Отчет по хакатону

Stardust Crusaders

22 ноября 2020 г.

Содержание

Резюме

`Основной github`: ссылка

`Обработанные данные`: ссылка

Написаны четыре модели, были проведены простые тесты (о каких-то статистических величинах говорить рано), построены гипотезы относительно природы данных.

Постановка задачи

В компании вводятся так называемые “барьеры безопасности”, анализ которых сможет предотвратить негативные происшествия. Однако аналитика сопряжена с рядом проблем:

- 16-20 тысяч сообщений на естественном языке в месяц. Объем регистрируемых сообщений об опасных действиях/опасных условиях с производственных площадок по компании
- 5 рабочих дней затрачивается на сбор обработку сообщений сотрудниками блоков и ДО. Длительный процесс не даёт возможность оперативно реагировать на сообщения о работоспособности производственных барьеров
- Человеческий фактор при классификации. Субъективный результат классификации сообщений и как результат – отсутствие сквозной аналитики (между ДО)

Сейчас существует цифровой проект «Определение барьеров производственной безопасности из сообщений об опасных условиях-опасных действиях».

В рамках проекта разрабатывается инструмент для классификации записей об опасных действиях и условиях, поступающих от ДО и подрядных организаций в различных информационных системах (контур безопасности, ОУ ОД БРД, АЗИМУТ и другие).

Решение позволит повысить достоверность данных, на которых принимаются решения о работоспособность барьеров и делается оценка рисков ПБ. Появление новой аналитики позволит сфокусировать усилия сотрудников на предупреждение «пробоя» барьера промышленной безопасности. Сократятся трудозатраты на обработку заявки и создание аналитики с 5 дней до нескольких часов.

В файле представлены данные для оценки и первичного анализа (файл мы пришлем всем командам, которые выбрали данное направление в субботу, 21 ноября после начала хакатона)

- В графе place — место, где произошел прецедент промышленной безопасности
- В графе precedent — сам прецедент. Что такое прецедент — опасное действие, условие.

Замечания по данным:

- Иногда в графу прецеденты респонденты заносят сразу несколько опасных действий и условий
- Данные содержат аббревиатуры, термины, неточные выражения, опечатки и сленг

Необходимо:

- Построить модель кластеризации. В дальнейшем эти кластеры помогут провести аналитику для сопоставления прецедентов с классификацией опасных действий и условий — с барьерам безопасности.
- Объяснить получившиеся кластеры: почему они такие. Сделать вывод, можно ли с помощью них упростить какую-либо полуавтоматическую разметку данным по классам

Глобально можно выделить несколько возможных кластеров:

- Прецеденты, связанные с вождением, и управлением транспортным средством (ТС)
- Прецеденты, связанные с использованием исправного инструмента
- Прецеденты, связанные с работой на высоте и с подъемом грузов
- Прецеденты, связанные с защитой (СИЗ (средства индивидуальной защиты), СИЗОД (средства индивидуальной защиты органов дыхания), перчатки, куртки и тп)
- Все остальное

Используемый датасет

Был предоставлен компанией-заказчиком.

Описание решения

Особенности обработки данных

Исходный `.xlsx` файл расщеплен на 5-столбцовый `.csv` файл, которые и уже обрабатывался. Дополнительно был построен словарь всех слов и словарь с подсчетом частот

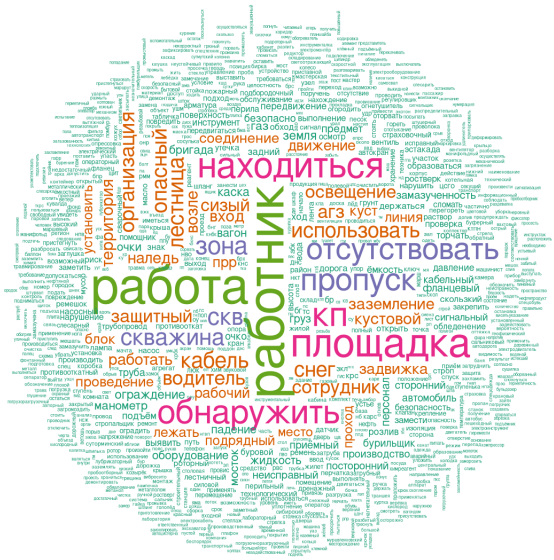

Рис. 1: Облако слов для очищенных данных, построенное с помощью языка R

При первичной очистке данных были удалены строчки таблицы начиная с `62000` заканчивая по `84000`, так как эти строки не соответствовали общему формату данных.

При обработке данных в первых трех моделях использовался `PyMoprhy`: в связи с чем происходила потеря значений аббревиатур и терминологии.

covid-19 встречался в сообщениях как на латинице, так и кириллице, но первый вариант так же был удален для использования `PyMoprhy`.

Для последних моделей использовался словарь аббревиатур и сторонние специализированные корпусы.

Модели

В рамках поставленной задачи были построены 4 модели кластеризации:

- Ручная кластеризация по ключевым словам;
- Кластеризация с помощью K-mean и Word2Vec, с использованием платформы KNIME;
- Кластеризация с помощью K-mean и Word2Vec, на питоне с дополнительным препроцессингом данных;
- Кластеризация с помощью K-mean и Word2Vec с внешним корпусом текстов и TensorFlow.

Краткое сравнение моделей

	Ручная кластеризация	KNIME	Python	Advanced
Сложность настройки	Низкая	Низкая	Средняя	Высокая
Гибкость модели	Низкая	Низкая	Средняя	Высокая
Точность	Средняя	Высокая	Средняя	?
Масштабируемость	Нулевая	Средняя	Средняя	Высокая

Таблица 1: Сравнение четырех моделей

Краткий обзор идей для анализа

На текущий момент мы предложили упорядочивать данные по следующей структуре:

Рис. 2: Структура данных

и исследовать особенности каждого уровня. В частности, искать сущности-*аттракторы*, у которых какие-то слова или другие свойства сообщений (`msg`) встречаются чаще, чем в среднем. Так, мы искали предприятия, на которых плохо соблюдается профилактика covid-19: `grep` по строчкам очищенной таблицы содержащим слова “ковид“ “пандемия“ показал, что организация “НоябрьскЭнергоНефть“ отправила большую часть сообщений с этими словами.

Полноценный поиск аттракторов не удалось выполнить. Так же, в рамках существуешь структуры данных предлагалось исследовать наличие пустых сообщений или сообщений дубликатов.

Структура репозитория

В корневом каталоге содержатся:

- `prepare.py` — модуль, который очищает и подготавливает данные для работы
- `reports/` — каталог с отчетами и презентацией.
- `keywords-model/` — каталог с моделью ручной кластеризации
- `word2vec-model/` — каталог с моделями базирующихся на Word2Vec и K-means, написанных на python.

Так же есть каталог `data/` который не отображается в github‘е. В `data/` находятся следующие файлы:

- `precedents.csv` — сырые данные преобразованные в `.csv`.
- `fixed-precedents.csv` — расщепленная таблица.
- `documents.csv` — таблица `fixed-precedents.csv` в которой все столбцы сведены в один для Word2Vec.
- `frequency.dic` — уникальные слова в нормальной форме для поиска ключевых слов и построения облака.
- `stopwords.txt` — список стоп-слов.

На том же гугл-диске находится папка `clustering` где расположены файлы моедли KNIME.

Результаты

Ручная модель работает исправно и позволяет делать предсказания:

`Input: "Рабочий взял оголенный электрокабель" Output: "Прецендеты связанные с электричеством"`

Модель на основе Word2Vec выделила порядка 14 кластеров, которые требуют дополнительной обработки — часть кластеров близка по смыслу друг к другу, поэтому их можно объединить.

Так же, не до конца удлаось понять смысл кластеров — кажется, что длz этого надо взять случайную выборку представителей каждого кластера и на основании этого попытаться присвоить какую-то осмысленную категорию.

Распределение:

- Ручная модель, очистка данных, отчеты - Михайлов Михаил
- Модель word2vec, очистка данных - Данил Ельцов