Встреча 10.01.2018

1)Частотность: частотные списки по частям речи; читать про коллокационные меры: LL для частей речи, сравнение с референц-корпусом, TTR (**Аня)**
2)word2vec для определения тематики текста и поиска “синонимов” (**Саша)**
3)“Типичные” ошибки (синтаксис прежде всего). Формализация поиска типичных ошибок (сравн. констр., субъект деепричастной клаузы и т.д.). (**Настя**)
4) Синтаксис: (**Маша**)
Sintagrus
rusyntaxs Можно на гитхабе или в установить на десктоп. Туда загружается текст и на выходе получается нлп(?) файл, в котором есть морфологическая, синтаксическая разметки.
Синтаксису обучали по майндпарсеру.
Universal Dependensies.
UDpy - критикуют его сегментацию, но он умеет из текста сделать qll с морфологической и синтаксической разметкой.

Несколько ключевых моментов, которые не позволяют точно их сравнить. Например, синтаксических лейблов в UD - 37, а в Синтаксисе меньше.
Syntagrus соединяет все через союзы и предлоги, а в UD все идет от глаголов.
В UD есть пунктуация, которая работает как тоже токены. В синтагрусе она живет внутри разметки.
В синтагрусе - моды, связанные с элипсисом, это пустые токены, типа нулевой связки, допустим, у нас какой-нибудь гэпинг и глагол во второй клаузе отсутствует. В UD раньше вообще строго запрещали пустые токены, а сейчас существует базовая репрезентация, в которой вообще никаких токенов быть не может.

Интуитивное ощущение, что смотришь - и ошибок мало?
В UD из теории графов - это такие плотные звезды с несколькими узлами, а в синтагрусе более крупные звезды, в которых больше хабов и которые не такие плотные. С точки зрения структуры это очень разные структуры.

Руссинтагс - меняет исходный файл? Что-то выкидывает при предобработке? Бывает расхождение между изначальным текстом посимвольное.
Стенфордпарсер - это не решение из коробки. Это так же, как майндпарсер.

+ дополнительная задача - взять из каждого домена по 10 предложений, сделать небольшой корпус, обработать кириной моделью, сделать визуалку, посмотреть на деревья, оценить, что получается

Мы двигаемся по еженедельному плану, чтобы отчитываться каждую неделю.

Корпуса и поддомены, которые мы создадим, будут не для распространения, но наши и мы решим, что с ними будем делать.

В ближайшее время мы должны вписаться в гуглофайл кто за какую часть берется и сделаем первые шаги.