Встреча 15.03.2018

1) Саша представил результаты работы (для него неделя короткая - появилась возможность работать только с понедельника): готовые корпуса на основе киберленинки по 6 доменам (вскоре будут в папке на гугл-диске). После субботних тестов ворд2века нужно будет договориться о встрече с М.В. Соответственно, задача - работать с ворд2веком и (желательно) попробовать дженсимовский инструмент для сравнения сходства текста с корпусом.

2) Аня представила свои результаты по написанию скрипта https://github.com/MariaFjodorowa/catandthekittens/tree/develop/collocation_frequences и работе с коллокациями (от униграмм к шестиграммам). Задача: составить коллокационные списки по всему корпусу и поработать с мерами на нём же (хотя могут быть подводные камни в виде того, что меры (T-score) могут тащить много мусора на относительно маленьком объёме данных).

3) У Маши новостей нет. Задача: разметить весь корпус (маленький).

4) Настя заболела и не присутствовала.

Маша и Настя договариваются отдельно о своей встрече в паре с М.В. Копотевым.

Решили, что пятиграммы и шестиграммы оставляем.

Немного обсудили gensim и его возможности для сравнения текста с корпусом.

Обсудили курсовую: можно написать даже не одну статью - описание работы; а также методические работы (по выделению разных вещей), в которых можно описать методику работы с разными инструментами. Решили формат, скоро будет список и план. За черновой план берем наш абстракт для конференции в Праге. Наверное, делаем одну общую большую работу, но каждая “глава” будет в формате статьи, которые мы следом переделываем в доступный для публикации формат.

Обсудили возможность подключить ГИКРЯ, связываемся с Катинской А.Ю. (?)

Обсудили предложение корпуса ошибочных текстов, которые можно будет использовать в тестированиях (по 10 текстов от профессоров по нашим доменам). Следует их анонимизировать.