Практическое задание №1 для студентов кафедры СП. Осень 2019

Постановка задачи

Целью работы является разработка метода, позволяющего относить новостную публикацию (на русском или английском языках) к одной или более тематике в области информационной безопасности. Другими словами, каждому тексту новости необходимо поставить одну или более меток тем из предложенного списка. Если текст новости не относится ни к одной из тем, перечисленных в списке, такой новости необходимо поставить метку «Прочее»

Список меток тем

- «Угроза» – метка назначается новости, если текст новости содержит описание (упоминание) одной или более угроз безопасности. Под угрозой безопасности понимается совокупность факторов и условий, которые могут создать опасность в отношении информации.
- «Уязвимость» – метка назначается новости, если текст новости содержит описание (упоминание) одной или более уязвимостей ПО. Под уязвимостью понимается недостаток в системе (ПО), используя который, можно намеренно нарушить целостность системы или вызвать некорректную работу системы.
- «Эксплойт» – метка назначается новости, если текст новости содержит описание (упоминание) одного или более эксплойта. Под эксплойтом понимается описание (в виде программного кода или последовательности действий) способа эксплуатации уязвимостей ПО.
- «Инцидент» – метка назначается новости, если текст новости содержит описание (упоминание) одного или более события нарушения информационной безопасности (реализации угроз).
- «Вредоносное ПО» – метка назначается новости, если текст новости содержит описание (упоминание) вредоносного ПО. Под вредоносным ПО понимается программное обеспечение, предназначенное для получения несанкционированного доступа к ресурсам ЭВМ или к информации, хранимой на ЭВМ, с целью использования ресурсов или причинения вреда.

Примеры разметки новостей приведены в приложении А.

Решение задачи

Практические аспекты

Решения проверяются удаленно. Интерфейс автоматической системы находится по адресу: https://2019-1.tpc.ispras.ru.

Решения должны быть написаны на языке Python (версия 3.6.3). Можно использовать все стандартные библиотеки, а также

- NLTK - инструменты для обработки текстов
- scikit-learn - алгоритмы машинного обучения
- numpy - работа с многомерными массивами
- tensorflow, pytorch – библиотеки для работы с искусственными нейронными сетями

В случае необходимости использования дополнительных библиотек, сообщите об этом организаторам практикума (библиотеки будут добавлены для всех студентов).

Доступ в Интернет на проверяющей машине закрыт.

Теоретические аспекты

В рамках решения первого задания практикума предполагается использование методов машинного обучения с учителем. Для обучения метода требуется придумать признаки и дать ему на вход правильные примеры - обучающий корпус.

Считается, что чем больше обучающий корпус, тем лучше работает алгоритм. Однако создание большого обучающего корпуса - довольно трудоемкая задача, непосильная одному человеку. Поэтому предлагается создать его с помощью коллективной работы. Чтобы облегчить эту работу, была разработана система разметки: https://2019-1.tpc.ispras.ru.

Разметка обучающего корпуса

Для разметки корпуса необходимо зарегистрироваться на сайте https://2019-1.tpc.ispras.ru. Пожалуйста, вводите правильные данные, так как именно они будут использоваться при выставлении зачетов.

Далее система будет показывать тексты новостных статей. Для каждого предложенного текста необходимо назначить одну или более метку в соответствии с текстом новости. Метка «Прочее» должна быть назначена, если статья не относится ни к одной из предложенных тем. Примеры разметки новостей приведены в приложении А.

В систему загружено 3000 текстов новостей (длина каждой новости не менее 1000 и не более 3000 символов). Каждому человеку предлагается разметить не менее 100 случайно выбранных новостей. Система перестает предлагать новость для разметки, если она была размечена 3 разными людьми. Информацию о различиях в разметке можно использовать при обучении алгоритмов.

После того, как будут размечены не менее 40 новостей, появится кнопка, позволяющая скачать размеченные новости, и станет доступна загрузка решений.

Рекомендуется размечать максимально честно, так как от этого будет зависеть качество всех моделей.

Тренировочный корпус

Тренировочный корпус будет доступен для скачивания в формате json. Для извлечения информации из этого файла рекомендуется использовать стандартную библиотеку Python с одноименным названием.

Для синхронизации обучения и тестирования в течение недели, корпус будет состоять из отзывов, размеченных автором классификатора, плюс все новости, размеченные в течение предшествующей недели.

Тестирование

Вместе с кнопкой скачивания тренировочного корпуса появится ссылка на форму для загрузки файла и личную страницу со статистикой. На личной странице находится статистика со всеми результатами в т.ч. результатами последнего тестирования (дата, описание, метрика качества).

Загрузка решения

Загружаемый файл должен представлять собой zip архив с любым именем. Архив должен обязательно содержать:

- Решение в файле ***solution.py***. В файле должен содержаться класс **Solution**. В классе должны присутствовать методы
    - *train(self, train_corpus: List[Tuple[str, str, Dict[str, Set[str]]]]) -> None*, где train_corpus – это список троек <заголовок новости; текст новости; разметка новости>. Разметка новости представляет собой отображение идентификатора автора разметки в множество меток тем.
    - *predict(self, news: List[Tuple[str, str]]) -> List[Set[str]]*, который получает на вход список пар <заголовок новости; текст новости> и возвращает список множеств меток.
- (пустой) файл ***__init__.py*** (требования к пакетам Python)
- описание применяемых методов в файле ***description.txt***. Пожалуйста, напишите подробное описание, какие методы и признаки использовались. Это описание будет выложено вместе с решением после завершения курса.
- все используемые ресурсы, необходимые для корректной работы метода.

Результаты тестирования появятся на личной странице, как только закончится обучение и тестирование. При загрузке нового классификатора обучение будет производиться на корпусе из новостей, размеченных автором решения, плюс все новости, размеченные в течение предшествующей загрузке недели. В течение недели студенты не видят прогресс своих коллег и могут посмотреть только свой результат.

Ограничения

- каждую неделю можно послать не более 7 решений. **!Внимание!** Итоговое тестирование будет проводиться на последнем загруженном решении
- размер загружаемого архива не должен превышать 15Мб
- Время тестирования одного решение (обучение + предсказание) не должно превышать 30 минут
- На проверяющей машине доступно 16Гб оперативной памяти

В связи с первым ограничением, для тестирования на локальной машине рекомендуется использовать метод перекрестной проверки.

Оценка качества

Для оценки качества используется micro-averaged F1-мера, вычисляемая как среднее гармоническое micro-averaged precision и micro-averaged recall.

$$F_1 = \frac{2PR}{P+R};$$

$$P = \frac{\sum_{c \in C} tp_c}{\sum_{c \in C} tp_c + fp_c};\ \ R = \frac{\sum_{c \in C} tp_c}{\sum_{c \in C} tp_c + fn_c},$$

Где tp_c, fp_c, fn_c – true positive, false positive, false negative для класса c.

Baseline решение

В качестве baseline решения предлагается модель, основанная на искусственных нейронных сетях (реализована с помощью пакета `torch`).

В качестве признаков используются TF-IDF слов текстов. Признаки подаются в MLP (многослойный персептрон), выходной слой - sigmoid на каждый класс. В качестве функции потерь используется усреднение бинарной кросс-энтропии по каждому классу.

В процессе обучения подбираются гиперпараметры (скорость обучения и cut-off для TF-IDF).

Язык текста определяется по наличию кириллических букв в нем.

Для токенизации используется пакет `nltk`, для определения границ предложений модели `punkt` из `nltk`.

Приложение А. Примеры разметки новостей

Название

Все пользователи Google Chrome в серьезной опасности. На кону данные банковских карт и пароли

Текст

Наиболее популярным в мире веб-браузером является Google Chrome, база активных пользователей которого уже давно превысила отметку в 2 млрд. Именно поэтому злоумышленники стараются делать ставку на него для продвижения своего вредоносного программного обеспечения. Сегодня, 20 мая 2018 года, сотрудники компании Proofpoint, которая занимается изучением безопасности в интернете, сообщили о новом вирусе под названием Vega Stealer. Он угрожаем всем пользователям популярного интернет-обозревателя.

Как сообщают эксперты, обнаруженный зловред занимается кражей данных банковских карт, а также логинов и паролей. Вся эта информация хранится в базе Google Chrome, получить доступ к которой расширения или какие-либо дополнения не могут, но ПО под названием Vega Stealer – может. Оно представляет из себя доработанную версию трояна August Stealer, который был распространен еще в декабре 2016 года.

Новый вредоносный скрипт, как правило, передается по электронной почте. Пользователям рассылаются фейковые электронные письма под видом таковых от крупных компаний, вроде Apple, Google, Microsoft, Intel и прочих. Предлагают ознакомиться с содержимом прикрепленных файлов под различными предлогами, одним из которых является документ «brief.doc». При его открытии автоматически срабатывают макросы, которые запускают троян Vega Stealer, а он, в свою очередь, незамедлительно начинает собирать данные и передавать их злоумышленникам.

Все данные передаются на удаленные сервера, а на выполнение всех задачи зловреду нужно около одной минуты. Пока пользователь изучает документ, троян выкачивает с компьютера все важные сведения. Представители Proofpoint отмечают, что программное обеспечение Vega Stealer обладает большой гибкостью, поэтому оно легко заражает десятки тысяч компьютеров ежедневно. Если так пойдет и дальше, то вскоре троян станет гораздо более известным и популярным.

Эксперты советуют с большой осторожностью относиться ко всем электронным письмам от неизвестных отправителей, а также ни в коем случае не открывать приложенные файлы, потому как в противном случае можно лишиться данных своей банковской карты, а вместе с тем и всех денежных средств, которые на ней хранятся. Точно такая же участь может постигнуть логины и пароли от различных веб-сайтов.

Ранее Google добавила в браузер Chrome специальную функцию, благодаря которой на экран телевизора можно транслировать фильмы, сериалы и музыку.

Метки тем

- «Угроза» - эта метка должна быть назначена в связи с тем, что в статье описывается угроза кражи данных банковских карт

- «Вредоносное ПО» - в статье описывается троян Vega Stealer, который является вредоносным программным обеспечением (предназначен для несанкционированного доступа к информации)

Название

В Firefox пропатчена атакуемая уязвимость 0-day

Текст

Mozilla в экстренном порядке выпустила Firefox 67.0.3 для Linux, macOS и Windows, чтобы устранить уязвимость, уже пущенную в ход злоумышленниками. Поскольку Firefox обновляется автоматически, пользователи смогут установить патч, просто перезапустив браузер.

Согласно бюллетеню разработчика, уязвимость нулевого дня CVE-2019-11707 проявляется как путаница типов данных (type confusion). Ошибка может возникнуть в ходе работы с массивами в JavaScript, а точнее, при применении метода pop() к объектам Array. (Метод pop изменяет исходный массив, удаляя из него последний элемент, и возвращает значение этого элемента.)

Злоумышленник может воспользоваться уязвимостью, вынудив пользователя открыть в браузере страницу с вредоносным JavaScript-сценарием. В случае успешной атаки ее автор сможет захватить контроль над системой жертвы.

В бюллетене также сказано, что в Mozilla имеются данные об использовании уязвимости в целевых атаках. Степень ее опасности оценена как критическая.

Новую проблему в Firefox обнаружил участник проекта Google Project Zero Сэмьюэл Грос (Samuel Groß) — тот же исследователь, который два года подряд успешно атаковал браузер Safari в рамках конкурса Pwn2Own — в 2017 году и 2018-м.

Firefox был обновлен до версии 67 совсем недавно, в мае. С ее выпуском Mozilla пропатчила 21 уязвимость. В браузер также добавили опцию блокировки скриптов, нацеленных на майнинг криптовалюты либо отслеживание пользователей по цифровым отпечаткам.

Метки тем

- «Угроза» - эта метка назначена новости в связи с описанием угрозы захвата контроля системы
- «Уязвимость» - эта метка назначена в связи с описанием уязвимости CVE-2019-11707
- «Эксплойт» - в статье описан способ эксплуатации уязвимости CVE-2019-11707 (применение метода pop() к объектам Array)

Метка «Инцидент» в данном случае не должна быть назначена, поскольку не описывается ни одного события нарушения информационной безопасности (упоминается лишь наличие данных об использовании уязвимости в целевых атаках)

Название
ФСБ: атаки хакеров не могут вызвать выход из строя выборных сайтов

Текст
Силовые структуры России отразили атаки хакеров проникнуть на открытые интернет-ресурсы, освещающие выборы. Об этом сообщили сегодня источники в правоохранительных органах, отметив, что "значительного увеличения хакерских атак в день выборов не произошло".

"В выборной системе отработана надежная техническая защита. В рамках российского законодательства соответствующие службы отслеживают ситуацию. Лица, причастные к нарушениям российского законодательства, могут быть наказаны в уголовном порядке", - сказал представитель Центра общественных связей ФСБ России.

Он заверил, что "система защиты выборных сайтов построена так, что атаки хакеров не могут вызвать ее выход из строя".

Метки тем
- «Прочее» - эта метка назначается в связи с тем, что статья не относится ни к одной теме из списка.

Название
Онлайн-магазин Sony Ericsson взломан хакерами

Текст
Хакеры атаковали онлайн-магазин канадского отделения компании Sony Ericsson.

По словам представителя Sony, в результате взлома были похищены личные данные 2 тысяч пользователей. Затем информация была опубликована на сайте "The Hacker News".

Как сообщается, данные включали имена, адреса электронной почты и пароли в зашифрованном виде. Номера кредитных карт клиентов Sony Ericsson в украденный массив информации не попали.

Метки тем
- «Инцидент» - в статье описано событие нарушения информационной безопасности (кража личных данных 2 тысяч пользователей)

Название

Найден способ блокировки Bad Rabbit

Текст

Эксперты нашли способ заблокировать активность нашумевшего шифровальщика Bad Rabbit, терроризирующего СМИ России и Украины.

Специалисты предлагают следующую последовательность действий:

создать файл C:\\windows\\infpub.dat

выставить этому файлу права «только для чтения».

По словам исследователей, это позволит сохранить ваши файлы нетронутыми даже в том случае, когда шифратор попал в систему.

Напомним, что раннее мы писали о больших масштабах деятельности этого вредоноса. Так, например, ESET зафиксировала сотни атак BadRabbit на Россию и Украину. Bad Rabbit добрался даже до Новой Газеты.

Метки тем

- «Вредоносное ПО» - статья содержит описание способа блокировки активности вируса-шифровальщика Bad Rabbit
- «Инцидент» - в статье упомянуты события атак вируса Bad Rabbit.