Всем добрый вечер. Презентацию видно? Я готов начинать.
Меня зовут Мукасеев Евгений и сегодня я хочу вам рассказать про мой проект, который я назвал «Тему, пожалуйста». **СЛАЙД**
Немного предисловия. Я работаю в финансах и мы с коллегами очень часто готовим аналитические отчеты на разные темы. Иногда темы спускаются руководством, а иногда тему надо придумать самому. И это совсем не простая задача. Я захотел эту задачу как-то упростить и автоматизировать. Поэтому конечная точка моего проекта - это модель, которая будет предлагать тему для более глубокого исследования с презентацией руководству. Но как это сделать? Было несколько подходов, например собрать все темы докладов на оперативном совещании и обучить модель этими данными. Но интересная тема - это такая, которую никто еще не рассказывал. Поэтому было решено использовать новости. К слову сейчас Банк имеет договоры с агентствами, которые делают подборку с упоминанием Банка. Но во-первых, в моей команде на это нет бюджета, а во-вторых, можно попробовать же сделать лучше, чем предлагают эти агенства. **СЛАЙД**
Как обрабатывать новости? Новостной агрегатор от Яндекса имеет семь с половиной тысяч партнеров, которые генерируют ему **СЛАЙД** по сто тысяч новостей в день. **СЛАЙД**
Но мне не нужны все эти новости. Я выделил несколько критериев для себя: новости должны быть свежие, только финансовой тематики, ну и желательно популярными. **СЛАЙД**
БОльшая часть моего проекта - это сбор данных, т.к. готового датасета у меня нет. Как я это делал? Я выбрал для начала новостную ленту сайта банки.ру и пресс релизы на сайтах Альфабанка и ВТБ. **СЛАЙД**
На банках новости публикуются по 50 штук на отдельных страницах. И все новости сдвигаются, когда выходит новая. Приходится в скрипте перебирать все страницы и выгружать информацию по каждой новости. Я выгружал заголовок, текст, дату публикации, количество просмотров и комментариев, источник публикации. Если новость уже есть в базе, я обновляю данные о просмотрах и комментариях. Также я отдельно собираю датасет с обновлением просмотров, чтобы сделать модель по предсказанию количества просмотров в будущем. Этот алгоритм лежит в отдельном скрипте, чтобы его можно было запускать отдельно от основного Парсинга. На сайте Альфы нужно выбирать год и месяц, что проще. Для обновления можно выгружать только текущий месяц. **СЛАЙД**
Как я уже говорил, я собираю отдельно статистику просмотров последних десяти новостей, чтобы используя нейронную сеть с памятью LSTM попытаться предсказать просмотры в будущем. Сеть неплохо справилась, RMSE сети с оконным методом на тестовой выборке составило 12,7. **СЛАЙД**
Дальше я попробовал обратить внимание на текст заголовков и новостей. В работе использовал библиотеку nltk, делал токены из каждого слова, считал вхождения различных слов, убирал окончания. Из этой статистики вручную выбрал топ-100 слов, по которым в дальнейшем хочу отделять новости. Также с помощью библиотеки UMAP попытался отобразить тексты и заголовки.

Интересно, что на графике можно выделить отдельные области статей. В будущем буду подробнее изучать эти области. **СЛАЙД**
Какие возникали трудности? Их кстати было не мало. Например, сайт банки.ру разрывал соединение, после нескольких подключений. Для решения этой проблемы пришлось перейти на Гугл колаб. Видимо потом они решили, что это Гугл их скаинурет. ВТБ практически сразу забанил айпи адрес, пришлось использовать прокси. Кстати хороших бесплатных прокси практически не найти. Отдельно хочется отметить различное время публикации новостей. Из-за чего нельзя напрямую сравнивать просмотры. Для решения этой проблемы использую отношение просмотров к времени с момента публикации. Ну основная проблема - это текст. Мало того, что статьи с эмоджи и html разметкой, так еще и обработка окончаний - это отдельная песня. **СЛАЙД**
Что готов на данный момент?
Самостоятельно собран датасет с более, чем 120 тысячами новостей.
Обновление просмотров возможно 2 раза в минуту в течение часа.
Собран bagofwords на 100 единиц.
Работает модель на предсказание просмотров.
Ну и есть визуализация кластеризации новостей. **СЛАЙД**
Выводы после разработки проекта неутишительные. Своими руками можно сделать MVP и довольно интересный, при этом используя в основном уже готовые библиотеки. Заголовки новостей однородны, но тексты можно разделить на кластеры, скорее всего по тематике. **СЛАЙД**
Какие планы на будущее? Хочу добавить новые источники: Открытие, Тиньков, Газпромбанк и 1prime.ru. Хочу проанализировать собранный словарь из интересных слов, собрать по ним статистику. А также реализовать в полной мере фронт часть. Пока это email рассылка, но хочу перенести это в месенджер Телеграмм. **СЛАЙД**
Спасибо за внимание, готов ответить на вопросы.