Что делаем:

Cat and kittens: корпус русских академических текстов (CAT) и инструментарий, позволяющий сравнивать студенческие тексты (kittens) с эталонными из корпуса.

Задачи:

1) **Сбор корпуса:**
   Из чего собираем: журналы из белого списка Вышки, материалы крупных конференций (Диалог), диссертации (магистерские и выше), учебники и монографии, отобранные соседним проектом (РКИ).
   Тематики:
   1. Социология и история;
   2. Политология, международные отношения,
   3. Юриспруденция;
   4. Лингвистика и прикладная лингвистика;
   5. Экономика;
   6. Психология и педагогика.

   Размер корпуса: объём каждого домена определяется генеральной совокупностью, но не должен быть менее 300000 токенов.
   Как производится сбор и обработка: тексты краулятся с сайтов-каталогов научных работ (например, cyberleninka) и очищаются при помощи алгоритма, оставляющего только текст статьи и метаданные.
   Состав метаданных: жанр, домен, автор, год написания, источник.
2) **Разметка корпуса:**
   Состав разметки: частеречная (MyStem), синтаксическая (RuSyntax).
   Общий формат – xml/txt.
3) **Реализация алгоритма** сравнения студенческих текстов с эталонным корпусом (будет сделано в третьем-четвертом модуле).
   - **Параметры сравнения:**
     - Сравнение с реф. корпусом:
       - Повторяющиеся леммы (топ-частот по частям речи, отличающиеся от распределения по большому корпусу)
       - морф. и синт. ошибки: (несуществ. формы “столи”, ошибки согласования “большая стол”).
     - Сравнение с тематическим корпусом:
       - Неспецифически научные слова и коллокации (топ-частот по частями речи и топ-частот биграмм и

триграмм, “гапаксы” (слова уникальные для проверяемого текста)
        - термины
- **Результаты:**
    - Часть результатов относится ко всему тексту: TTR, близость к домену X. (таблица в шапке анализа)
    - Большая часть результатов подсвечивается в тексте:
        - С вариантами замены:

— Нестандартная синт. связь “большая стол”: вариант стандартной замены

— С вариантами замены: нестандартное слово — синонимы по word2vec’у — коллокации по LL для этих слов.

        - Без вариантов замены:

— “Этот слово редко встречается в текстах на эту тему. Убедитесь, что оно употреблено правильно.”

- **Тестирование**: оценка асессорами. Можно дать 10 грязных текстов (~500 слов) асессорам (3-5), они размечают ошибки и затем мы сравниваем с результатом нашей программы. Из результатов асессоров нужно оставлять не 1 вариант, а несколько. Одна из мер качеств: точность - ложный alarm/точный; полнота - на все ли выделенные реакции человека есть реакция программы. (четвёртый модуль)
- Сделать веб-интерфейс для взаимодействия с корпусом (первый модуль второго курса).

Задачи на ближайшее время:

1) Разметка корпусов: каждый из участников работает с несколькими доменами, собирает корпус каждого из них и организует разметку
Сбор корпусов: выполнено, достигнут базовый уровень в триста тысяч токенов для каждого из доменов.
Разметка: будет закончено в середине-конце января, выполняется каждым из участников для своих доменов.

Следующие задачи должны быть выполнены к концу февраля-началу марта:

2) Частотность: составление частотных списков по частям речи (Аня)
3) Применение word2vec для определения тематики текста и поиска “синонимов” (Саша и Маша)
4) Выявление типичных ошибок (синтаксис прежде всего). Формализация поиска типичных ошибок (сравн. констр., субъект деепричастной клаузы и т.д.). (Настя)