**Изменения в основном файле:**

1) заполнена колонка односторонняя/двусторонняя переписка: теперь даже если письмо одно, в этой колонке оно помечается, как односторонняя переписка. Если есть ответное письмо, то запись имеет следующий вид:

|  |  |  |
| --- | --- | --- |
| письмо | двустороннее | И. И. Иванову. (1). 1880. |

2) Убраны мелкие опечатки фамилии, которые были случайно обнаружены при работе, а также вычищены: двойные пробелы, буквы латиницей (были, наверно, в распознанном слое pdf). Поставлены точки там, где их не хватало, или наоборот – убраны лишние.

3) Все данные колонки «персоналии», имеющие дополнительный характер (сын, мать, великий князь), для облегчения обработки были перенесены в круглые скобки, например: Константину Константиновичу (вел. кн.). (1). Б. д.

Есть ячейки, где в круглые скобки вынесена вторая фамилия (псевдоним или девичья): Э. Ф. Тютчевой (Пфеффель). (6). 1870.

И есть ячейки с квадратными скобками, следующего вида: Редактору [А. Сомову]. (1). 1866.

В двух данных случаях приоритет отдавался авторскому именованию адресата – все то, что до скобок (в квадратных скобках по канону библиографической записи дается восстановленная информация – так записано и у Муратовой). Поэтому при обработке скриптом эти фамилии вместе с другими данными в скобках будут отсекаться в колонку «примечания». Возможно, такие случаи надо будет обрабатывать как-то отдельно, пока я не задумывалась над этим.

Во избежание путаницы были удалены скобки в ячейках следующего вида: [И. И. Иванову]. (1). 1800. – здесь автор по умолчанию не указал адресат письма, но он был восставлен библиографом.

4) Если какие данные неизвестны, то теперь вместо них ставится «?», а не просто пустые строки, например:

?. (670). ?. – неизвестен адресат и годы переписки.

**Скрипты:**

**Для облегчения отладки и понимания процесса писался не один код, а несколько маленьких. Нужно запускать последовательно. Можно указать входной и выходной файл в качестве аргументов запуска. Ниже информация о работе каждого из них.**

**0\_regex\_csv\_latin\_to\_ru.py** – для вычистки латинских букв, чистит, но не все и не везде. Не может полностью вычистить число «3», когда распознается вместо «З», букву латинскую «X» вместо русской «Х», а также строчные буквы. Поэтому такие случаи чистились руками.

**1\_regex\_personalii\_csv.py** – по регулярному выражению разделяет колонку персоналии на «адресат», «количество» писем, «даты».

**2\_regex\_csv\_split\_adresat.py** – делит колонку адресат на собственно «адресат» и «адресат\_примечание» (см. изменения в основном файле, пункт 3).

**3\_regex\_csv\_adr\_extract\_commadot.py** – делит сложные группы адресатов, написанные через «;», на 2 строки, например:

А. Г., Г. Н., К. Г., Н. Г., М. И. Белинским; А. П., Д. П., Е. П., П. П., Ф. С. Ивановым. (41). 1829—1835.

При этом данные предыдущих колонок копируются в строку ниже:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Выносит все в колонку «адресат\_тчкзпт».

**4\_regex\_csv\_adr\_extract\_commas.py** – разделяет сложные группы инициалов под одной фамилией на отдельные строки, например: А. Г., Г. Н., К. Г., Н. Г., М. И. Белинским (тот же самый пример, что выше). При этом данные предыдущих колонок копируются в строку ниже:

Изображение выглядит как стена, внутренний, электроника, стол

Автоматически созданное описание

Выносит все в колонку «адресат\_одиночн».

В идеале нужно бы еще ставить фамилию в мужской или женский род в зависимости от пола адресата, но это не всегда сразу известно и требует ручной обработки, поэтому пока «А. Г. Белинским».

**5\_regex\_csv\_adr\_extract\_io.py –** разделяет колонку «адресат\_одиночн» на инициалы (адресат\_ио) и все, что не инициалы (фамилия\_неделимо) – фамилии, сложные описательные адресаты (например, «Бывшим студентам Казанского университета»).

**6\_adr\_pymorphy.py –** работа с анализатором морф pymorphy2, приводит колонку «фамилия\_неделимо» в именительный падеж. Прописаны исключения.

**author\_split\_csv.py –** старый код, пока не чистила и не правила, должен был делить самую первую колонку «авторы» на фамилию, имя, отчество отдельно.

**Минус – в конце работы имеем очень много колонок, но лишнее всегда можно удалить. А так - хорошо видны ошибки, возникающие на каждом этапе.**