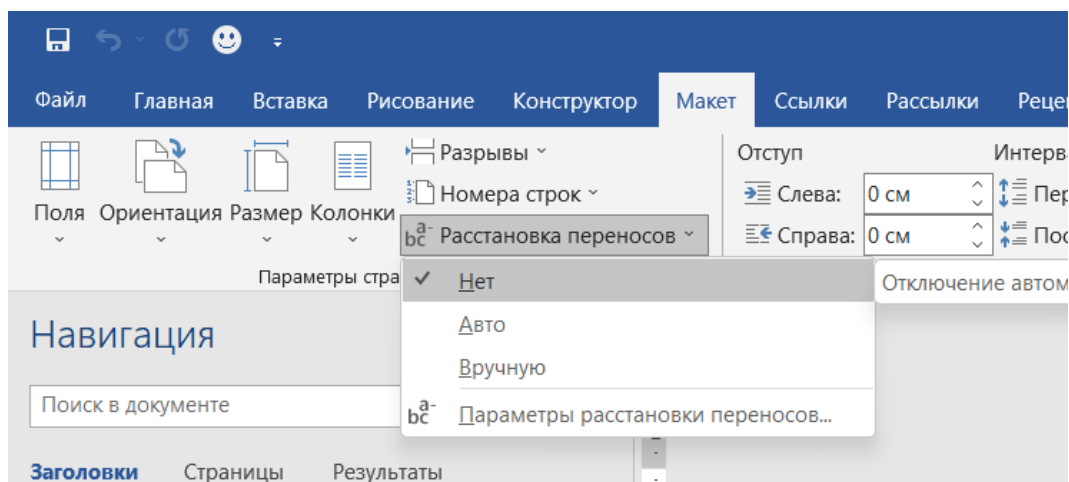


Рис. 20: Отключение переносов. Способ 1.

- Также есть вариант воспользоваться поиском в верхней части панели:

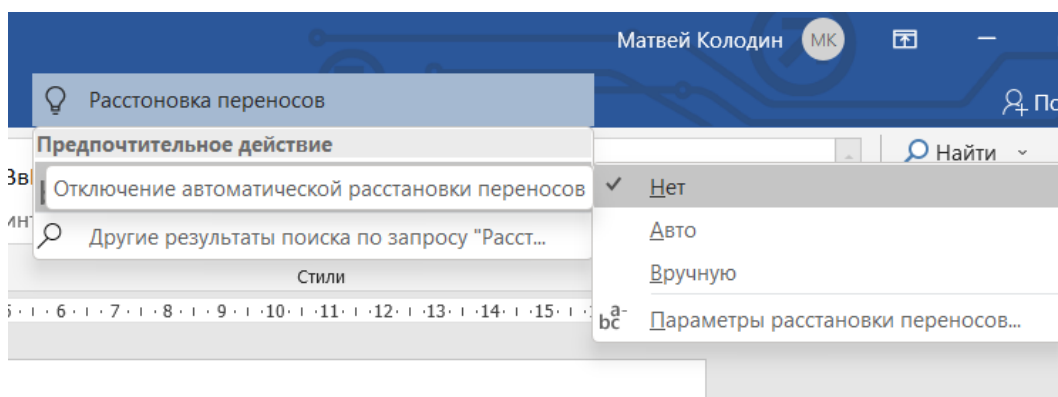
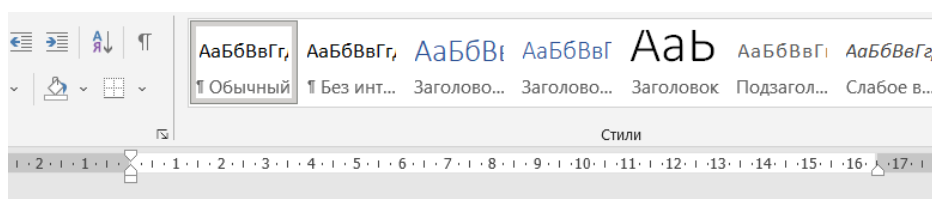


Рис. 21: Отключение переносов. Способ 2.

После чего в созданный Word-документ копируем содержимое из текстового файла.

Чтобы сделать это переходим в текстовый файл, нажимаем «Ctrl»+«A» (в результате чего должен быть выделен весь текст), после «Ctrl»+«C» (тем самым скопировав содержимое), и наконец в Word-документе «Ctrl»+«V» (чтобы вставить содержимое).

Результат должен выглядеть примерно так:



САМООЦЕНКА И ЦЕННОСТНЫЕ ПРЕДПОЧТЕНИЯ РОССИЙСКОЙ МОЛОДЁЖИ

Аннотация. В данной статье рассматриваются особенности социального статуса современной молодёжи в её объективном и субъективном измерении. Объективные социальные статусы молодёжи охарактеризованы через положение её представителей в ключевых иерархиях по уровню образования, профессиональным позициям и уровню дохода. Показано, что уровень образования российской молодёжи за последние двадцать лет заметно вырос; социально-профессиональная структура молодёжной группы также претерпела определённые изменения, отразившиеся в снижении в её составе доли неработающих, а также рабочих различного уровня квалификации. Пространство социально-профессиональных статусов молодёжи не демонстрирует значимых отличий по сравнению с россиянами средних и старших возрастов, и это же характерно для моделей доходной стратификации этих групп. В целом, говоря о пространстве объективных статусов молодёжи в современном российском обществе, можно констатировать, что молодёжь не оказывается в этом отношении в ущемлённом положении в силу специфики этапа жизненного цикла - наоборот, уже в молодёжной группе можно наблюдать сложившуюся конфигурацию социальной структуры, характерную и для россиян трудоспособных возрастов в целом (что отражается, в частности, в схожей доле среднего класса в составе этих групп). Что касается

Рис. 22: Полученный текст в word-документе.

Теперь проверим орфографию полученного текста - для этого в строке поиска вводим "Правописание в результате чего будет открыта вкладка, в которой можно будет проверять спорные моменты:

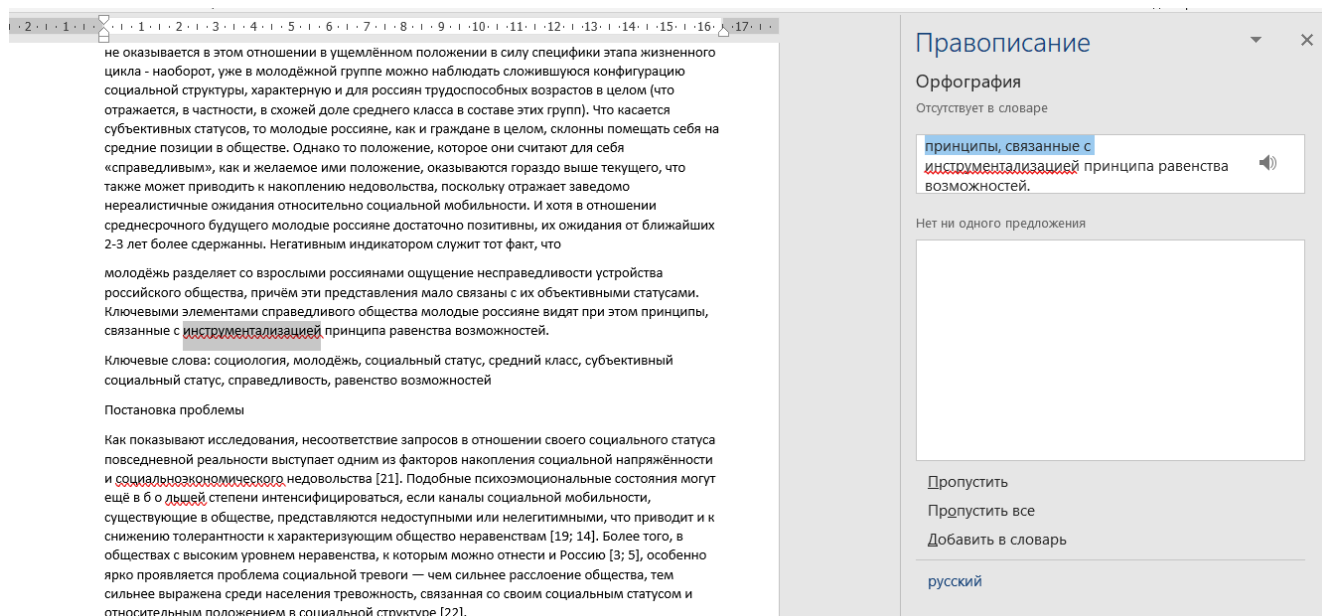


Рис. 23: Использование инструмента «Правописание».

Я крайне рекомендую также смотреть текст, может быть не вчитываться, но просматривать. Потому что при обработке бывают ошибки по типу «б о льшей», как на рисунке.

После проверки орфографии остается лишь заменить ссылки. Для этого нажимаем «Ctrl» + «F», в результате чего открывается окно навигации. Далее нажимаем на маленькую галочку (вниз), рядом со значком поиска и выбираем «Заменить...»:

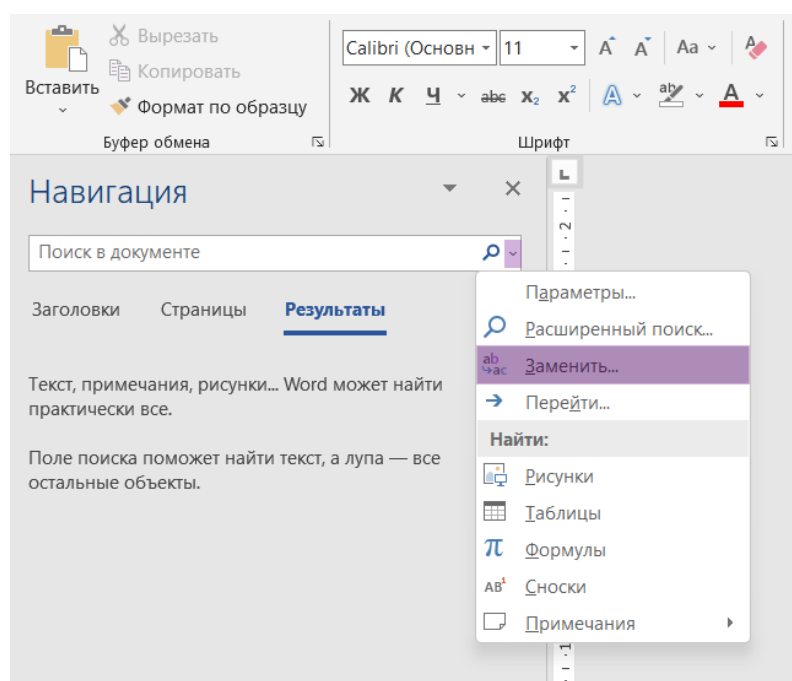


Рис. 24: Использование инструмента «Заменить...».

Далее нажимаем на кнопку «Больше», в результате чего будут доступны полные возможности инструмента:

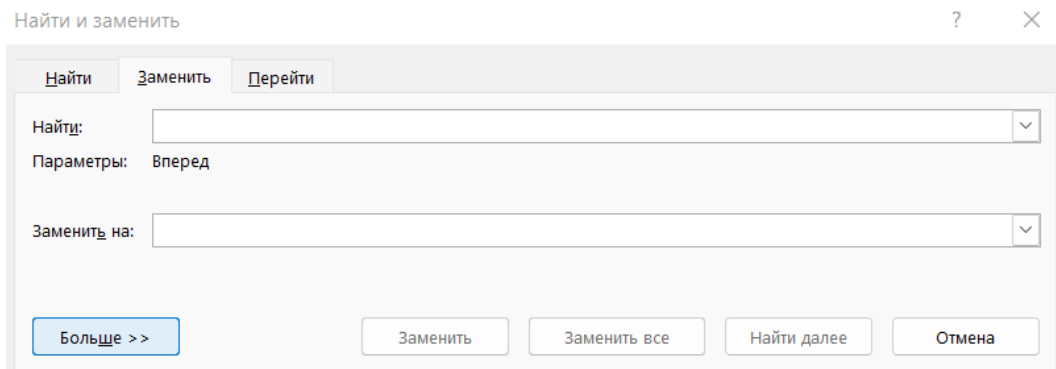


Рис. 25: Кнопка «Больше» в окне.

В развернутом списке ставим галочку напротив «Подстановочные знаки», после чего в поле «Найти» вводим [[]*[]] (эта маска позволяет найти весь текст заключенный в квадратные скобки). Поле «Заменить на » остается пустым.

После чего нажимаем кнопку «Заменить все». В результате чего будут выполнены замены, и получен документ с «чистым» текстом:

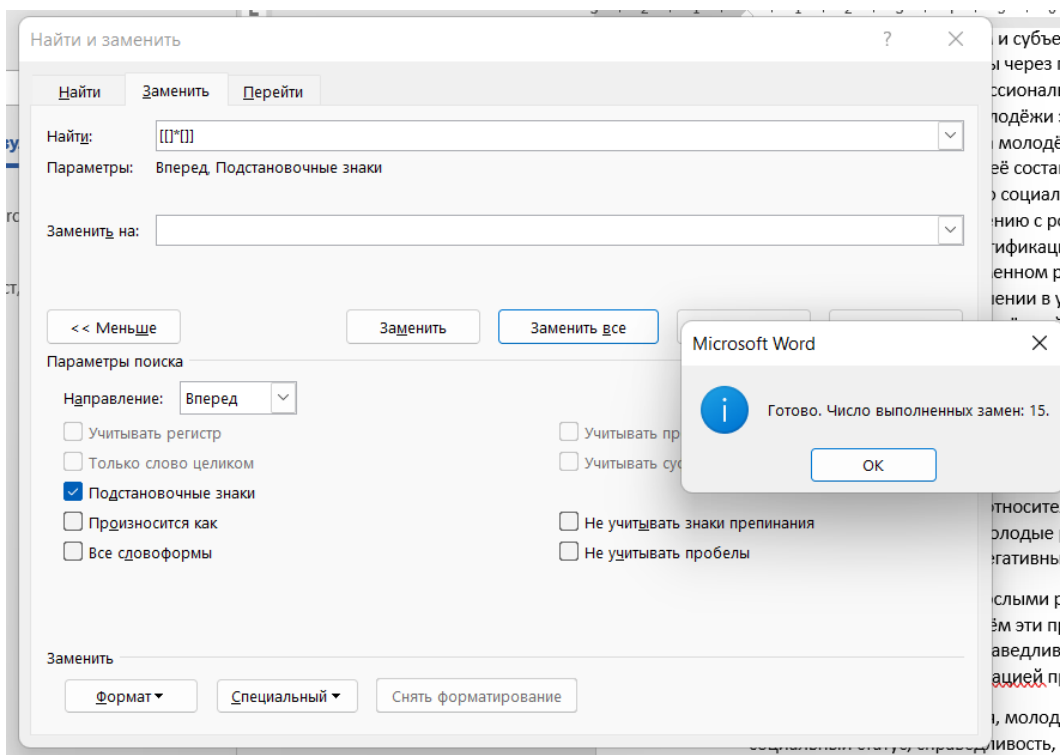


Рис. 26: Окончание удаления ссылок.

Теперь копируем весь текст документа («Ctrl»+«A» - для выделения, «Ctrl»+«C» - для копирования), после чего создаем новый txt-файл, в который копируем «чистый» текст («Ctrl»+«V» - для вставки).

Файл сохраняем под таким же названием, что и pdf-файл. После обработки всех pdf-файлов мы получаем готовую текстовую коллекцию:

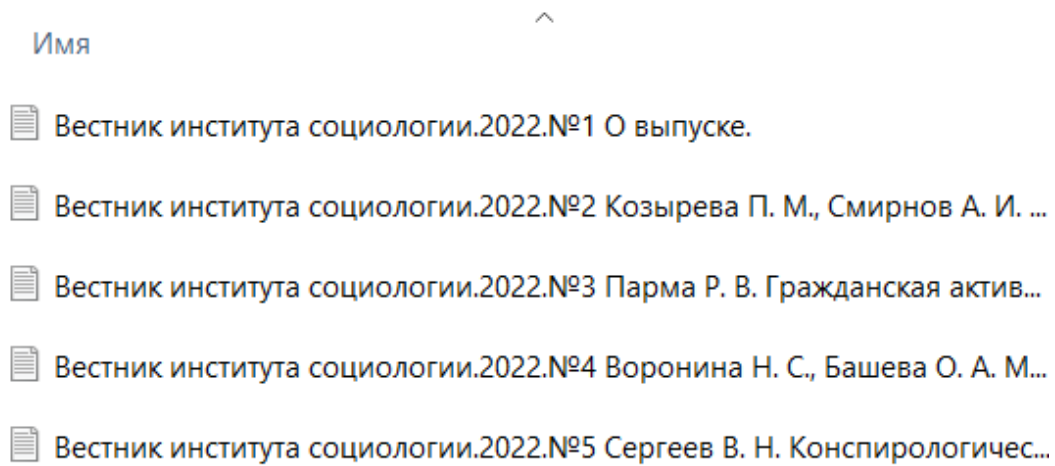


Рис. 27: Итог обработки текста.

Также необходимо создать копию получившийся коллекции, но без части «Название».

Это необходимо, так как, к сожалению, Google Colab не поддерживает обработку файлов со слишком длинными названиями. Таким образом, получаем копию коллекции:

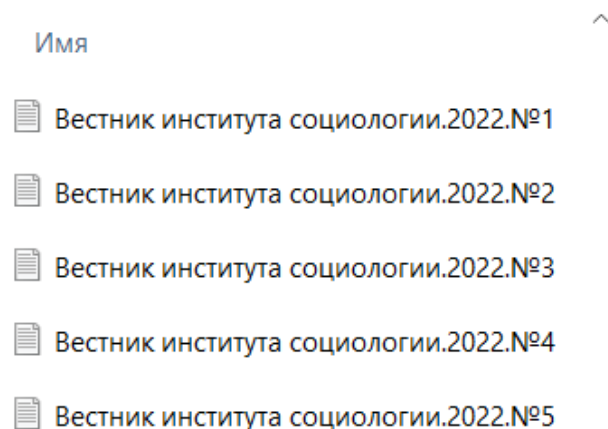


Рис. 28: Копия текстовой коллекции (с коротким именем).

Теперь наступает финальный этап решения задачи - лингвистическая обработка текста.

2.3 Преобразование текстовых коллекций в лингвистический набор данных.

Перед тем как начать работу с текстовой коллекцией, нужно создать Google-аккаунт (**существующий также подойдет!**). Далее переходим по ссылке: [Google Colab](#). Теперь авторизируемся, нажимая кнопку «Войти» в верхнем правом углу:

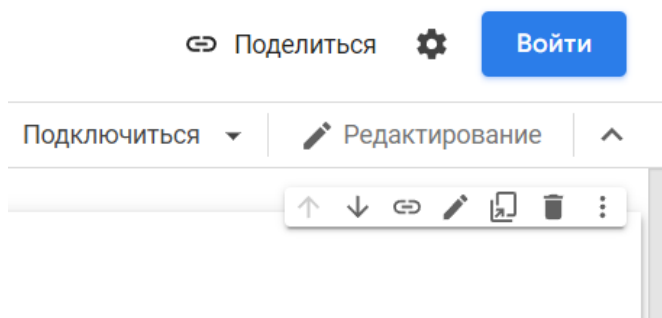


Рис. 29: Авторизация в Google Colab.

После окончания авторизации можно переходить по данной ссылке, в результате чего будет открыт блокнот с кодом:

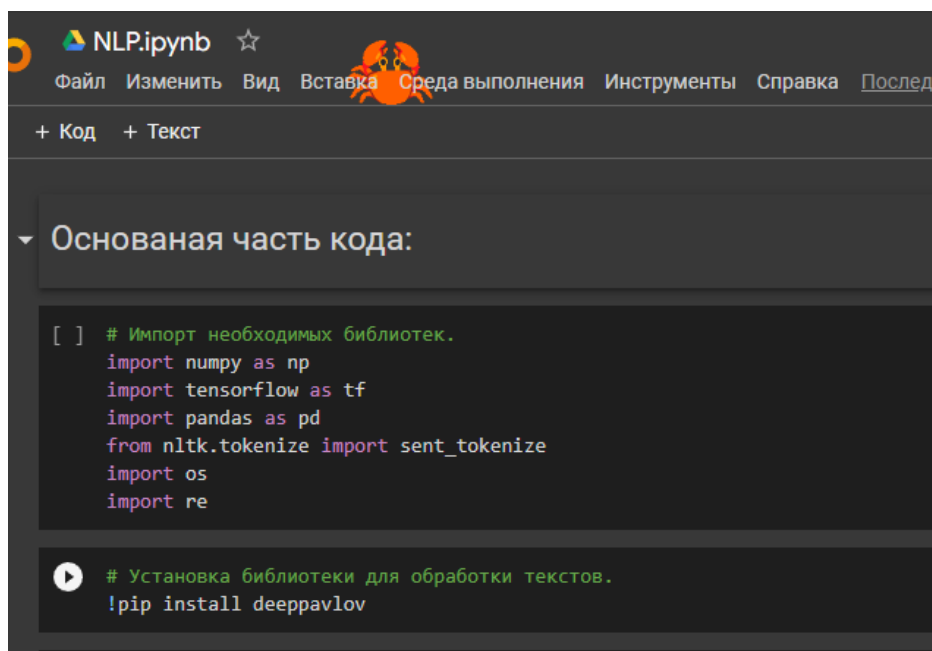


Рис. 30: Открытие кода в Google Colab.

Важно: поскольку Google Colab выделяет место на сервере для работы кода, то, если закрыть страницу на довольно долгий промежуток времени (думаю, что он составляет около 20 минут) сеанс будет окончен, и запускать все придется заново. Это не очень долго, но рекомендую не закрывать страницу или отключать ваш ноутбук/компьютер на продолжительное время. В случае чего - заново нажмите на ссылку. Но лучше этого избегать, чтобы не возникало ошибок!

Теперь мы начинаем взаимодействовать с ним! Нажимаем на кнопку «Подключиться» в верхнем правом углу и ждем, когда процесс закончится (обычно занимает около минуты).

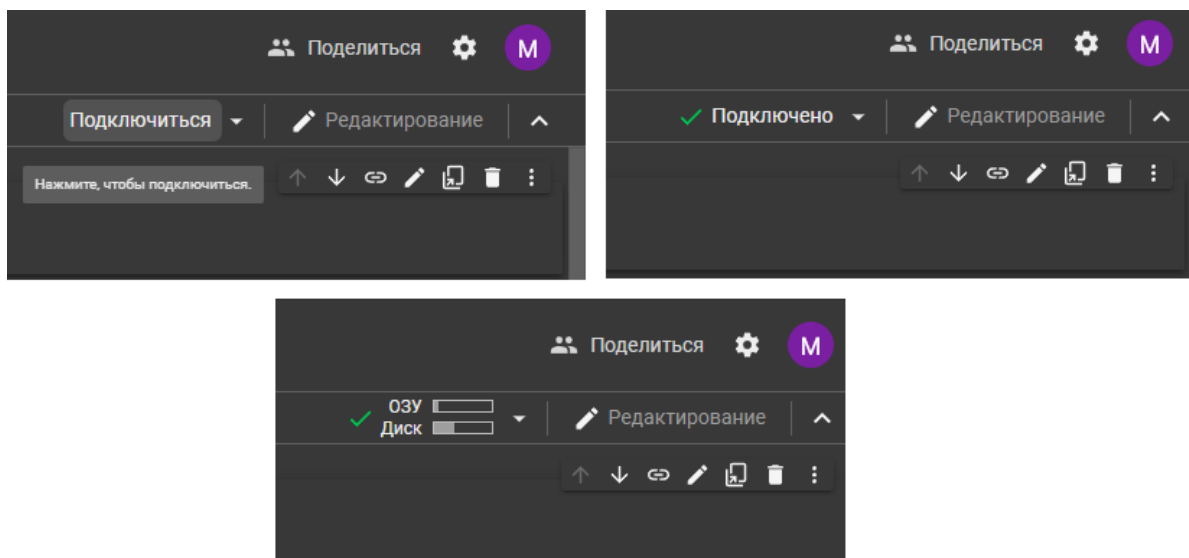


Рис. 31: Подключение к серверам.

Теперь мы можем запускать ячейки! Для этого достаточно навести на верхний левый угол ячейки и запустить код в ячейке:

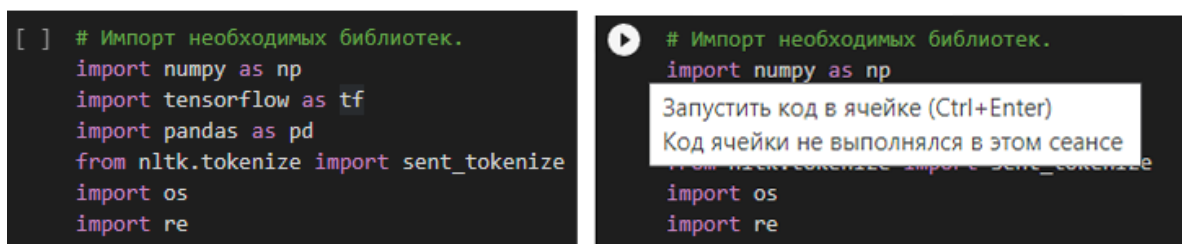


Рис. 32: Запуск кода в ячейках.

При первом запуске также будет нужно согласиться на выполнение кода, нажав кнопку «Выполнить»:

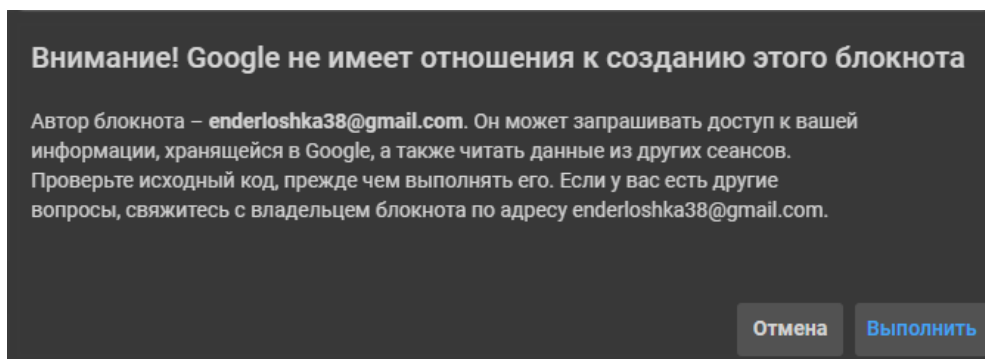
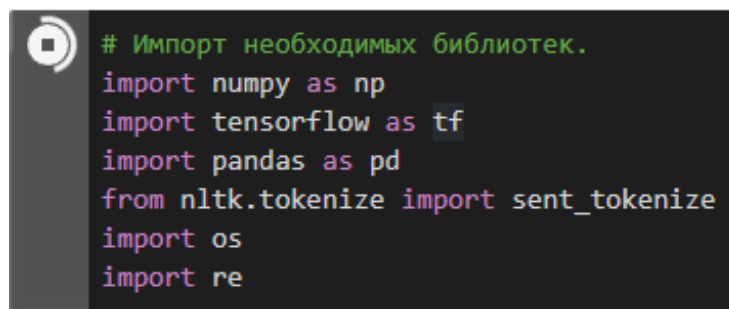


Рис. 33: Согласие на запуск.

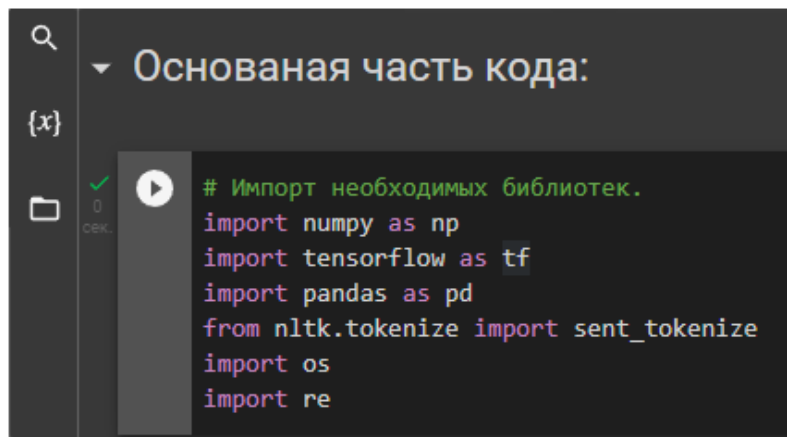
После чего код будет выполняться некоторое время. Процесс компиляции кода можно опознать по квадрату, находящемуся рядом с ячейкой:



```
# Импорт необходимых библиотек.  
import numpy as np  
import tensorflow as tf  
import pandas as pd  
from nltk.tokenize import sent_tokenize  
import os  
import re
```

Рис. 34: Процесс выполнения кода в ячейке.

Код в ячейке будет скомпилирован и успешно запущен, если рядом с ячейкой появится маленькая зеленая галочка:



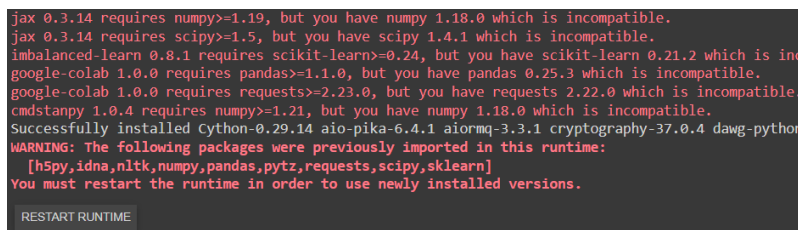
```
Основанная часть кода:  
# Импорт необходимых библиотек.  
import numpy as np  
import tensorflow as tf  
import pandas as pd  
from nltk.tokenize import sent_tokenize  
import os  
import re
```

Рис. 35: Окончание выполнения.

После того, как мы научились компилировать ячейку, начинаем выполнять весь код в строгом порядке:

1. Импортируем необходимые библиотеки (запускаем первую ячейку).
2. Устанавливаем библиотеки для обработки текстов.

Вторая ячейка будет довольно долго компилироваться (около минуты), и по окончании выдаст предупреждение. Переживать не нужно - просто перезапустите страницу!



```
jax 0.3.14 requires numpy>=1.19, but you have numpy 1.18.0 which is incompatible.  
jax 0.3.14 requires scipy>=1.5, but you have scipy 1.4.1 which is incompatible.  
imbalanced-learn 0.8.1 requires scikit-learn>=0.24, but you have scikit-learn 0.21.2 which is incompatible.  
google-colab 1.0.0 requires pandas>=1.1.0, but you have pandas 0.25.3 which is incompatible.  
google-colab 1.0.0 requires requests>=2.23.0, but you have requests 2.22.0 which is incompatible.  
cmdstanpy 1.0.4 requires numpy>=1.21, but you have numpy 1.18.0 which is incompatible.  
Successfully installed cython-0.29.14 aio-pika-6.4.1 aiormq-3.3.1 cryptography-37.0.4 dargpy-0.4  
WARNING: The following packages were previously imported in this runtime:  
[hspy,idna,nltk,numpy,pandas,pytz,requests,scipy,sklearn]  
You must restart the runtime in order to use newly installed versions.  
RESTART RUNTIME
```

Рис. 36: Предупреждение при установке DeepPavlov.

P.S. После перезапуска заново запускать ничего не нужно! Поскольку прошел короткий промежуток времени, то просто остается подождать, пока ваш компьютер снова подключится к серверу (как на рисунке 31).

Но, в случае чего, чтобы быть уверенными, подождите одну минуту, чтобы Google Colab снова проставил зеленые галочки рядом с ячейками. Это будет означать, что все хорошо, и они уже выполнены.

Также после перезагрузки сопровождающий текст (как на рисунке 36) к итогам компиляции может как удалиться, так и остаться.

В случае если он остался, я рекомендую для визуального комфорта нажать на ячейку (в любое место), после чего в правом верхнем ее углу нажать на 3 вертикальные точки, а там уже нажать на «Очистить выходные данные»:

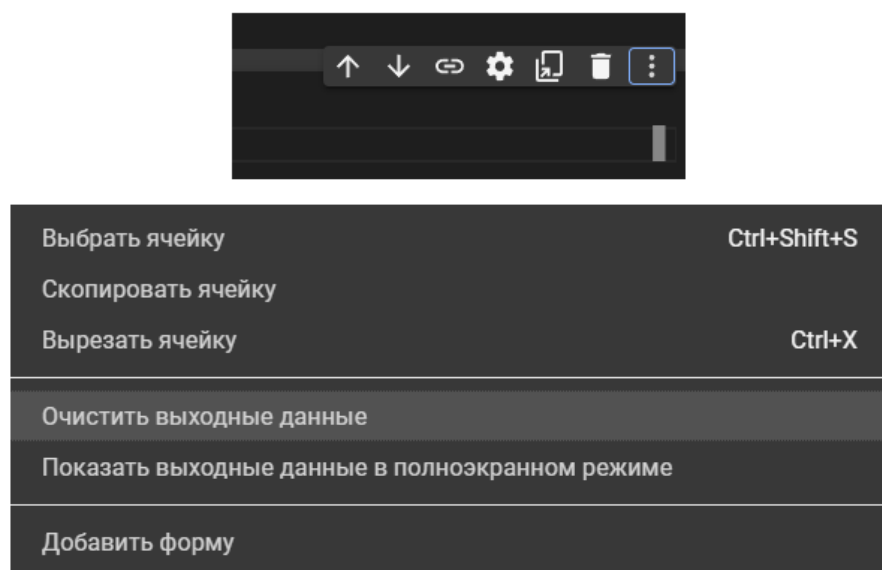


Рис. 37: Очистка выходных данных.

Также можно делать и дальше, для вашего удобства.

3. Скачиваем надстройки (третья ячейка).

После чего очищаем выходные данные и перезагружаем страницу.

4. Создаем папку с результатами.

5. Создаем папку для хранения текстовых данных.

После чего нажимаем на значок слева в виде папочки:

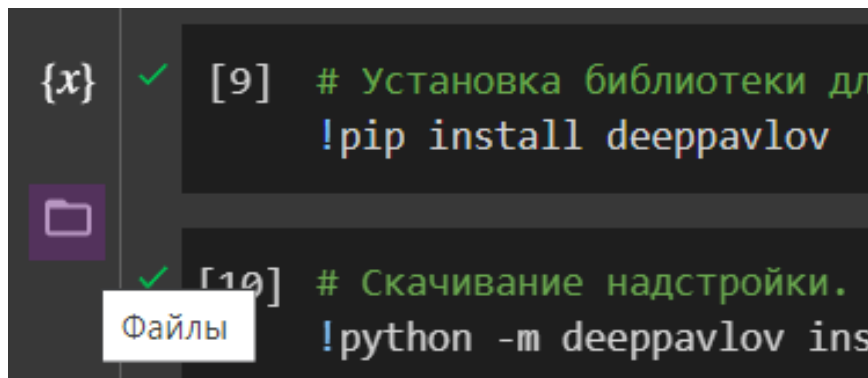


Рис. 38: Открытие файлов виртуальной машины.

После открытия файлов жмем на папочку с круговой стрелкой, чтобы обновить содержимое директории. В результате чего появятся две новые папки:

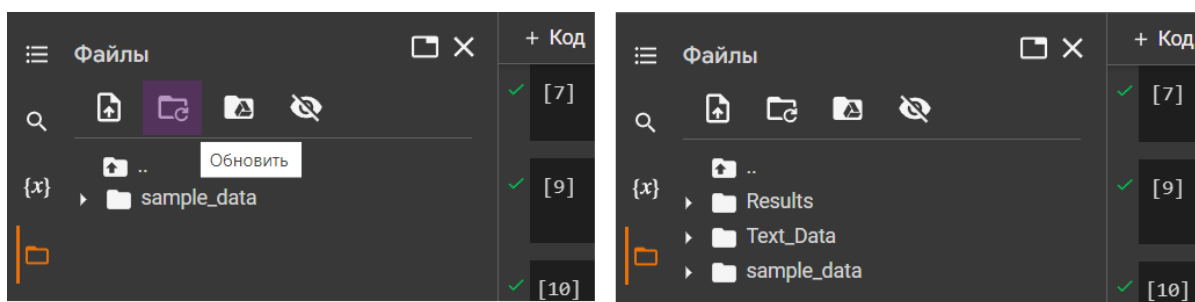


Рис. 39: Обновление содержимого папки.

Теперь наводим стрелочку на папку с именем «*Text_Data*», нажимаем на кнопку с тремя точками и выбираем «Загрузить»:

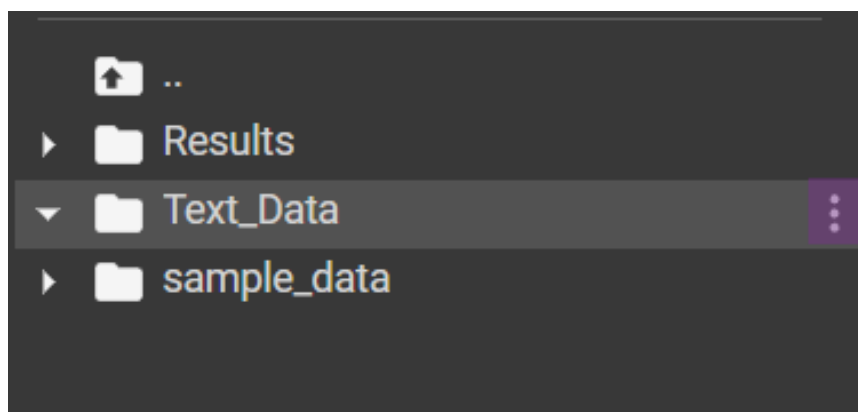


Рис. 40: Выбор файлов для загрузки в папку.

После чего нам нужно загрузить все файлы из готовой текстовой коллекции (с **укороченными названиями!**). Чтобы сделать это быстро, нужно зайти в папку, после чего нажать сочетание клавиш «Ctrl»+«A», в результате которого все файлы будут выделены.

Процесс загрузки должен выглядеть примерно так:

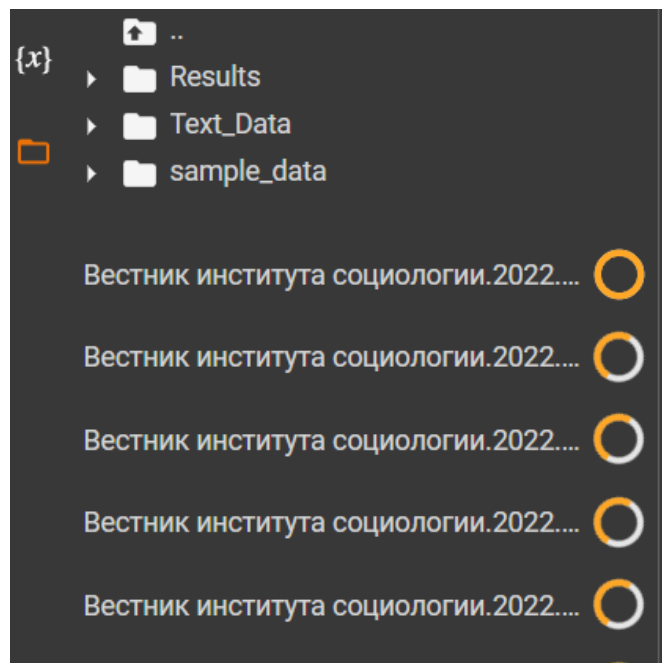


Рис. 41: Загрузка файлов.

Теперь нужно проверить, туда ли загрузились файлы. Нажмите на папку «*Text_Data*», после этого станет видно ее содержимое:

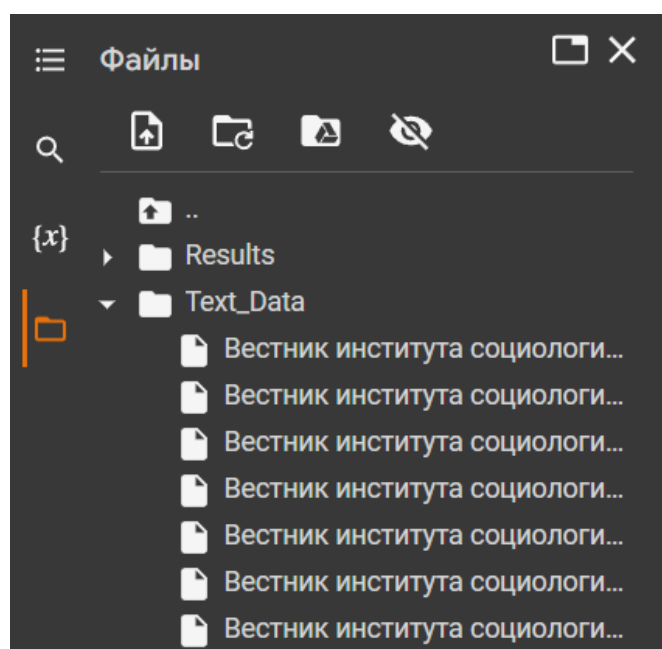


Рис. 42: Проверка содержимого.

После того, как мы убедились, что все файлы загружены, можно переходить к последней ячейке (**но не запускайте ее сразу, сначала прочитайте пояснение!!**)

6. Перед тем, как запустить выполнение ячейки, нужно изменить одну вещь в тексте кода. Комментарии к коду выделены зеленым цветом в самой ячейке-

ке. Необходимо пролистать до «**#Добавляем строку данных в meta-данные.**»:

```
# Добавляем строку данных в meta-данные.  
TTR = len(dict_lem) / len(dict_tkn)  
df_meta.iloc[FileIndex] = [FileIndex + 1, len(dict_tkn), len(dict_lem), len(tknz_sent), TTR, "Sociologia"]
```

Рис. 43: Правка при создании мета-данных.

Единственное, что вам нужно сделать - это изменить текст в кавычках. Вместо «"Sociologia"» написать ту тему, которой посвящены ваши текстовые файлы, например, если файлы посвящены истории : «"History"». То есть будет это выглядеть так:

```
# Добавляем строку данных в meta-данные.  
TTR = len(dict_lem) / len(dict_tkn)  
df_meta.iloc[FileIndex] = [FileIndex + 1, len(dict_tkn), len(dict_lem), len(tknz_sent), TTR, "history"]
```

Рис. 44: Правка при создании мета-данных. Изменение темы.

Теперь можно запускать ячейку! Время компиляции зависит от количества и размеров файлов. В моем случае (10 файлов, в среднем 25 страниц в pdf-файле) время компиляции составило 2 минуты.

7. Запускаем последнюю ячейку, при помощи которой будут загружены результаты.

Теперь остается создать итоговую папку, в которой будут храниться результаты для разных коллекций. Можно, например, дать ей имя «All_Results». Далее создать внутри папку (пусть ее имя - «1»), в которую будут сложены папки с **исходными pdf-файлами, обработанные текстовые коллекции** (в двух вариантах названий), и **папка с результатами** (скаченные результаты необходимо распаковать. Для этого надо зайти в папку «content», и оттуда скопировать папку «Results»).

В результате получим примерно следующее:

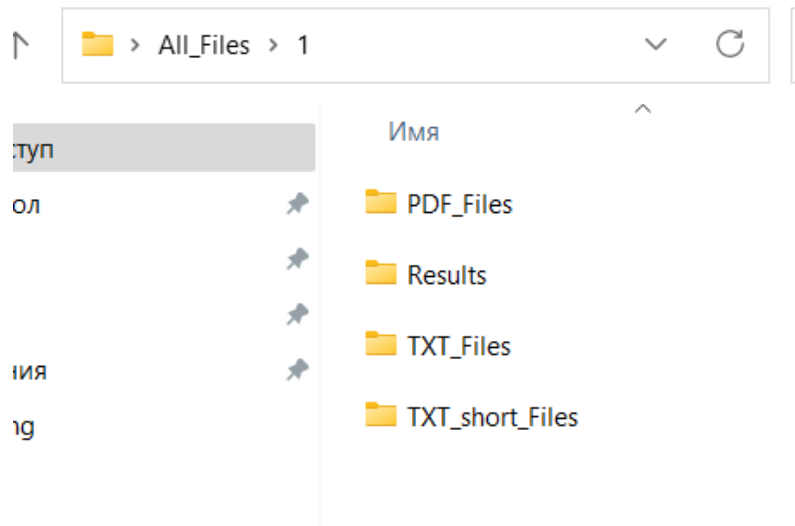


Рис. 45: Общий итог работы.

На этом решение задачи заканчивается. Но для последующей работы **крайне рекомендую** прочитать примечания!

2.4 Примечание

Прежде всего, примечаний будет два, каждое из них является важным для вашей комфортной работы:

1. Если необходимо обработать еще одну коллекцию, то во избежание ошибок рекомендую закончить сеанс с виртуальной машиной, и проделать компиляцию ячеек заново. Для этого сверху справа, рядом с надписью «Редактирование» (чуть левее) нажимаем на стрелочку и открываем «Управление сеансами», где и завершаем текущий:

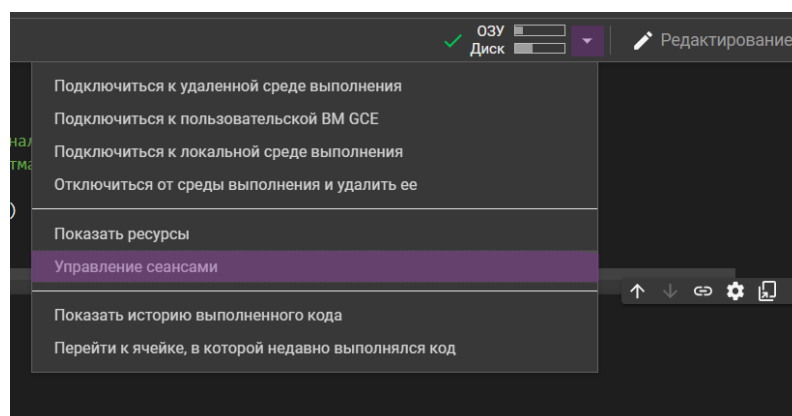


Рис. 46: Завершение сеанса.

После чего нужно будет снова подключиться (рисунок 31) и следовать той инструкции, что я написал.

2. В случае возникших вопросов писать в Telegram!