

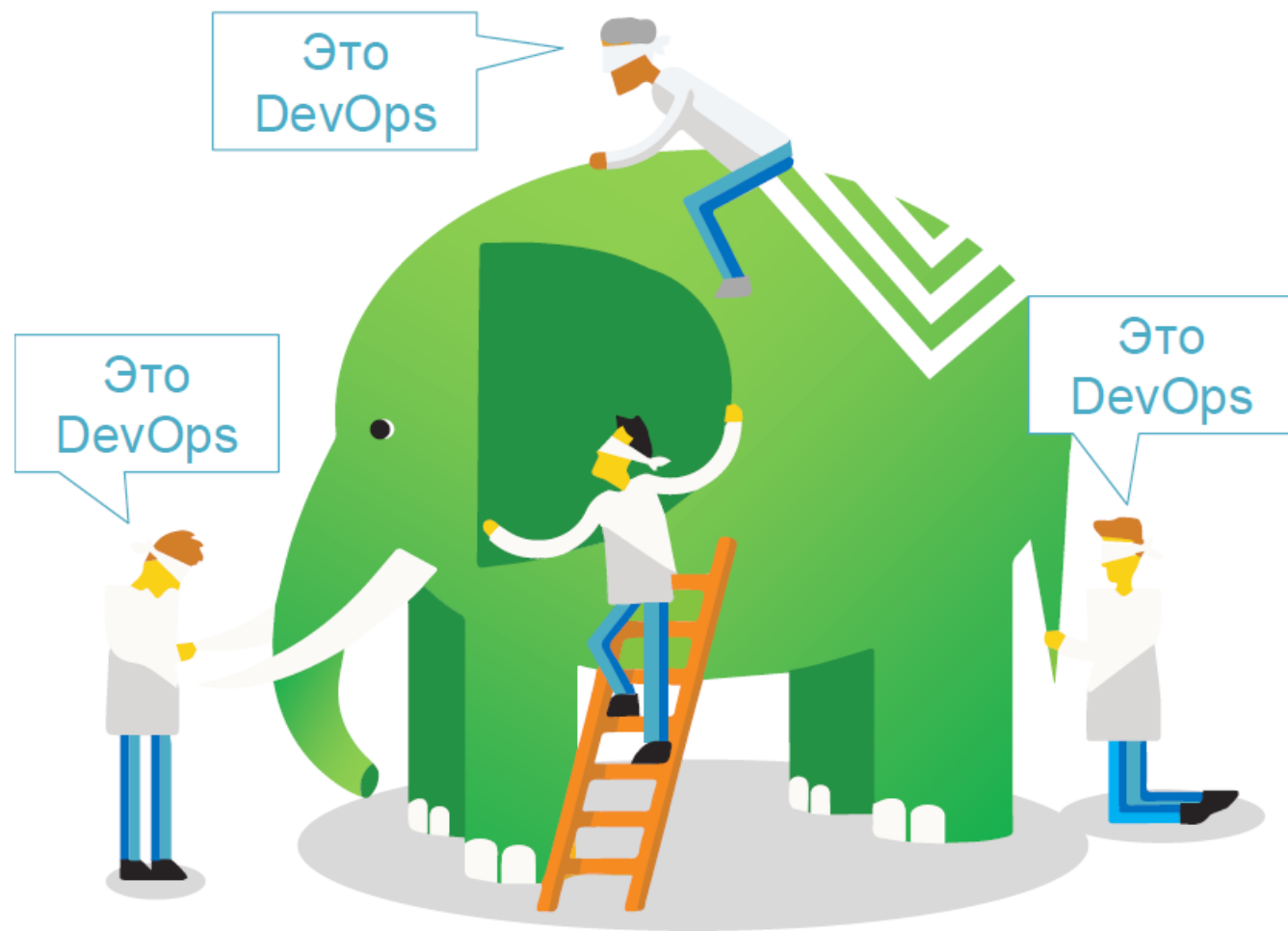


Как повысить эффективность разработки: хакатон по ML в СберТехе



MYuRuzhentsev.sbt@sberbank.ru

Руководитель направления по развитию ИТ-систем



Ключевые цифры о данных в Сбербанке

Прогноз прироста данных на **2018 год – 10 Пб** (сырых данных) -> **15,5 Тб в день**

Объем хранения данных в Hadoop ~ **2,5 Пб**, планируется к концу 2018 - **10Пб**

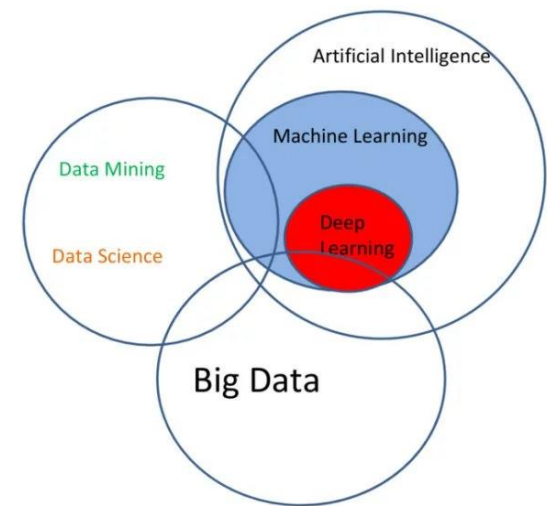
Hadoop кластер 250 узлов (6800 CPU cores, 110 ТБ memory, 10 Пб HDD)

Почему Machine Learning

Arthur Samuel (1959).

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Машинное Обучение: Область исследования, которая дает компьютерам возможность учиться, не будучи явно запрограммированными.



Ключевые особенности системы Machine Learning



Является частью платформы



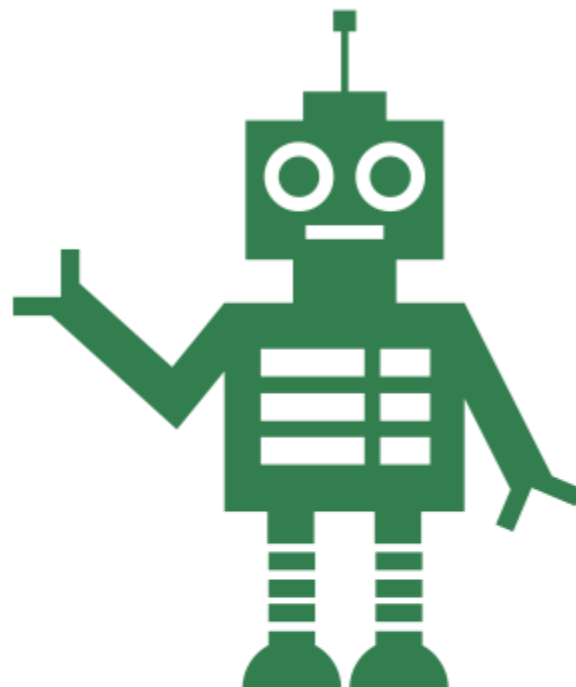
Обеспечивает полный жизненный цикл модели



Позволяет автоматически дообучать модели на лету



Используются технологии с открытым кодом



Где взять дата сет ?



Безопасность



Очевидная ценность, размер достаточен и данные не слишком “грязные”



Вариант проведения Хакатона

1 часть – Сбор качественных и заведомо полезных данных самими участниками

2 часть – участники упражняются в вытаскивании максимально точного и качественного результата моделей

Плюсы

- качество/ценность полученных результатов
- простота выявления победителей

Минусы

- продолжительность проведения
- отсутствие результата "здесь и сейчас"



Вариант проведения Хакатона

Участники упражняются в вытаскивании максимально точного и качественного результата моделей

Плюсы

- получение результата "здесь и сейчас"

Минусы

- Если датасет простой, то большая вероятность "близких результатов" команд

Исходные данные

- JIRA – внутренняя сеть (текстовые поля с историей плюс вложения)
- JIRA – внешняя сеть (текстовые поля с историей плюс вложения)
- ЦУП – этапы планирования внедрения дистрибутива





- лучше планировать загрузку команд в ходе исправлений на разных этапах разработки
- формировать структуру релизов в части hot fix
- планировать работу в спринтах — определять story points, зарезервированные для исправления проблем
- в целом снижать количество багов в программном коде
- сократить время выхода продуктов на рынок
- увеличить предсказуемость решений и эффективность тестирования



- Полезность для СБТ
- Техническая реализация - уместное использование технологий
- Подготовка данных
- Дополнительные источники
- Оформление результатов

Как проводить хакатон ?



2 недели в “свободном” режиме



Чтобы мощности хватило на все вычисления - решили
быстро развернуть кластер в **OpenStack**



Мы хотели соблюсти принципы
честной игры, и не давать
преимущества тем, кто уже работает
на выбранном стенде перед теми,
кто впервые решил прикоснуться к
машинному обучению



80CPU 200Gb 1.5Tb - звучит как одноюнитовый сервер с упором на
вычисления.



«работает» ли Hadoop в OpenStack ?

cloudera



«работает» ли Hadoop в OpenStack ?



cloudera



«работает» ли Hadoop в OpenStack ?

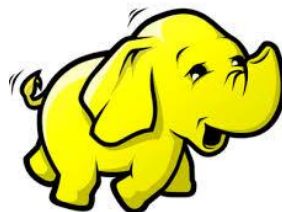
❗ <https://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/>

- 12-24 1-4TB hard disks in a JBOD (Just a Bunch Of Disks) configuration
- 2 quad-/hex-/octo-core CPUs, running at least 2-2.5GHz
- 64-512GB of RAM
- Bonded Gigabit Ethernet or 10Gigabit Ethernet (the more storage density, the higher the network throughput needed)

- ❖ *Персональные рабочие места для DS с GPU*
- ❖ *Обучение нейронных сетей на полных данных в кластере*
- ❖ *Applications As Service – настроенный Jupyter Notebook с NVIDIA P100*



Почему был выбран Hadoop ?



SQL анализ на
данных Hadoop



Вычисления в
памяти



Языки



ETL

Informatica Big
Data
Management

Доставка
данных

Sqoop

Приложения



Security

Kerberos
Sentry

Управление ресурсам кластера

YARN

Распределенная файловая система кластера



OpenSource – ПО, распространяемое с открытым исходным кодом, как свободное для изучения, переиспользования и модификации исходного кода конечными пользователями

JupyterHUB - создает
параллельные инстансы на всех
участников



- Разрабатывать и запускать скрипты на Python, R
- Выполняет код как локально, так и на кластере Hadoop (PySpark)





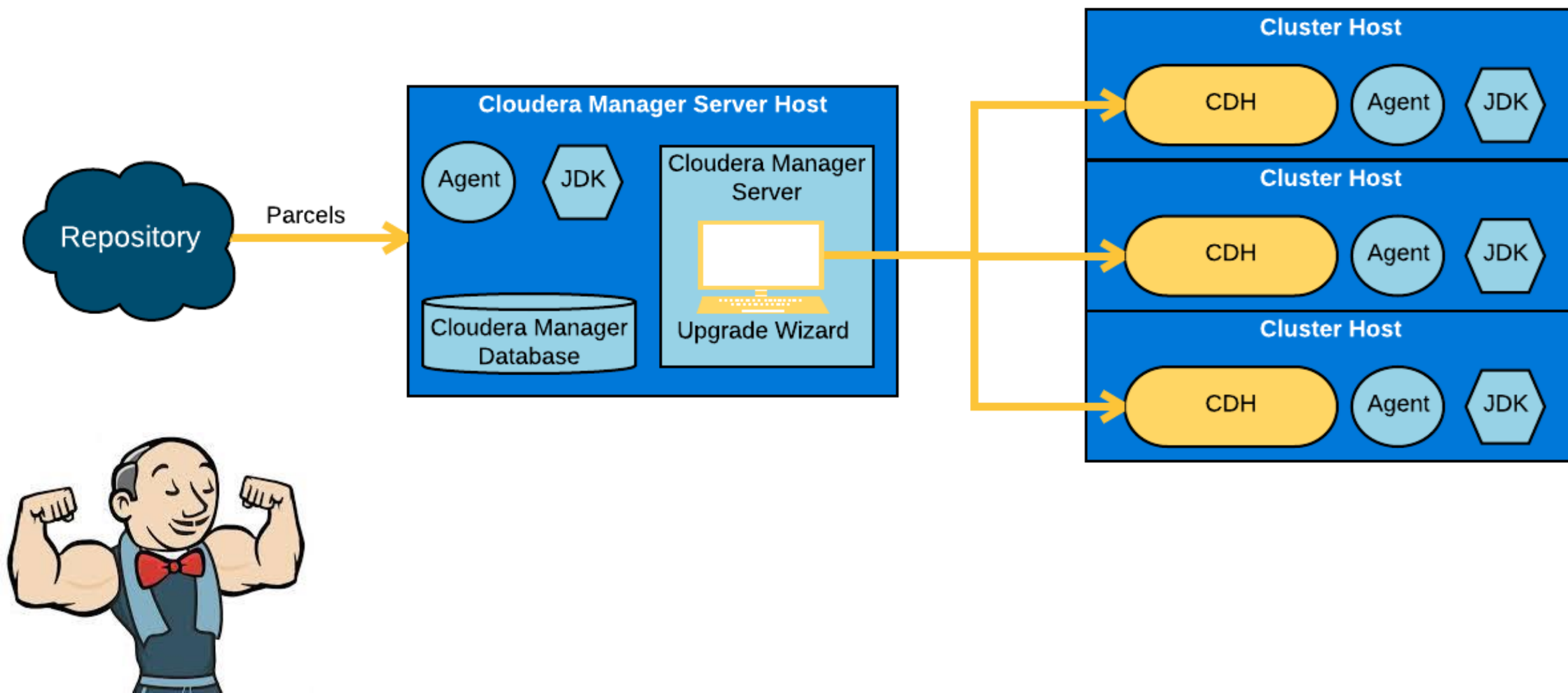
Python 3.5.3

Python 3.5.3 PySpark 2.1

R 3.4 (Spark 2.1)



Data
Science





120 участников от новичков до экспертов

Призовой фонд — 300.000 рублей

Опыт — бесценно

Официальная цель: получить MachineLearning-решения, которые будут предсказывать сроки реализации доработок и критичность багов.

Лучшие идеи планируется внедрить в СберТехе для повышения качества производственных процессов.



Проблемы с кластером (недовольство) некоторых команд
Что делать с датасетом ? Покажите примеры!

- Борьба за ресурсы
 - Новички - считают на edge ноде
 - Продвинутые - жрут все ресурсы, используя ruspark
 - Самые наглые - прибивают чужие задачи
 - Из JN можно выполнить ssh команды



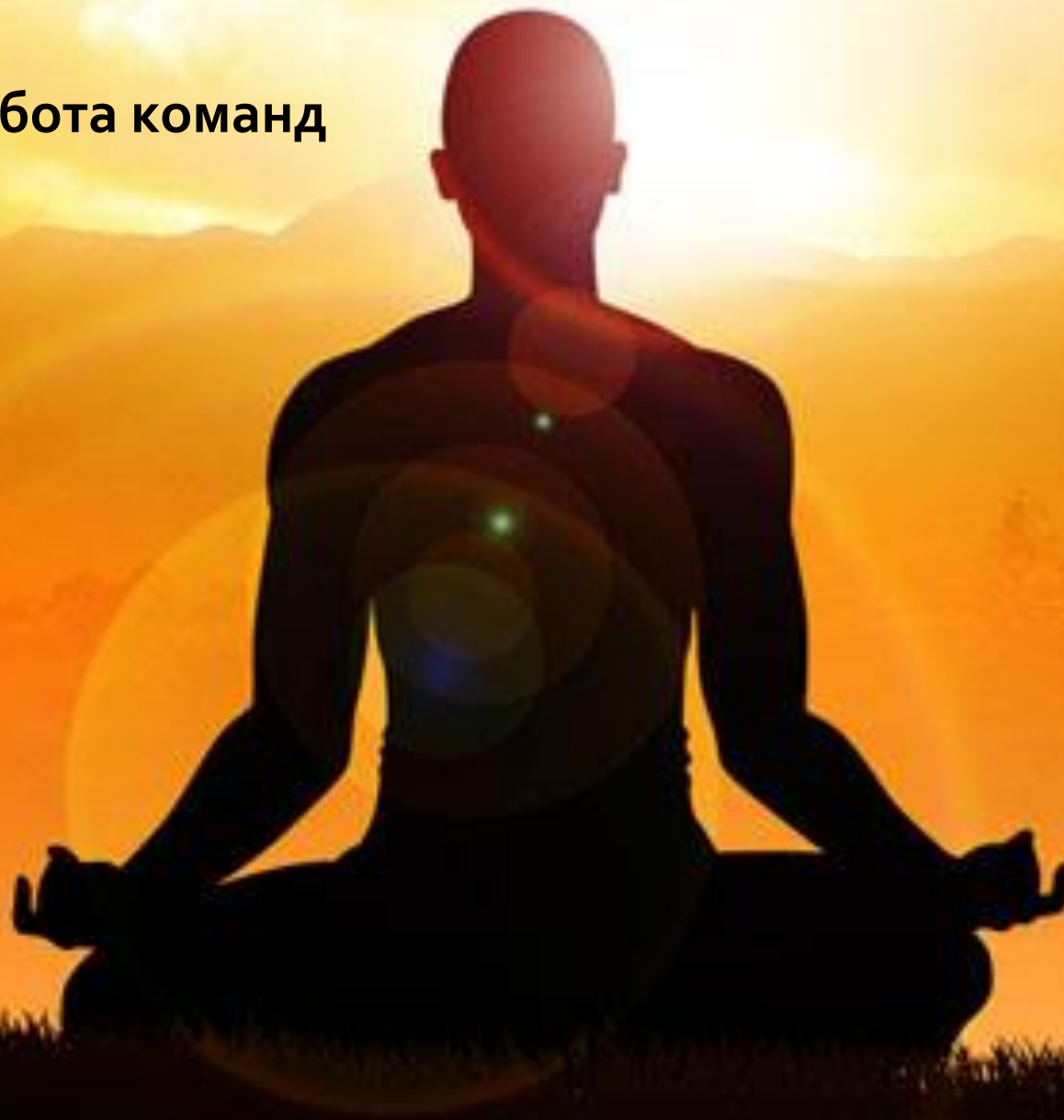


- Виды очередей
 - Статические очереди
 - Динамические очереди



- Как поступили
 - Автоматом создаются очереди
 - По сути - одна очередь
 - Для тех кто просил - создавали отдельные

Кропотливая работа команд



Итоги

Пятерка лучших была приглашена на ДФ для очной презентации результатов

III место: прогнозирование рисков

Иннополис

Цель: Создание системы прогнозирования на основе нейронной сети для минимизации рисков при управлении ИТ-проектами.

Обучение модели

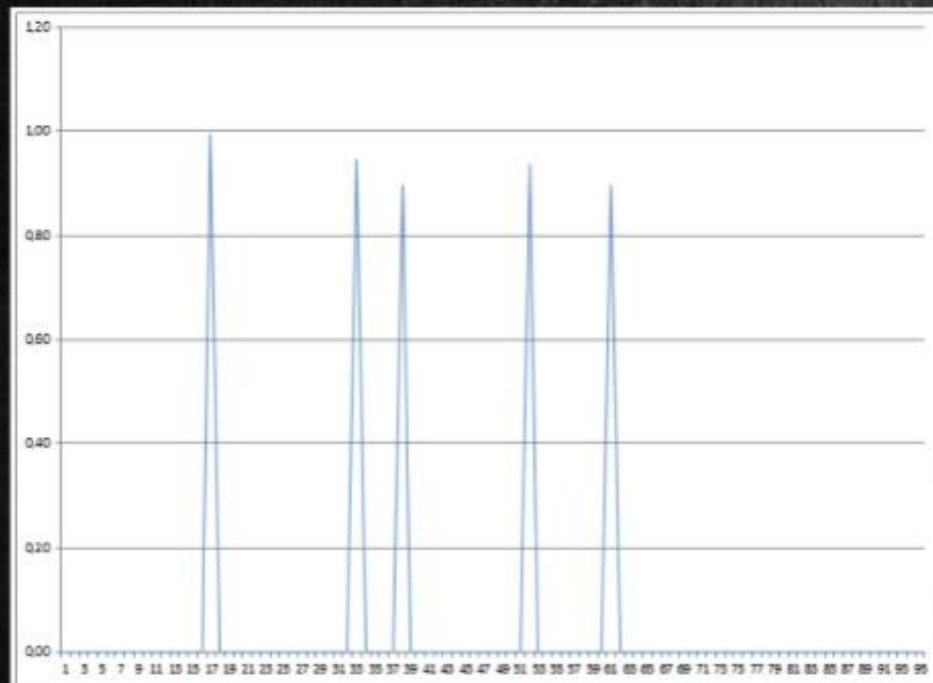


Рисунок. Состояния, в которых побывала/будет задача близки к 1, остальные неотличимы от 0

По каждой задаче воссозданы все изменения, полей с момента её создания. По всем задачам получено 346тыс. состояний на временной шкале. Для обучения случайным образом выбрано 3% из этого массива. В выборку попали задачи на разных этапах своего существования. 1,5% от общего объема контрольная выборка, не пересекающаяся с данными для обучения. Для проверки точности обучения использованы известные даты смены состояния задачи.

Например, на вход нейросети подается задача на этапе разработки, а на выходе, расчётная дата закрытия проекта сверяется с известной датой закрытия по этой же задаче.

Полученная модель достаточно точно предсказывает появление различных состояний.

Анализ результатов



Рисунок. Показывает, на сколько дней отклоняется прогноз по дате от фактической. Пик во второй половине графика показывает большое отклонение для задач на ранних этапах.



Рисунок. Показывает, что отклонение не зависит от объема задачи и имеет равномерное распределение, что подтверждает корректность метрик

Итоги

Выводы:

Модель показала, точность прогнозирования в диапазоне от 91 до 94%.

Модель подходит для выявления на ранних этапах задач, которые могут содержать ошибки, что приведет к их повторному открытию. Прогноз позволяет заранее принять управленческие решения для предотвращения потерь времени.

Модель подходит для прогнозирования даты наступления различных событий, в том числе даты закрытия задачи. Отклонение от 0 до 37 дней.

Модель имеет ошибку Loss в районе 0.54, что достаточно велико, но в то же время приемлемо, т.к. подходит для задач на разных этапах, в т.ч. сразу после открытия.

Рекомендации:

Для повышения точности прогнозирования необходимо повысить дисциплину ведения учета в Jira, как источнике первичных данных.

Необходимо очистить справочники от мусорных данных для исключения ошибок ведения и дублирования значений.

Необходим доступ к дополнительным справочникам для получения более точных данных, например важна дата приёма сотрудника, его возраст и стаж.

Построение модели с привязкой к git позволит получить еще более точные результаты.

II место — оптимизация производственных процессов

Москва

Задачи:

- предсказать итоговый приоритет бага, используя информацию о нем.
- на основе данных о распределении багов во времени предсказать число багов

Данные

В качестве исходных данных выбраны баги из Jira сегмента ALPHA

Данные содержат описание **67 593** багов в статусе **Завершён** с 2011 по 2017 год

Описание каждого бага содержит **более 500** возможных полей, с итоговыми значениями на момент закрытия бага

Changelog хранит информацию о всех изменениях полей

Что не
используется

Что
используется

Отбор
признаков

Отбор признаков

Задача классификации требует отбора признаков для построения модели

Целевой признак:

- **Priority** - приоритет бага

Отобранные признаки:

- Summary - краткое описание бага
- Description - подробное описание бага
- Created - дата создания записи о баге
- Creator - автор записи о баге
- Customfield_10501 - функциональная подсистема
- Customfield_14009 - тип бага
- Customfield_14007 - этап обнаружения бага
- Project.key - код системы, в которой обнаружен баг

Priority

Summary

Description

Created

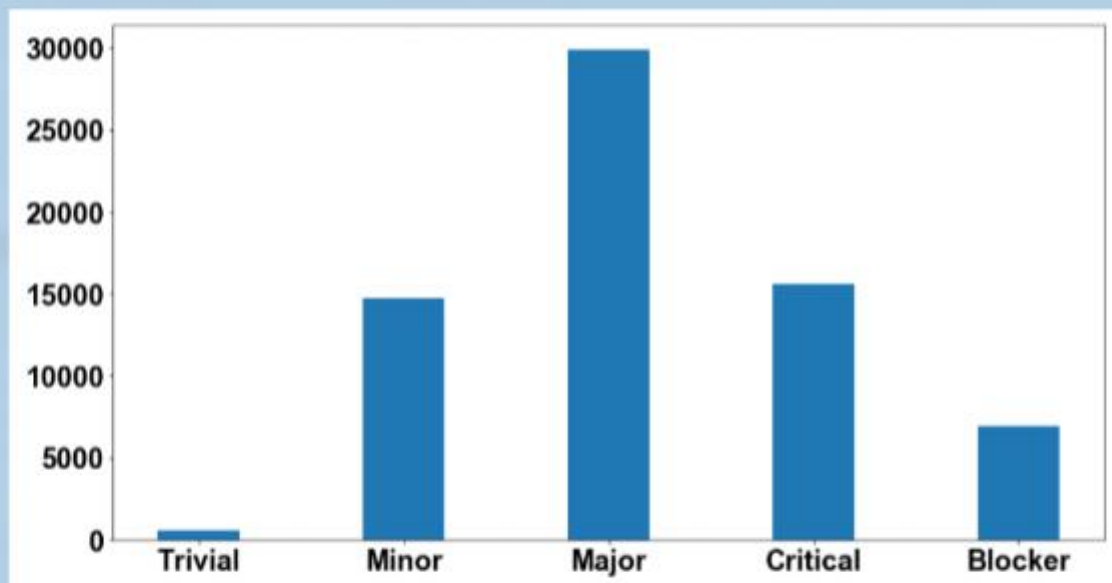
Остальные
признаки

V Крамера

Итог

Priority

Приоритет бага
Целевой признак модели



Меняется в процессе работы по
устранению бага

Initial
Priority

Initial Priority

Исходное значение признака Priority,
выставленное при создании записи о баге

Может быть получено из контейнера Changelog

Initial priority

Priority		Trivial	Minor	Major	Critical	Blocker
	Trivial	361	65	67	31	10
	Minor	60	13196	1142	232	66
	Major	31	1176	27282	1132	284
	Critical	9	231	1314	13620	373
	Blocker	0	91	207	465	6148

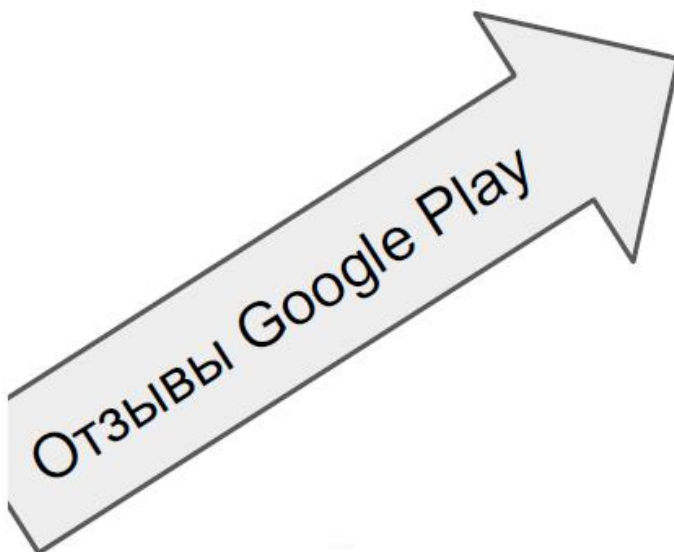
Совпадает с priority с точностью 89,7%

Преобразуется с помощью
One-hot кодирования

I место — автообработка багов

Москва

Цель: Создание pipeline автообработки багов в проектах Сбербанк-Технологии.



Отзывы пользователей



Артём Наумов 24 мая 2017 г.

★★★★★

После обновления кроме взаимодействия с Android Pay ожидал, что в приложении появится своя бесконтактная оплата, а оно только отсылает в гуглоприложуху мой номер карты.. И если я уже добавил карту, зачем предлагать снова её добавить в AP? Запуск приложения стал дольше раза в три, зачем надо каждый раз запускать проверку антивирусом? Добавьте возможность досрочного частичного погашения кредита.
Xiaomi MI 5, Android 7.0



Вячеслав Коперсак 3 июля 2017 г.

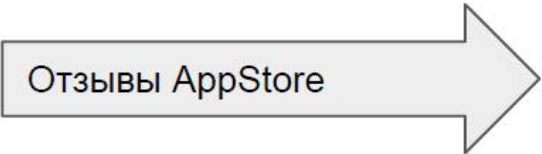
★★★★★

Очень лаконичное и приятное приложение! Хотя есть ряд замечаний! Столкнулся с проблемой встроенного антивируса, находит вирус но удалить не может, думаю эта проблема связана с частичным открытием root прав. Еще есть просьба добавить в копилку галочку для пополнения ее от всех перечислений на карту, включая сторонние счета! Спасибо за удобный функционал! Разработчики молодцы!

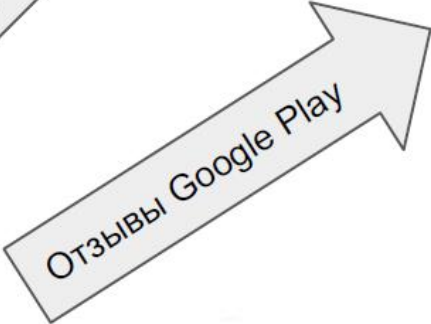
Отзывы
пользователей с
оценкой



Размеченная
выборка отзывов



	text	name_user	date	rating	review_positive
0	Приложение не обновляется и не скачивается за...	Sasha Pershina	9 июля 2017 г.	1.0	0
1	Хорошее приложение облегчает платежи	Oleg S	9 июля 2017 г.	5.0	1
2	Нет придела совершенству. Практичное приложен...	Евгения Ключникова	9 июля 2017 г.	4.0	1
3	Нормас работает		9 июля 2017 г.	4.0	1
4	Удобное приложение, основные функции отрабаты...		9 июля 2017 г.	5.0	1



Отзывы покупателей

Выпустил ★★★★★
от Minigore2

Ахах
Выпустил неделю назад MasterCard
Для пользования ApplePay
Не успел забрать карту
И наконец-то выкатывают для Visa
Ладно) будет две карты)
Спасибо☺

СУПЕР!!! ★★★★★
от Velvii

Вы большие молодцы! Супер приложение и даже Apple Pay для карт Visa появился!!!

Visa Apple Pay ★★★★★
от RomK@

Вот теперь 🙏 Изменяю рейтинг как и обещал!

Отзывы пользователей



Артём Наумов 24 мая 2017 г.

★★★★★

После обновления кроме
взаимодействия с Android Pay



Вячеслав Коперсак 3 июля 2017 г.

★★★★★

Очень лаконичное и приятное
приложение! Хотя есть ряд
замечаний! Столкнулся с
проблемой встроенного

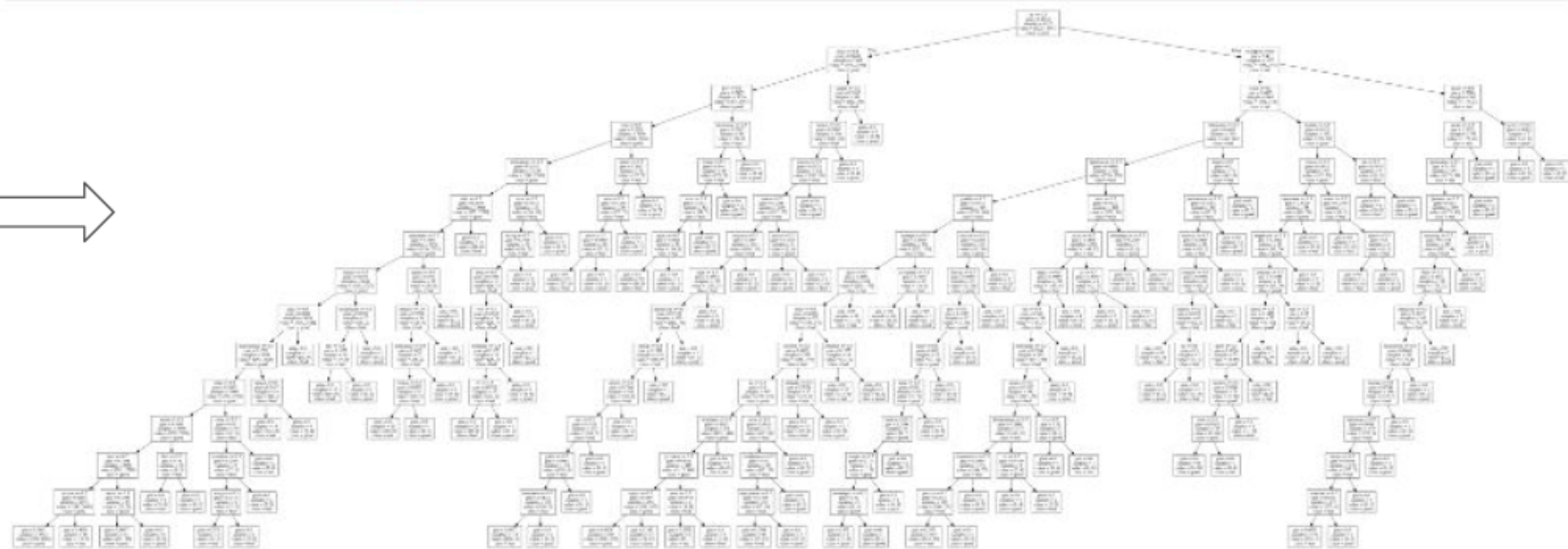
Векторизированное
текстовое представление

DecisionTreeClassifier

accuracy = 86%,
precision = 88%

Классифицированные
отзывы пользователей

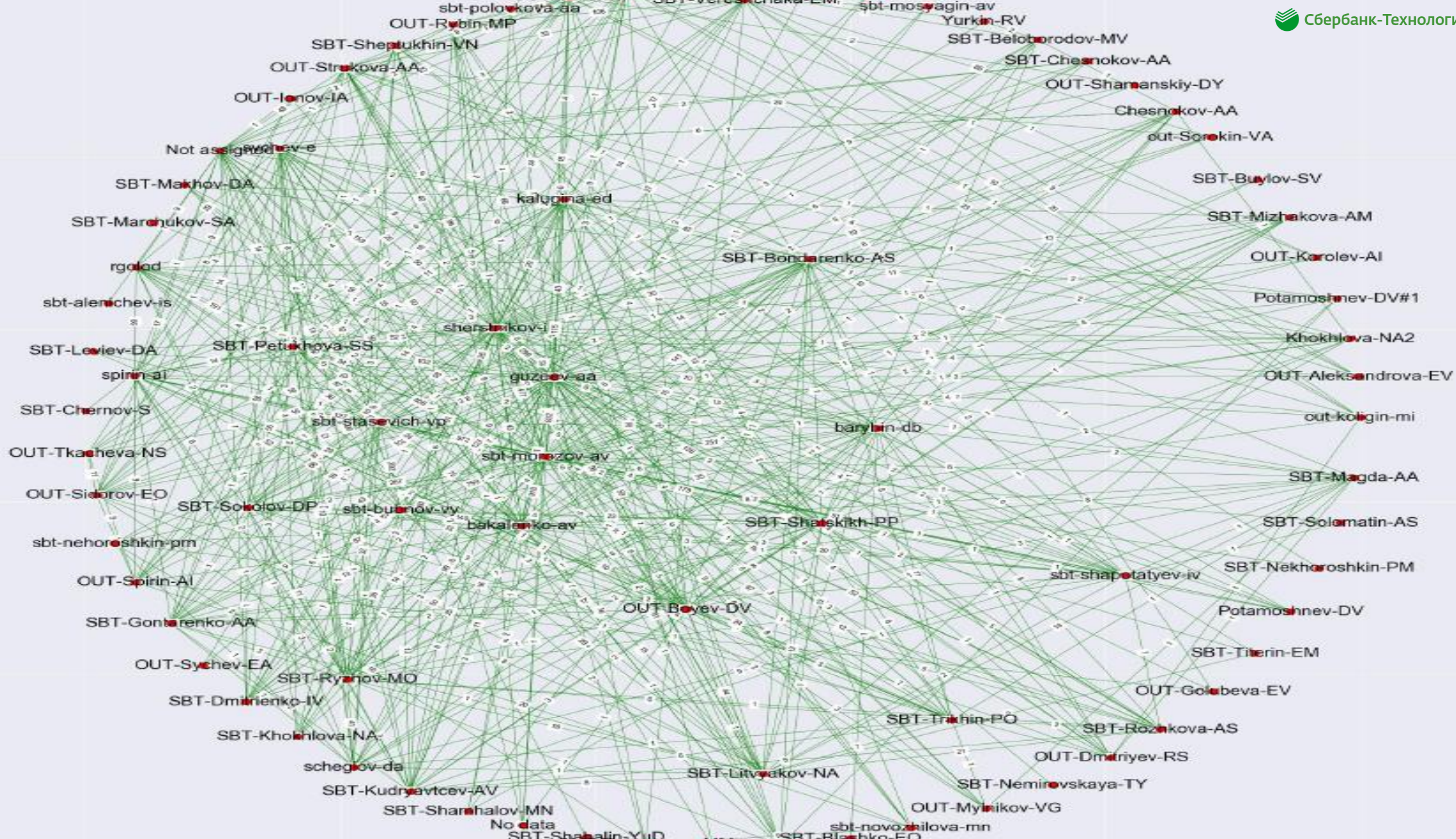
0 1 0 1 1 0 0 1 0 1
1 0 1 0 0 1 1 0 1 0
1 0 1 0 1 0 1 1 0 1
0 1 0 1 1 0 0 1 0 1
1 0 1 0 0 1 1 0 1 0
1 0 1 0 1 0 1 1 0 1
0 1 0 1 1 0 0 1 0 1
1 0 1 0 0 1 1 0 1 0
1 0 1 0 1 0 1 1 0 1
0 1 0 1 1 0 0 1 0 1



В результате данного этапа получен классификатор отзывов. Он необходим для уменьшения шума в данных, в ситуациях, где пользователи указали неверную оценку (не соответствующую отзыву)

DecisionTreeClassifier строит дерево принятия решений на основании критерия джени. Грубо говоря, на каждом шаге строится разбиение выборки по признаку, по которому данное разбиение наилучшее. Итог данного этапа - выделение вероятно негативных отзывов.

Стек: python, sklearn, nltk, stopwords



как победить на хакатоне по ML

СПАСИБО !

Михаил Руженцев
MYuRuzhentsev.sbt@sberbank.ru

Презентации команд

<https://drive.google.com/file/d/0B3WFrRsNpUNAcEMxNE1peHNXQIE/view>
<https://prezi.com/view/7YCssok7yNpAcLyhKy9u/>
<https://drive.google.com/file/d/0B3WFrRsNpUNAb1Jvckl0VjBBd28/view>