Встреча 01.03.18

Отчитались по выполненному.

Аня: Сделала частотные списки биграммов и триграммов для своих доменов. Задача остаётся та же, необходимо добавить частотные списки с учётом pos-тегов.
Маша: Ru-syntax удалось запустить из-под линукса, выложила на гитхаб код, который напускает его сразу на целую директорию с текстами. Сейчас занимаюсь собственно напусканием. Пока только по своим доменам, остальных текстов у меня пока неттак выглядит размеченная статья. Сейчас на обработку 15 примерно таких статей уходит около 10 минут. Обработка одной статьи занимает до 40 секунд только в мальтпарсере, + какое-то время на TreeTagger и mystem, + какое-то время на запуск ru-syntax`a. Единственный вариант оптимизации, который мне видится - запоминать длину каждой статьи в токенах, склеивать их в один док, запускать ru-syntax на нем, а потом делить выдачу. Причем наша токенизация должна быть точно такой же, как у ru-syntaxa. И это не спасение.Также может быть быстрее, если делать это не в виртуальной машине.
Настя: Поработала с учебником Сидоровой и Савельева, выписала с опорой на них характерные особенности академического текста и сделала предположения как мы можем их выделять. Пока что грубо можно выделить три группы:

1. Структура, если мы выделяем структуру документа - несоответствие “котенка” эталонному “коту” одного с ним жанра. Или, например, отсутствие во введении какого-то из обязательных элементов целей, задач, актуальности и т.п. (Эта задача для поздней работы, хотя находить ошибки такого типа было бы интересно, пока мы с этим не работаем).
2. Закономерности на уровне синтаксиса.
3. Закономерности на уровне лексики, словообразования. Сформулировать для Ани задачи (например, подсчет глаголов настоящего времени, которых в академическом тексте должно быть больше, чем других). Обсудить с Аней, потом показать Михаилу Вячеславовичу

Саша: Сломался компьютер :(Но Саша понимает, что делать дальше :)

Организационное:
На следующей неделе, возможно, не будет общего созвона, т.к. Михаил Вячеславович уезжает на школу.