Cold start.

- Задача: из полного набора предложений отобрать только те, где могут встречаться названия организаций.

- Как это делаем:
  1. С помощью парсера natasha (http://natasha.readthedocs.io/ru/latest/) размечаем все предложения:
     - «1» = в предложении встречается название организации
     - «0» = в предложении не встречается название организации
  2. В размеченных предложениях удаляем все ключевые слова типа «организация», «компания», «ООО», «ОАО», кавычки и т.д.
     - Не удалять сами названия компаний
     - Попробовать их тоже удалить
  3. На полученном датасете (у предложения есть лейбл «1» или «0», и удалены все ключевые слова) обучаем нейросеть. Так она научится определять наличие организации в предложении только по контексту.
  4. Подаем на обученную нейросеть исходные датасеты (без всяких удалений и неразмеченный). Для каждого предложения из этих датасетов нейросеть определит, есть ли в нем упоминание организации. Причем она сделает это исключительно по контексту.
  5. Для Толоки из каждого датасета выберем топ 200 предложений по скору нейросети.