22.03.2018

Настя рассказала о своей работе: был составлен список типичных слов и словосочетаний (с опорой на учебники по функциональной стилистике). Кроме того, существует список признаков академического текста, которые нужно проверить на русинтаксе. Задача: проверить каждый признак на истинность и соответствие нашему корпусу. То есть, экспериментально доказать, правда ли это признак Академического стиля и потом с опорой на это выявлять.

Аня рассказала о законченной работе по составлению списков коллокаций и различных мер.

Можно составить общий список всех встречающихся коллокаций, если они встречаются чаще n раз, MI больше, то мы можем оценивать появление её. Нам нужен рейндж/разброс. Сделать ранжирование? Насколько вероятно, что один из доменов даёт всплеск при сливании их в подсчете коллокаций?

Работа с коллокациями. Смотрим на текст, ищем подозрительные коллокации, предлагаем замену с помощью машинного обучения.

Задача Насте: Общую информацию о тексте: ваш текст состоит из такого-то количества слов, TTR такой-то, TTR по домену такой-то, ридабилити по домену такой-то а у вашего текста такой-то. Для этого мы используем corrected TTR, чтобы не зависеть от длины текста (для этого меряются для тысяч слов отдельно). Есть готовые формулы.

Маша: сделана синтаксическая разметка. Немного поговорили про базу данных (индексирование, параметры оптимизации).

Саша: рассказал о своих word2vec. Нужно сделать помимо сравнения текстов еще модель со словами для генерации замен. Нужно посмотреть о влиянии границ предложений на ворд2век модели.