**Информация о команде**

KIPFA (или KIPIK, я подзабыл уже)

Учувствую я один

* Зайцев Никита Валерьевич
* 89160018648
* Telegram - @TheZaychik
* Email – nikitos945max@gmail.com

**Идея решения**

В итоге идея решения раскололась на две вариации:

Первая подразумевает прямой парсинг табличек с википедии, однако из-за разности табличек, с любой страницей это гарантированно работать не будет. В целом, те страницы, которые записаны в pages2.txt парсятся корректно

Вторая идея – это прямой парсинг по викидате, а точнее по position held. Однако данный способ более «ручной», так как вам нужно отдавать парсеру именно должность, а она на разные периоды страны может быть разная.

Так же обработка страниц ведется 8 восьми параллельных процессах. Их кол-во можно увеличить.

**Использованные сторонние библиотеки и инструменты**

BeautifulSoup4 – парсит сами html-ки

Qwikidata – интерфейс к API спаркла викидаты, просто упрощает жизнь (хотя можно по-хардкору поднимать хром с помощью селениума и парсить прямо из сессии браузера, но это долго и ресурсоемко)

Остальные библиотеки встроены в язык.

**Общее описание собранного датасета на данный момент**

На данный момент –2436 строк, однако возможно частичное дублирование данных. Во многих временных полях стоит N/A, это означает что данных по времени правления нет. СSV-разделитель – запятая.

Сбор происходил по странам:

Россия, Белоруссия, Великобритания, Германия, Золотая Орда, США, Япония, Испания, Италия, Франция, Рим, Кастилия.

**Описание текущего прототипа решения**

На данный момент пасрер умеет брать категорию викидаты, формировать спаркл запрос, получать и парсить списки правителей, а так же смотреть таблицы на страницах со списокм правителей, переходить по каждому из них, и брать ссылку на викидату.

Если говорить чуть глубже, то сначала берется список категорий, затем из каждой категории с помощью спаркла достается json со всеми правителями в этой категории, позже уже каждого правителя начинает обрабатывать основная парсинг-функция

Во втором случае берется таблица, и по строкам проходится на наличие ссылок, проверяется не является ли ссылка картинкой, и в таком случае совершается переход. Иногда парсер ошибается, но чаще всего на основных типовых таблицах все работет нормально

**Испытываемые сложности командой на данном этапе**

По большому счету есть одна – это ввести проверку на дублирующийся строки. Но искать подходящий алгортим (а не простой счетчик) уже времени нет.ы

**Планы развития решения**

В дальнейшем можно ввести проверку дублирующийся значений.