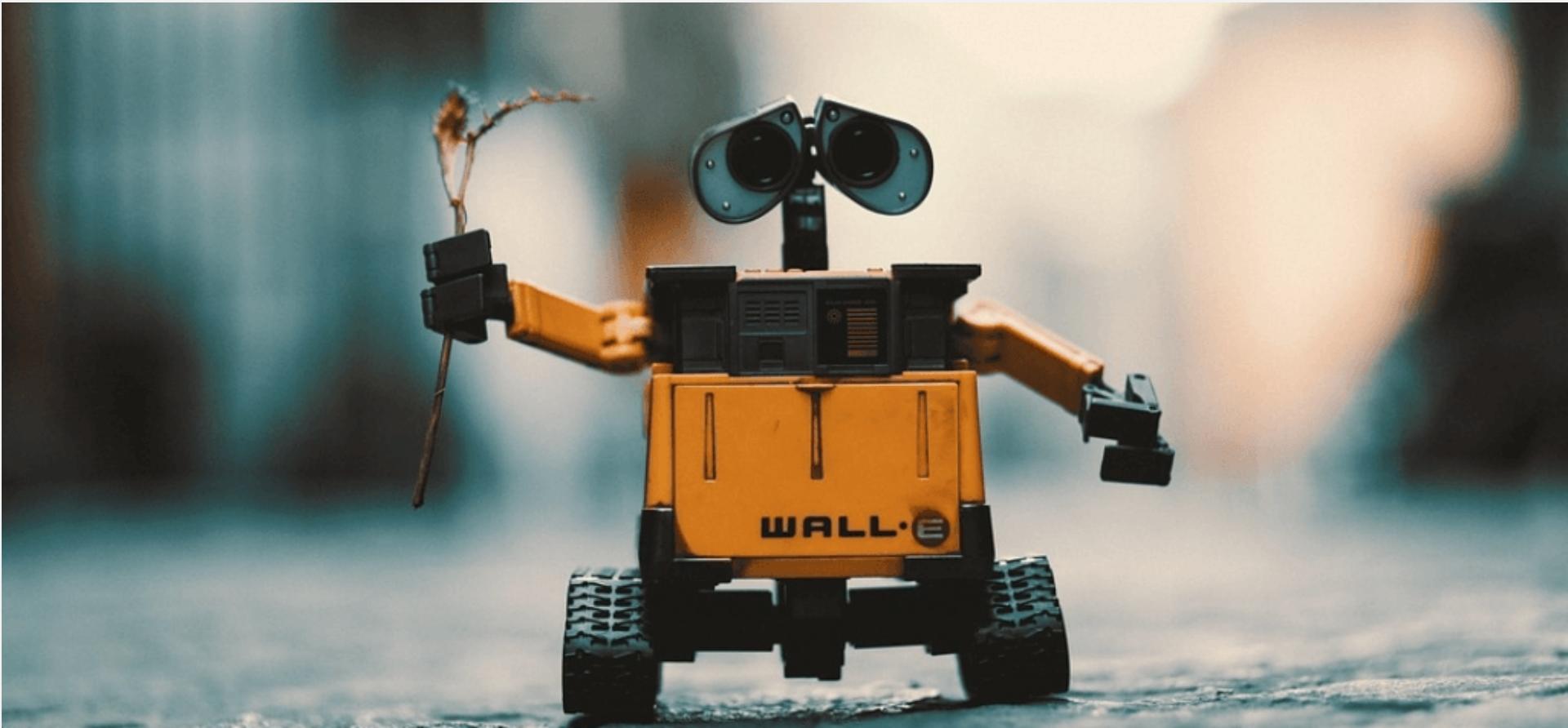


Почта Mail.Ru: ML Hands-On

Лекция 1
Осень 2020

Дмитрий Меркушов

Data Mining: Hands Off

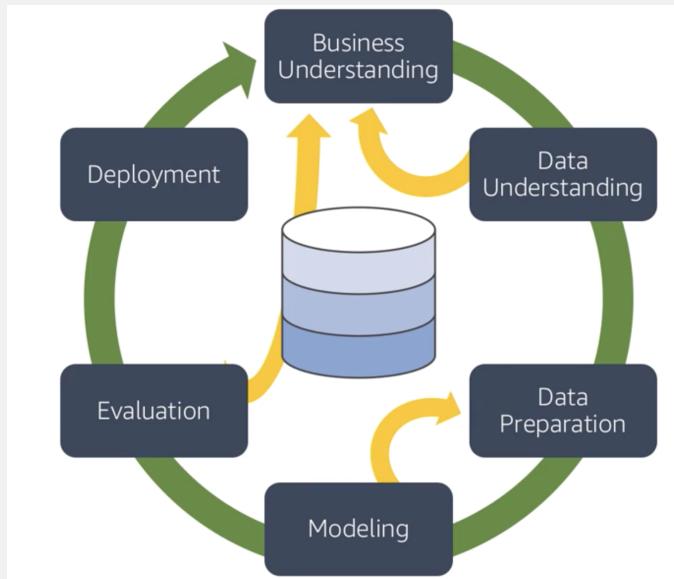


Duolingo: Глоссарий курса



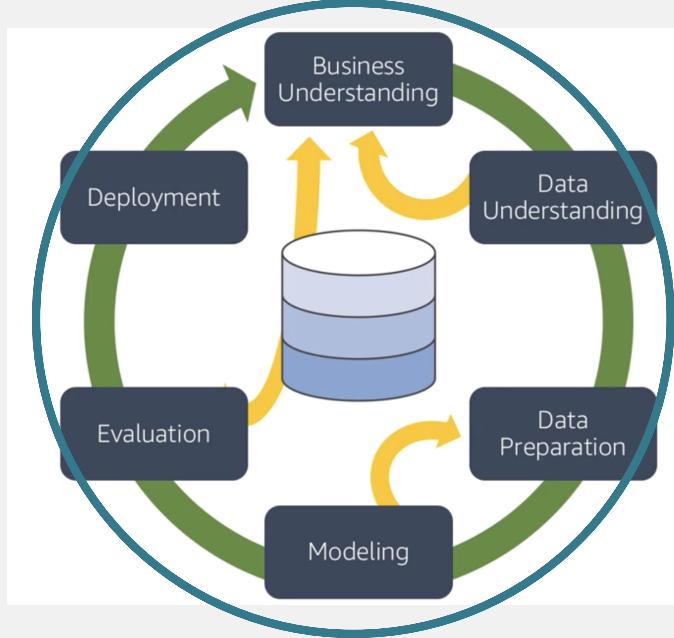
- Будет много специфических терминов – англицизмов
- Неполный список
 - Loss - функция потерь
 - Preprocess - предобработка данных для задач машинного обучения
 - Inference / Predict – предсказание модели на объекте
 - Overfit – переобучение
 - Hold Out – отложенная выборка
 - False – ошибка модели (false positive/negative)
 - Detect – положительное срабатывание классификатора
 - Assessor – человек, помогающий в разметке выборки
 - Domain Knowledge – владение предметной областью применения моделей
 - Sample – равномерный срез процента данных с продуктового потока
 - Feedback – обратная связь от пользователей модели(клики, лайки, шервы)
 - Production - боевое окружение вашего сервиса
 - Deploy – запуск модели в боевое окружение (на пользователей)
- А вообще, лучше учите **язык**, чем отдельные термины ;)

Data Mining: Life Cycle



- Бизнес формулировка
- Проверка гипотез
- Сбор данных для обучения
- Модель
- Оценка системы
- Вывод в бой

Data Mining: Life Cycle



- Бизнес формулировка
- Проверка гипотез
- Сбор данных для обучения
- Модель
- Оценка системы
- Вывод в бой
- + Поддержка

Почта: Категоризация



Почта: Категоризация



Входящие - Почта Mail.Ru Блог Почты Mail.Ru А вы защитили свою почту? Andrey
Защищено | https://e.mail.ru/messages/inbox/

Mail.Ru Почта Мой Мир Одноклассники Игры Знакомства Новости Поиск Все проекты ▾
a.shamne@inbox.ru ▾ выход

@mail.ru БЕТА Письма Контакты Файлы Темы Ещё

Написать письмо Удалить Спам Переместить Ещё

ВХОДЯЩИЕ

- Социальные сети
- Рассылки
- Важное
- Гитара
- Друзья
- Рыбалка
- Семья
- Фото
- Отправленные
- Черновики
- Архив
- Спам
- Корзина

ОЧИСТИТЬ ОЧИСТИТЬ

Удалить Спам Переместить Ещё

РАССЫЛКИ Biglion Москва, The New York Times, Tatyana Shamne, Штрафы ГИБДД М, PlayStation, iHerb, Бандеролька, Kickstarter, PlayStation, PlayStation, AliExpress Order, Tatyana Shamne, OZON.ru, OZON.RU, secure.payment@...

ФОРМАТЫ Fwd: Документы, Начало периода, Штраф ГИБДД погашен, Благодарим за покупку, Представляем доставку, Перемены всегда к лучшему, Projects We Love: The White Stripes, Благодарим за покупку, Благодарим за покупку, Andrey Shamne, the sender, Fwd: Welcome, Ваш заказ 22971728-0027 в OZON.ru оплачен, Информация о платеже, Информация о платеже

ВЫ НЕДАВНО ИСКАЛИ

От: info@twitter.com
От: Sony@email.sonyentertainmentnetwork.com
От: playstation@eu.playstationmail.net
От: follow-suggestion-noreply@quora.com
От: a.sergeev@corp.mail.ru

КАТЕГОРИИ

- Заказы
- Финансы
- Регистрации
- Путешествия
- Билеты
- Штрафы

Расширенный поиск

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки

Почта: Категоризация



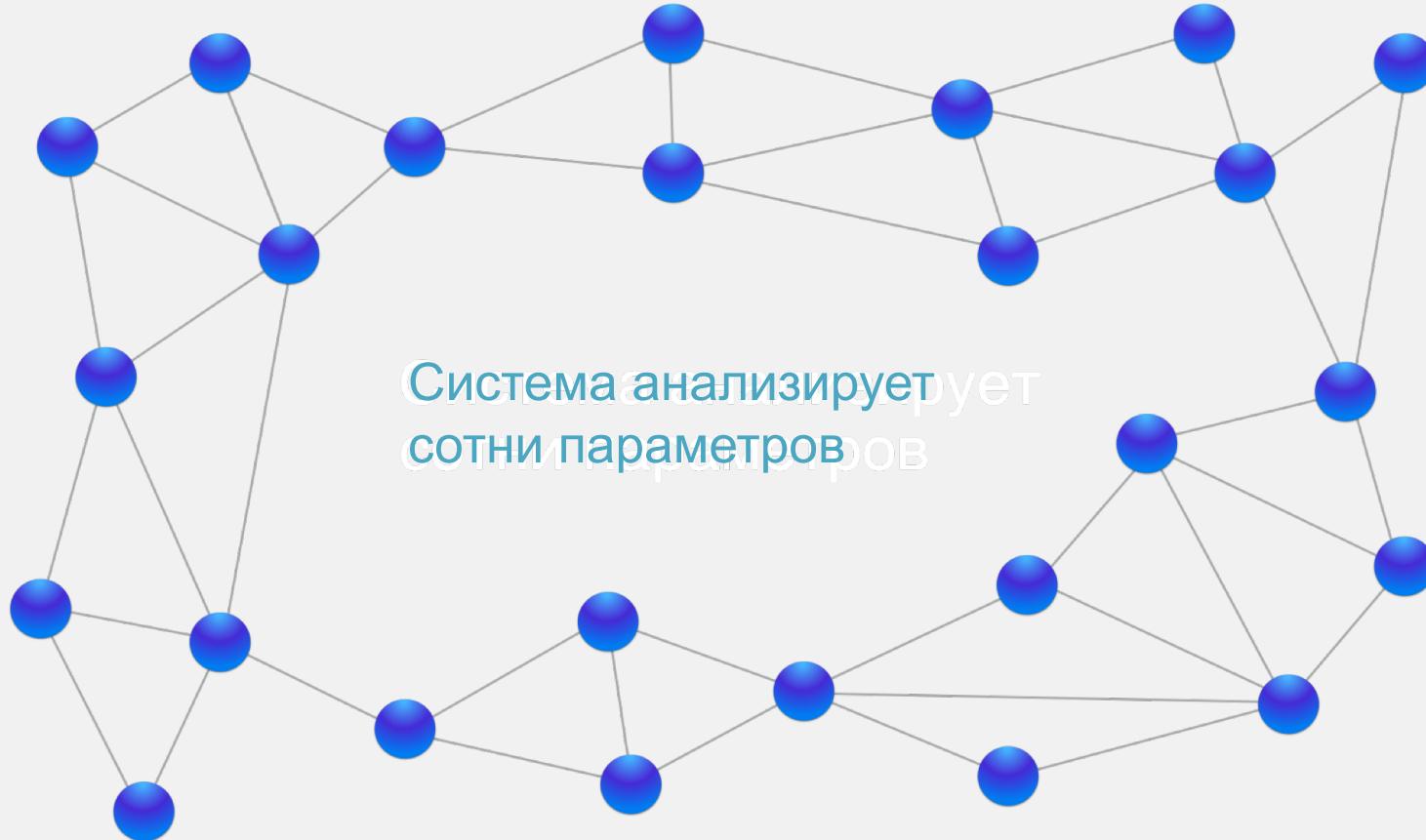
- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки
- Модель
 - Разное, от простого к сложному
 - Bag of words, word2vec классификаторы

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки
- Модель
 - Разное, от простого к сложному
 - Bag of words, word2vec классификаторы
- Оценка
 - Алгоритмические – метрики на выборках
 - Технические – оценка precision/recall на семплах
 - Продуктовые – число регулярных детекторов, число фолзов системы

Антиспам: Marshal



Антиспам: Marshal



Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов

Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения

Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения
- Сбор данных
 - Очень много данных сессий пользователей
 - Выделить явно хорошие, явно плохие
 - Профит – выделить граничные плохие, возможно ассессорами

Антиспам: Marshal



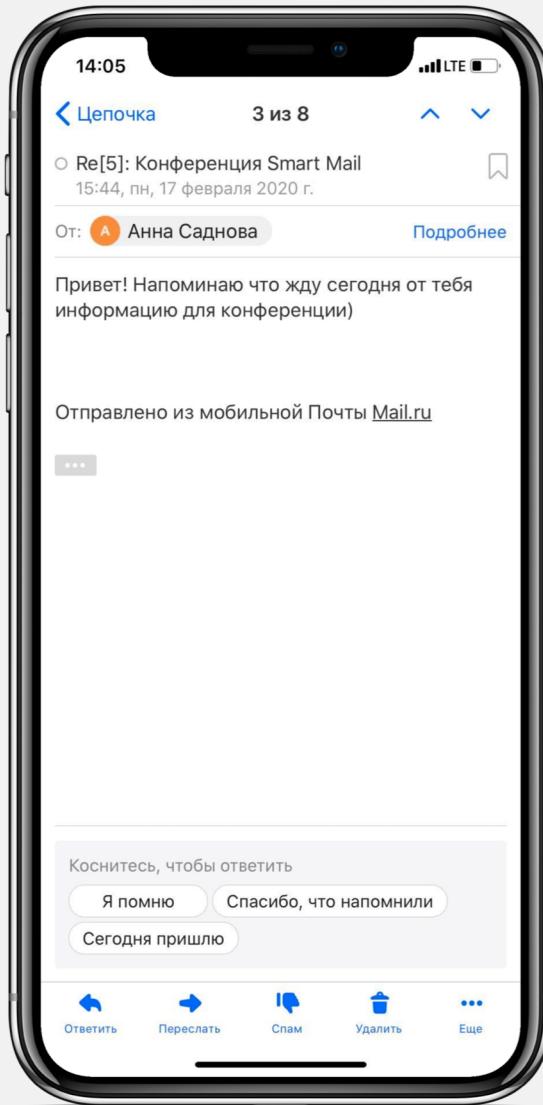
- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения
- Сбор данных
 - Очень много данных сессий пользователей
 - Выделить явно хорошие, явно плохие
 - Профит – выделить граничные плохие, возможно ассесорами
- Модель
 - Пробуем разные, от простого к сложному
 - Pattern Mining: Apriori, FP-Growth

Антиспам: Marshal

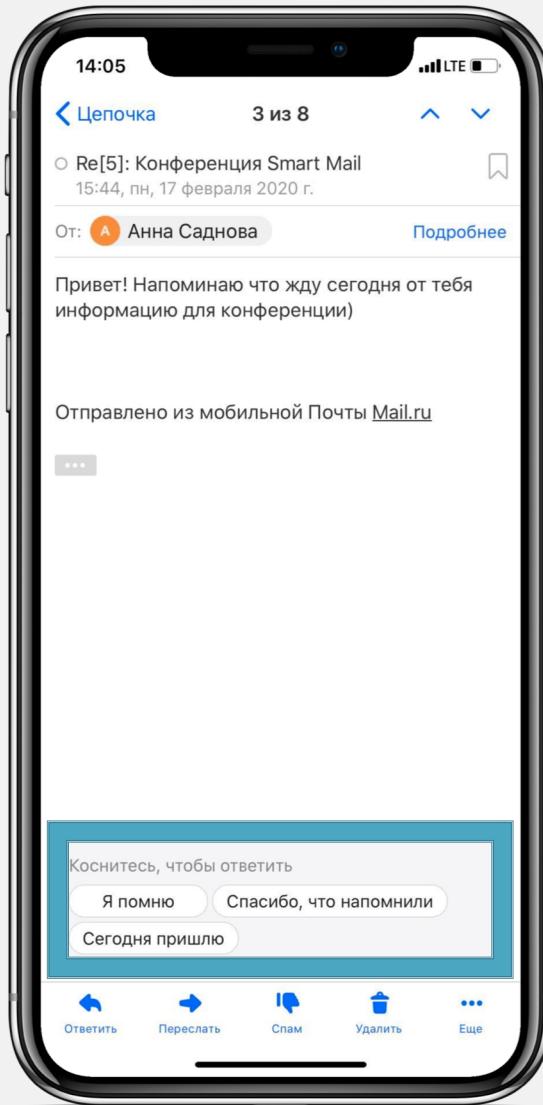


- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения
- Сбор данных
 - Очень много данных сессий пользователей
 - Выделить явно хорошие, явно плохие
 - Профит – выделить граничные плохие, возможно ассесорами
- Модель
 - Пробуем разные, от простого к сложному
 - Pattern Mining: Apriori, FP-Growth
- Оценка
 - Технические – метрики на выборках
 - Технические – оценка precision/recall на семплах
 - Продуктовые – число регулярных детекторов, число фолзов системы

Почта: SmartReply



Почта: SmartReply



Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first

Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set

Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set
- Сбор данных
 - Выделяем большой корпус
 - Прочищаем от шлака, цензурируем
- Модель
 - Seq2seq, DSSM

Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI based
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set
- Сбор данных
 - Выделяем большой корпус
 - Прочищаем от шлака, цензурируем
- Модель
 - Seq2seq, DSSM
- Оценка
 - Технические – метрики на выборках, семплах prec@3
 - Продуктовые – % пользования фичей, % показанных ответов

Антиспам: SafeClick



Антиспам: SafeClick



Антиспам: SafeClick



- Проблема
 - Шумный фидбек от пользователей



Антиспам: SafeClick



- Бизнес формулировка
 - Получать достоверную оценку сервиса пользователями
 - Обучать ML-системы на надежных данных

Антиспам: SafeClick



- Бизнес формулировка
 - Получать достоверную оценку сервиса пользователями
 - Обучать ML-системы на надежных данных
- Проверка гипотез
 - Есть «ядро» пользователей, которые не шумят
 - Достоверный фидбек можно выделить репрезентативным образом

Антиспам: SafeClick



- Бизнес формулировка
 - Получать достоверную оценку сервиса пользователями
 - Обучать ML-системы на надежных данных
- Проверка гипотез
 - Есть «ядро» пользователей, которые не шумят
 - Достоверный фидбек можно выделить репрезентативным образом
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки

Антиспам: SafeClick



- Бизнес формулировка
 - Получать достоверную оценку сервиса пользователями
 - Обучать ML-системы на надежных данных
- Проверка гипотез
 - Есть «ядро» пользователей, которые не шумят
 - Достоверный фидбек можно выделить репрезентативным образом
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки
- Модель
 - Главное в этой задаче – качественные признаки
 - Сначала линейная модель, для повышения точности перешли на градиентный бустинг

Антиспам: SafeClick



- Бизнес формулировка
 - Получать достоверную оценку сервиса пользователями
 - Обучать ML-системы на надежных данных
- Проверка гипотез
 - Есть «ядро» пользователей, которые не шумят
 - Достоверный фидбек можно выделить репрезентативным образом
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки
- Модель
 - Главное в этой задаче – качественные признаки
 - Сначала линейная модель, для повышения точности перешли на градиентный бустинг
- Оценка
 - Алгоритмические – метрики на выборках
 - Технические – оценка precision/recall на семплах
 - Продуктовые – число регулярных детекторов, число фолзов системы

<https://new.mail.ru>
пробуйте