Встреча 25.11.2017

Мы обсудили разметку ошибок КРУТ и решили, что будем использовать их (с их разрешения) разметку ошибок. Возможно, этот корпус нам понадобится для обучения алгоритмов поиска ошибок в студенческих („kitten”) текстах.

Также у нас появился скрипт, который чистит от библиографий, оглавлений и оставляет чистый текст. Мы проверили его на одном из наших источников (учебнике по международным отношениям).

Мы обсудили метаразметку: у нас должны быть жанры (учебники, диссертации, статьи), темы (поддомены), авторы, год написания, источник (?), статус автора (?).

Мы составили план до субботы:

1) Александр определяет размер генеральной выборки, собирает статьи согласно «белому списку» ВШЭ.

2) Анна чистит тексты и определяет чистоту их.

3) Анастасия размечает.

4) Мария исследует выгруженный корпус киберленинки.